

Universität Heidelberg

Holistic Evaluation of Text Summarization

Eine Heidelberger Dissertation

JULIUS MAXIMILIAN STEEN

THIS DISSERTATION IS SUBMITTED FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY

First supervisor: Prof. Dr. Katja Markert

Second supervisor: Prof. Dr. Simone Paolo Ponzetto

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Katja Markert for her invaluable support and advice throughout this journey. I would also like to thank Simone Paolo Ponzetto for agreeing to be the second assessor of this thesis and Anette Frank for her advice and collaboration on the work on faithfulness that ended up as the fifth chapter of this work.

I am very grateful to my colleagues Juri Opitz, Frederick Riemenschneider, Letitia Pârcălăbescu, Moritz Plenz, Xiyang Fu, Michael Staniek, Michael Hagmann, Nathan Berger, and Marius Fracarolli for their support and their companionship during many lunch breaks in my time at the Institute for Computational Linguistics. Especially big thanks go to Michael and Juri for their helpful feedback on drafts of this thesis.

I will forever be grateful to Sibylla and to my family, especially to my mother Birgit and to my grandmother Ingrid, for putting up with me during the conception of this thesis.

Finally, I would like to thank my father, Frank. While he never got to see the conclusion of my long academic journey, I would not have started it without him.

Abstract

Text summarization is a vital tool to make large bodies of texts easily digestible for human readers. To develop powerful automatic summarization systems, researchers rely on evaluation protocols to assess the progress made by newly proposed summarizers. However, as we will show in this thesis, current protocols are not always sufficient to provide reliable feedback on summarizer performance across a wide range of quality dimensions. In this work, we will thus aim to develop a framework for *holistic* evaluation of text summarization that covers a broad range of quality dimensions and evaluation settings. In addition to this holistic coverage of quality dimensions and settings, two criteria will guide our investigations: *Reliability*, which ensures evaluations lead to comparable results across different settings, and *cost-efficiency*, which is critical to ensure evaluations can be run frequently and exhaustively.

We will begin our investigation at the “gold standard” of summarization evaluation, the human evaluation study. Here, we will show weaknesses in current practices that jeopardize their reliability. Our work will formulate concrete proposals to improve current practices to create both more reliable and cost-efficient human studies. Since even cost-efficient human evaluation is still prohibitive for many use cases, we will then turn our attention to *automatic* evaluation, starting with an assessment of common meta-evaluation practices. We find that current practices are at risk of leading to unreliable conclusions on evaluation metric performance. We will use these insights to conduct an in-depth meta-evaluation of automatic summary coherence measures. In the final two parts of this thesis, we will then focus on automatic evaluation for two important quality dimensions, which have only recently started to receive attention in text summarization: *Faithfulness* and *Bias*. For faithfulness, which is the degree to which a summary correctly reproduces facts from the input, we find that currently proposed metrics are usually computationally expensive. This motivates us to search for a cost-efficient automatic faithfulness metric. Finally, we find that social *bias*, which is a frequently studied phenomenon in other NLP tasks, has not yet been systematically investigated for text summarization. We will thus provide both abstract definitions as well as practical automatic metrics to assess the presence of bias in summarization systems.

As a whole, our work will provide researchers and users who are interested in the performance of summarization systems a toolbox to cost-efficiently and reliably assess summarizers across key quality dimensions.

Contents

1	Introduction and Research Questions	1
1.1	Summarization Evaluation	1
1.2	Research Questions	4
1.3	Contributions	5
1.4	Thesis Overview	6
1.5	Published Work	6
1.6	Published Code	6
2	An Overview of Evaluation in Text Summarization	9
2.1	Text Summarization: Tasks and Datasets	9
2.2	Text Summarization Systems	14
2.2.1	Extractive Summarization	14
2.2.2	Abstractive Summarization	16
2.3	Human Evaluation of Text Summarization	17
2.3.1	Extrinsic Evaluation	17
2.3.2	Intrinsic Evaluation	18
2.4	Automatic Evaluation Metrics	22
2.4.1	Reference-based Metrics	22
2.4.2	Reference-free Metrics	25
2.5	Discussion	26
3	Human Evaluation of Summarization Systems	29
3.1	Motivation	29
3.2	Background and Related Work	32
3.2.1	Reliability as a Proxy for Validity	32
3.2.2	Human Evaluation with Crowd Workers	37

3.2.3	Methods for Judgement Elicitation	39
3.2.4	Null Hypothesis Significance Testing	41
3.3	Literature Survey of Evaluation Practices	44
3.3.1	Methods	46
3.3.2	Statistical Analysis	46
3.3.3	Design	47
3.4	Coherence and Repetition Annotation	47
3.5	Ranking vs. Likert	51
3.5.1	Reliability	52
3.5.2	Cost Efficiency	54
3.6	Statistical Analysis and Type I Errors	55
3.6.1	Mixed-Effects Models	56
3.6.2	Ordinal Regression	57
3.6.3	Hypothesis Testing using Generalized Mixed Models	58
3.6.4	Modelling our Annotations	58
3.6.5	Demonstrating the Dangers of Ignoring Grouping Factors	60
3.7	Study Design and Study Power	62
3.7.1	Overall Number of Annotators	62
3.7.2	Annotator Distribution	63
3.8	Five Years Later: Have Practices Changed?	66
3.9	Discussion	69
4	Meta-Evaluation: A Case Study in Summary Coherence	71
4.1	Motivation	71
4.2	Background	74
4.2.1	Meta-Evaluation	74
4.2.2	Measuring Coherence	78
4.3	Related Work	84
4.4	Selecting a Meta-Evaluation Dataset	85
4.5	A new Meta-Evaluation Metric	88
4.5.1	A new Evaluation Metric: Intra-System Correlation	89
4.5.2	System-level Confounders	90
4.6	Coherence Measures	92

4.6.1	Entity Grid and Extensions	94
4.6.2	Lexical Coherence Models	98
4.6.3	Supervised Linguistic Quality Model: SumQE	102
4.6.4	Unsupervised Linguistic Quality Models	103
4.7	Results	106
4.7.1	Detailed Intra-System Correlation Results	107
4.8	Bias Matrices	108
4.9	Coherence Measure Analysis	112
4.9.1	Correlation with Shuffle-Performance	112
4.9.2	GRUEN	112
4.9.3	Entity Driven Measures	114
4.9.4	Global Training vs. Pairwise Ranking	114
4.10	Discussion	117
5	Faithfulness Evaluation with NLI Models	119
5.1	Motivation	119
5.2	Background and Related Work	121
5.2.1	Faithfulness and Factuality	121
5.2.2	Metrics	123
5.2.3	Datasets and Meta Evaluation	126
5.3	Method Details	128
5.3.1	Task-adaptive Data Augmentation	128
5.3.2	Monte-Carlo Dropout	130
5.4	Experimental Setup	131
5.5	Results	132
5.6	Effect of Dialogue Adaptation	133
5.7	Phrase Selection Robustness	136
5.8	Phrase Ablation Experiments	137
5.9	Effect of Integrating Contradiction Scores	144
5.10	Bias Analysis	145
5.11	Cost Comparison to Other Approaches	148
5.12	Discussion	148

6	Social Bias Evaluation	151
6.1	Motivation	151
6.2	Background	154
6.3	Related Work	156
6.4	Defining Bias in Text Summarization	159
6.5	Bias Metrics	160
6.5.1	Inclusion: Word Lists	160
6.5.2	Inclusion: Entity Inclusion Bias	161
6.5.3	Hallucination: Entity Hallucination Bias	162
6.5.4	Representation: Distinguishability	163
6.6	Input Documents are Already Biased	164
6.7	Gender Bias Experiments	167
6.7.1	Dataset	167
6.7.2	Template Construction Algorithm	168
6.8	Metric Implementation Details	170
6.8.1	Entity Alignment	170
6.8.2	Identifying Gender of Hallucinated Entities	171
6.9	Summarizers	172
6.9.1	Models	172
6.9.2	Gender Bias Summary Statistics	173
6.10	Gender Bias Results	173
6.11	Validating our Metrics	175
6.11.1	Validation of our Alignment Algorithm	176
6.11.2	Summary Quality	178
6.11.3	Content Words	179
6.11.4	Induced Bias Detection	182
6.12	Gender Bias Analysis	182
6.12.1	Investigating Hallucination Bias	182
6.12.2	Investigating the Effect of Replacing Last Names	184
6.12.3	Investigating Distinguishability	184
6.13	Extension to Race Bias	186
6.13.1	A Dataset for Race Bias in Summarization	186
6.13.2	Results	187

6.13.3 Investigating Distinguishability	187
6.14 Discussion	189
7 Conclusions and Future Work	191
7.1 Conclusions	191
7.2 Outlook	193
Bibliography	197
A Summarization Dataset References	255
B Survey	257
B.1 Categories	257
B.2 Survey Files	258
B.3 Files for the Repeat of the Survey	260
C Coherence Measures: Implementation Details	261
C.1 Extended Entity Grid (EEG)	261
C.2 Entity Graph (EGR)	261
C.3 Neural Entity Grid (NEG)	262
C.4 Graph-based Model (GRA)	262
C.5 Unified Coherence Model (UNF)	263
C.6 Coherence Classifier (CCL)	263
C.7 BARTScore (BAS)	263
C.8 GRUEN (GRN)	264
C.9 SumQE (SQE)	264
D NLI Model Augmentation Training Details	265
D.1 Hyper-Parameters	265
D.2 Training	266
E Dataset Bias in BEGIN-v2	267
F Bias Experiment Topic Assignment Heuristic	269
G Dataset Statistics for Intersectional Biases	271

List of Tables

1.1	Links to repositories containing the code underlying the work in this thesis.	7
2.1	Dataset statistics and task descriptions for various summarization datasets.	12
3.1	Survey of 58 system papers with 95 manual evaluation studies (2017-2019).	45
3.2	Results of our annotation experiment.	52
3.3	Krippendorff's α with ordinal level of measurement and Split-Half Reliability for both annotation methods on the two quality dimensions.	52
3.4	Results of our new survey of 8 papers with 11 studies from EACL 2024 and NAACL 2024.	67
4.1	Results for the confounders and upper bound.	91
4.2	Training settings for the CMs under investigation.	92
4.3	Results on SummEval for all CMs.	107
4.4	Shuffle accuracies on CNN/DM for 1000 randomly sampled reference summaries.	113
4.5	Performance of GRUEN constituent measures.	113
4.6	Proportion of documents without any entity overlap, as well as average ratio of sentences without entity links per document for various datasets.	114
5.1	Manually curated list of dialogue phrases.	129
5.2	Dataset statistics for all constituent corpora in TRUE.	131
5.3	AUC scores for all models on TRUE.	132
5.4	AUC differences for individual modifications of Base	133

5.5	Kendall's τ correlations of gold labels/system scores with first-person pronoun occurrence.	134
5.6	Results of our phrase selection robustness analysis.	136
5.7	Statement augmentation phrases.	137
5.8	Average AUC over 15 different runs of ablation experiments with original and alternative augmentations for the append and prepend setting.	139
5.9	Average changes in entailment-only score of models trained with different augmentations relative to the ANLI-only models.	141
5.10	Correlations of token overlap between generation and input with model predictions for Orig. Stmt. and ANLI on CNN/DM derived corpora.	142
5.11	Correlations of the position of the first error and model scores on unfaithful instances for all augmentation settings and ANLI.	144
5.12	Original and intra system ROC AUC scores for the four datasets where model information is available.	146
5.13	Performance vs. cost analysis.	148
6.1	Male and female word lists reproduced from HELM.	157
6.2	Ten most male/female associated words in CNN/DM and XSum, with z-scores.	165
6.3	Number of documents and % of female identifiers per topic and word list inclusion scores of our simulation experiment.	167
6.4	Pronouns used for gender classification in Wikipedia articles.	171
6.5	Prompts used for the Llama-2 models.	173
6.6	Summary statistics for summaries generated on C_{loc} and C_{glob} for gender bias, and on original documents.	174
6.7	Results of our bias metrics.	175
6.8	Results of our manual annotation of entity alignments.	177
6.9	GPT-3.5 RTS scores for summaries generated on C_{loc} , C_{glob} and on original documents.	179
6.10	Extended word list used to identify candidate documents for annotation.	180

6.11	Results on our manually extended variants of C_{loc} and C_{glob} for gender bias with content words altered to conform to entity gender.	181
6.12	Inclusion bias scores on Llama-2 13b prompted to induce an inclusion bias towards female entities.	182
6.13	Ten most frequent PERSON named entities without source alignment in the generated summaries.	183
6.14	Results for entity metrics computed on C_{loc} for gender bias with last names altered.	185
6.15	GPT 3.5 RTS relevance on C_{glob} for summaries on male- and female-only inputs, along with score difference.	185
6.16	Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on race bias data with randomly assigned genders.	186
6.17	Bias scores for race bias with black/white associated names with different gender assignments.	188
6.18	Quality difference scores for race bias with random gender assignment.	189
A.1	References for datasets listed in Table 2.1.	255
C.1	Best hyper-parameters for the neural entity grid on DUC 03.	262
D.1	Hyper-parameters for training models with augmentations.	265
F.1	Words used for topic identification.	269
G.1	Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black male, white female).	272
G.2	Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black male, white male).	272
G.3	Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black female, white female).	273

G.4	Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black female, white male).	273
-----	--	-----

List of Figures

3.1	Screenshots of the annotator instructions.	48
3.2	Screenshots of the annotation interfaces.	49
3.3	Schematic representation of our study design.	51
3.4	Score distribution of Likert for both quality dimensions.	53
3.5	Time spent on annotation (in minutes) vs. correlation with the full-sized score.	54
3.6	Relation of Type I error rates at $p < 0.05$ to the total number of annotators for different designs.	61
3.7	Power for 100 documents and 3 judgements per summary with different number of total annotators.	63
3.8	Reliabilities of nested vs. crossed designs for Rank and Likert for both quality dimensions.	64
3.9	Power for $p < 0.05$ of nested and crossed designs for ARTagg and regression.	66
4.1	Distribution of human coherence scores for the 17 systems in the SummEval dataset.	88
4.2	Intra-system correlations of the best CMs as well as the human upper bound on the SummEval dataset.	108
4.3	Bias matrices for the best CMs.	110
4.4	Bias matrix for BAS with specific analysis for BART and Pegasus.	111
4.5	Histograms of the lengths of summaries generated by the summarizers in SummEval and their mean lengths.	115
4.6	Ranking accuracy between shuffled and original summaries of different lengths (in characters).	116

5.1	Histogram of the score distributions with and without $e-c$ for faithful and non-faithful instances.	145
5.2	Bias matrices for our best-performing model and three top competitors on TRUE.	147
6.1	Schematic overview of our approach for summary gender bias evaluation with an example generated by BART XSum.	152

Chapter 1

Introduction and Research Questions

1.1 Summarization Evaluation

Automatic text summarization is the automatic generation of concise, informative summaries from a given document or set of documents. As such, it plays an important role in helping users process and sort through the ever-growing amount of information available to them. As with any natural language generation (NLG) task, progress in text summarization is driven by a continuous assessment of the capabilities of newly proposed summarization systems using both human and automatic evaluation metrics. The quality of this evaluation in turn is what allows the research community to accurately quantify the progress made in the field and to identify promising directions for improvement. This is especially critical in a field that has seen, alongside many other areas in natural language processing (NLP), an exponential improvement in capabilities (Lewis et al., 2020; Zhang et al., 2020a; Goyal et al., 2022).

In summarization evaluation, much attention has been paid to the evaluation of summary *relevance* or *informativeness*, i.e. whether the summary reflects the important information in the input document(s). This is commonly measured by comparing a generated summary to one or more human-written reference summaries. This process is either done automatically, often using overlap metrics, such as ROUGE (Lin, 2004b), or in a systematic manual fashion, as in the PYRAMID (Nenkova and Passonneau, 2004) evaluation framework. However, as are

most NLG tasks, summarization evaluation is a multidimensional problem. For example, a summary that contains all relevant information in the input can still be poorly written and incoherent to the point it does not provide a benefit over the input documents. Traditionally, additional quality dimensions cover linguistic quality concerns, including grammaticality, readability, and coherence (Dang and Owczarzak, 2009a). With the rapid development of improved summarization systems, however, new quality dimensions have started to receive attention. Of particular concern is the *faithfulness* (Maynez et al., 2020) of summaries, which refers to the extent to which content in the summary is supported by the input.

The breadth of relevant quality dimensions and the rapid development of summarization systems lead us to ask whether current evaluation practices give an adequate picture of summarizer performance. In this thesis, we thus aim to take stock of the toolbox currently available for summarization evaluation and ask what is missing to give researchers and users a *holistic* view of summarizer performance. By the end of the thesis, we aim to have established a set of evaluation best practices and automatic methods that, in conjunction with prior work, allow us to assess the quality of summarization models across a wide array of quality dimensions. Two criteria will guide our exploration of this space: *reliability* and *cost-efficiency*.

By reliability, we mean that evaluation should be consistent across different settings and not affected by non-material changes to the evaluation setup. For example, resampling the documents used as input for a summarizer (from the same distribution) should not alter our conclusion about its performance characteristics. This is a common definition of reliability in NLP (see Riezler and Hagmann, 2024). Reliability is critical to ensure evaluation gives useful feedback on model capabilities. *Cost-efficiency*, on the other hand, is a more practical concern. Researchers typically only have a limited budget available to evaluate new summarizers. For an evaluation procedure to provide useful feedback, it must thus be cheap enough to run frequently. For automatic evaluation, that means it should need as few computational resources as possible. For manual evaluation, it means minimizing the amount of human effort required.

We will begin our investigation at what is often regarded as the *gold standard* (Gehrmann et al., 2023) of evaluation: Human judgements. As we will show,

human evaluation is poorly standardized in summarization. Current practices are usually neither cost-efficient nor are the conclusions drawn from them reliable due to inappropriate statistical tools being used in the analysis of raw results. We will establish a set of best practices for the selection of annotation methods, study design, and analysis for human evaluation studies.

While our approach will allow for cost-efficient human evaluation, the need for human labor still makes it expensive and time-consuming. We will thus then turn our attention to automatic evaluation. For automatic metrics to be a useful signal, they must themselves be evaluated for their correspondence to human judgements. As a first step, we are thus going to investigate current practices in this *meta*-evaluation. We will identify threats to the reliability of meta-evaluation in the form of confounding system properties and propose methods to remedy them.

With the fundamental framework in place, we will then identify three quality dimensions which have a strong need for improved automatic evaluation procedures: *Coherence*, *faithfulness*, and *bias*.

For **coherence**, we find that while a large number of coherence measures have been proposed, a lack of standardized meta-evaluation makes it difficult to select promising coherence measures which work well as evaluation metrics. We will use our meta-evaluation methodology to assess the state of summary coherence modelling and to identify promising directions for further improvements.

For **faithfulness**, we will show that, while there is a large number of metrics available, the best-performing ones are computationally very costly. This leads us to investigate whether we can close the gap to more expensive metrics by finding a cheap, yet powerful, faithfulness evaluation metric.

Finally, we identify the lack of summarizer social **bias** analysis as an important gap in current evaluation practices. While it has long been established that harmful social biases are present in the (pre-)training corpora for large language models (LLMs) (Barocas et al., 2017), which also underlie contemporary summarizers, it is unclear to which extent bias is propagated to automatic summaries. This motivates us to develop definitions for biased behavior in summarization and practical metrics to identify it.

Taken together, the work in this thesis both helps close critical gaps in automatic summarization evaluation and supports future evaluation with a set of

reliable best practices for both human evaluation studies and meta-evaluation.

1.2 Research Questions

We seek to answer the following research questions in this thesis:

Research Question I: How can we conduct cost-efficient and reliable human evaluation? Human annotation plays an important role as a gold standard evaluation for summarization systems. However, there are few standards for designing these studies in a way that leads to cost-efficient and reliable results. We investigate how different annotation methods for eliciting human linguistic quality judgements differ in reliability and efficiency. We also consider the question of how to analyze study results in a way that minimizes the risk of drawing erroneous conclusions about summarizer performance. Finally, we establish a set of guidelines to help researchers design their human evaluation studies.

Research Question II: How can we ensure reliable meta-evaluation of summarization metrics? Automatic metrics complement human evaluation by providing cheaper and faster feedback on summarizer performance. However, for them to be reliable stand-ins for human judgements, they themselves must be properly evaluated. We design methods for analyzing meta-evaluation results with a focus on reducing the impact of confounding factors that limit their generalizability. We will then use these methods to identify promising coherence measures for automatic summary coherence evaluation.

Research Question III: How can we conduct efficient automatic faithfulness evaluation? Recent summarization models suffer from a phenomenon called *hallucination*, where summaries introduce new facts that are unfaithful to the source. This reduces their reliability and trustworthiness in real-world use and calls for automatic metrics to identify unfaithful summaries. Recent faithfulness evaluation methods use complex, computationally expensive setups. We instead ask how small NLI-based models can perform competitively, without introducing costly inference time mechanisms, and conduct a meta-evaluation of our newly proposed metric.

Research Question IV: How can we identify social biases in summarization systems? As large language models become increasingly important for summarization, there is a risk of summarization systems exhibiting and amplifying social biases acquired during pretraining. We first ask what it means for a summarizer to be biased and develop a set of definitions for biased behavior in summarization systems. We then consider the question of how to operationalize them to quantify the presence of bias in summaries and how to ensure measurements are not confounded by biases already present in the input documents.

1.3 Contributions

The core contributions of this thesis are:

- A comparison of different methodologies for human evaluation in text summarization, as well as a set of best practices for their design and analysis
- A set of tools for the meta-evaluation of automatic summarization metrics
- A thorough meta-evaluation and an in-depth analysis of summary coherence measures, establishing their shortcomings and identifying pathways for their improvement
- The development of a lightweight method for summary faithfulness evaluation
- The development of definitions for biased behavior in summarization systems and the development of automatic metrics and approaches that allow us to identify instances of summarizer bias.

In summary, this thesis provides a comprehensive set of tools, as well as guidelines, to holistically assess the quality of summarization systems along multiple quality dimensions. Our contributions will help guide the development of summarization systems by providing reliable feedback on summarizer performance at comparatively low cost.

1.4 Thesis Overview

In the following Chapter 2, we will go into more detail on both the task of text summarization and its evaluation. We will then turn to the core contributions of this thesis, starting with an investigation of how to conduct cost-efficient and reliable human evaluation studies in Chapter 3. Starting from Chapter 4, we will consider the topic of *automatic* metrics. Here, we will first establish general methodology for *meta-evaluation*, i.e. the evaluation of automatic metrics themselves. We will then use these principles to investigate the performance of automatic metrics for summary coherence. In Chapter 5, we will develop a cost-efficient automatic metric for summary faithfulness based on augmenting a pretrained NLI model. Finally, we will investigate the question of bias in text summarization in Chapter 6. We will develop definitions for biases in text summarization, as well as practical tools to measure them. Finally, we will conclude the thesis in Chapter 7 with an outlook on summarization evaluation in the face of improving model capabilities.

1.5 Published Work

This thesis is largely based on works that have been previously published. Our work on human evaluation in Chapter 3 has been described in Steen and Markert (2021). Our meta-evaluation techniques and meta-evaluation of summary coherence measures in Chapter 4 have been published in Steen and Markert (2022). Chapter 5 is based on Steen et al. (2023), with extensions mostly focused on a more thorough meta-evaluation and analysis of our proposed metric. Finally, the work on bias in summarization in Chapter 6 has been published as Steen and Markert (2024).

1.6 Published Code

All code underlying the work in this thesis is available online. Table 1.1 gives an overview of the code repositories corresponding to the chapters of this thesis.

Chapter	Link
Chapter 3	https://github.com/julmaxi/summary_lq_analysis
Chapter 4	https://github.com/julmaxi/summary_coherence_evaluation
Chapter 5	https://github.com/julmaxi/with_a_little_push
Chapter 6	https://github.com/julmaxi/summary_bias

TABLE 1.1: Links to repositories containing the code underlying the work in this thesis.

Chapter 2

An Overview of Evaluation in Text Summarization

In this chapter, we will give a brief overview of the field of text summarization and its evaluation. The intent of this chapter is not to give a full history of the development of the field but rather to introduce the task and to help contextualize our work on evaluation. We will discuss the various variants of the text summarization task and the methods that have been proposed to tackle it. Since our focus is on evaluation, we will then pay particular attention to the quality dimensions and automatic metrics used in their evaluation.

2.1 Text Summarization: Tasks and Datasets

Jones (1999) defines text summarization as “a reductive transformation of source text to summary text through content reduction by selection and/or generalisation on what is important in the source” (Jones, 1999). This definition covers a large number of different notions of “transformation” and “importance”, all of which might be grouped under the umbrella of text summarization. Attempts have been made at a more formal definition: Peyrard and Eckle-Kohler (2017) propose that a good summary maximizes the information gain a user derives from it over the knowledge they had before reading it. However, we argue that, in practice, the easiest way to understand the task is by surveying the numerous datasets that have been created for developing and evaluating summarization systems over the years. In general, summarization corpora consist of individual instances, which

have either a single document or a cluster of multiple documents as inputs and one or more *reference* summaries as output.

Within this general design space, Jones (1999) identifies three sets of what she calls “context factors” which define a summarization task: *input factors*, *purpose factors*, and *output factors*. Input factors include the structure, the genre, and language of the input. Purpose factors describe the use of a summary. This includes the intended audience and the context of its use, i.e. what information the summary is intended to transport. Finally, output factors describe the format of the output, the output language, its style, and how much content it covers from the input.

Since the full set of factors considered by Jones is too detailed for a structured overview, we select a subset of categories to represent these factors:

For **input** factors, we report the length of the input and whether a corpus contains *single-document* summaries or *multi-document* summaries. This distinction is important, since multi-document summarization usually requires reasoning and aggregating information across multiple documents, whereas single-document summarization is, in some settings, solved well by relatively simple heuristics. For example, in single-document news summarization, a well-performing (extractive) summarizer can often be constructed by simply taking the first three sentences of the input (Over et al., 2007). However, we note that just because a dataset is a multi-document dataset, it does not necessarily require information aggregation. Wolhandler et al. (2022) note, for example, that in some multi-document summarization datasets almost all information is contained in a single document. Finally, we also list the genre of the inputs.

Purpose factors are much more difficult to categorize. A traditional distinction is that between *generic* and *focused* summarization, where the former implies a “general” summary and the latter implies a response to some stated information need. This includes summaries in response to a question (Kulkarni et al., 2020; Zhong et al., 2021) or a topic statement (Dang, 2005; Dang, 2006; NIST, 2007). It is important to note that a generic summary is not necessarily one without a specific purpose. Jones (1999) argues that even superficially generic summaries make implicit assumptions about their requirements. She gives the example of a paper abstract, which assumes that the reader is familiar with the field of the

paper. A similar argument can be made for many crawled datasets, like CNN/DM (Hermann et al., 2015). CNN/DM contains a list of “highlights” for each input article. While researchers are not privy to how these highlights are selected, it is reasonable to assume that there is a hidden query or communicative goal highlight authors had in mind when the summary was created.

Finally, with regard to the **output** factors, we consider the length of the generated summary. Here it is important to differentiate between a fixed-length budget and an empirical summary length. In a fixed budget setting, reference summaries are created with a known limit on the number of tokens, characters, or sentences in the summary. During evaluation, this allows the length of the automatic summarizer output to be limited as well, which avoids issues inherent in comparing summaries of different lengths, which we will discuss in Section 2.4. Where the process of summary generation is unknown, we can only observe the length of summaries empirically, which implies no strict length limitation.

The earliest datasets we will list here are those from the influential Document Understanding Conferences (DUC), which were held yearly from 2001 to 2007 and were run by the U.S. National Institute of Standards and Technology (NIST). DUC created several shared tasks for summarization that included new summarization datasets, which were later widely used. DUC also conducted human evaluation studies on the participating summarization systems. We list the datasets created for the three most recent conferences to allow for comparison to later developments.

Several trends become apparent from the resulting overview in Table 2.1:

1. Summarization research is overwhelmingly conducted on single-document generic news summarization, although novel datasets have introduced alternative genres beyond news.
2. There has been a shift away from small-scale, carefully curated datasets to large-scale, web-crawled datasets.

The latter has a particularly large effect on evaluation, as will also become apparent in this work. For example, the XSum (Narayan et al., 2018a) reference summaries, which are just the first sentence of news articles, are known to contain information not present in the rest of the document (Maynez et al., 2020). This in

Name	Size	Type	Genre	Doc. Len.	Task	Task Details	Sum. Len.
Arxiv	≈215,000	Single	Scientific Articles	4,938	Generic	Abstracts	220
BigPatent	1,341,362	Single	Patents	3,573	Generic	Patent Abstracts	116
BillSum	18,949/3,269	Single	Legislative Documents	1,666	Generic	Congressional Re-search Service	203
BookSum Full	405	Single	Fiction	112,885	Generic	Various Book Summarization Pages, Like Sparknotes, Cliffnotes, etc.	1,167
BookSum Chapter	12,630	Single	Fiction	5,102	Generic	See above	505
BookSum Paragraph	146,532	Single	Fiction	160	Generic	See above	41
CNN/DM	287,226/13,368/11,490	Single	News	781	Generic	Article Highlights from News Websites	56
DUC 2005	50 topics w/ avg. of 32 docs	Multi	News	23,486	Topic and granularity	Expert Summaries	250 (f)
DUC 2006	50 topics w/ 25 documents	Multi	News	18,235	Topic	Expert Summaries	250 (f)
DUC 2007	45 topics w/ 25 documents	Multi	News	14,220	Topic	Expert Summaries	250 (f)
Gigaword	≈4,000,000/2000	Single (Sentence)	News	31	Generic	Headlines	8
GovReport	19,466	Single	Government Reports	9,409	Generic	Expert Summaries	553
Multinews	44,972/5,622/5,622	Multi	News	2,103	Generic	Articles From a News Aggregation Service	264
Newsroom	995,041/326,954	Single	News	659	Generic	Metadata Embedded in News Article Websites	27
NYT	589,284/32,736/32,739	Single	News	549	Generic	Abstracts From the NYT Indexing Service	40
PubMed	≈133,000	Single	Scientific Articles	3,016	Generic	Abstracts	203
SamSUM	14,732/818/819	Single	Dialogue	121	Generic	Summaries Written by Linguists	23
XSum	204,045/11,332/11,334	Single	News	431	Generic	Introductory Sentence of the Article	23

TABLE 2.1: Dataset statistics and task descriptions for various summarization datasets. Lengths in tokens are rounded to the nearest integer. (f) indicates fixed length-constraints, all other values are empirical. Multiple values in the size column refer to the number of instances in the train/test or train/dev/test splits, respectively.

References for the datasets can be found in Appendix A.

turn leads to models trained on XSum hallucinating (i.e. introducing facts unsupported by the input) much more than models trained on, for example, CNN/DM (Hermann et al., 2015). We will discuss the topic of hallucinations in more detail in Chapter 5. The unreliability of reference summaries and the fact that all crawled corpora have only a single reference is also a challenge for traditional automatic metrics like ROUGE (Lin, 2004b), which were designed for multi-reference settings.

There are a number of other tasks that fall under the general umbrella of summarization, but which we do not consider in this thesis. We give a brief overview for completeness:

In **Update Summarization**, pioneered during DUC 2007 (NIST, 2007) and continued in TAC 2008 (Dang and Owczarzak, 2009a), systems must generate a summary from an initial set of documents and then generate an incremental update from a second set of provided documents published after the initial documents.

In **Timeline Summarization** (Chieu and Lee, 2004; Binh Tran et al., 2013; Tran et al., 2015; Martschat and Markert, 2018; Gholipour Ghalandari and Ifrim, 2020) systems are expected to identify important events in a large corpus of input documents and generate a structured list of date/summary pairs for each event. This necessitates specialized evaluation criteria that are temporally sensitive (Martschat and Markert, 2017).

Opinion Summarization concerns itself with summarizing user opinions expressed, for example, in reviews (Ganesan et al., 2010; Chu and Liu, 2019; Bražinskas et al., 2020), online discussions (Fabbri et al., 2021a), or tweets (Inouye and Kalita, 2011). We consider this task to be subtly different from the type of document summarization we primarily study in this thesis, since the objective of opinion summarization is generally not to identify the most *important* information in the input, but rather to give a good overview of the *distribution* of opinions in the text. This has, as we will discuss later in this thesis, an impact on what we expect for a summary to be *unbiased*.

Efforts have also been made to create **Multimodal Summarization** systems (Zhu et al., 2018; Verma et al., 2023), which consider image data along with the input text and may select images to accompany summaries.

Finally, **Cross-Lingual Summarization** (Ladhak et al., 2020; Takeshita et al., 2022; Fatima and Strube, 2023) is a special case where input and summary are not written in the same language. This requires systems to either use translation as a preprocessing step or to natively deal with multiple languages.

2.2 Text Summarization Systems

Text summarization systems can be coarsely classified into extractive and abstractive systems. Extractive summarizers build a summary by selecting sentences from the source documents, whereas abstractive summarizers attempt to generate a summary from scratch. Due to the difficulty of language generation before the advent of neural networks for language modelling, most early work followed an extractive paradigm.

2.2.1 Extractive Summarization

In very early work, Luhn (1958) proposes an algorithm for summarizing technical documents based on identifying sentences with words that are frequent in the source document. Edmundson (1969) uses a feature-based approach, combining several surface-based features to assess sentence importance. Before the advent of neural summarization systems, Nenkova and McKeown (2012) propose that these, and subsequent, works are broadly characterized by three distinct phases: content representation, sentence scoring, and sentence selection. However, in this overview, we will discuss the first two jointly, since for many – especially neural – systems, representation is an implicit step.

Many approaches to sentence scoring are based on the observation that summary-worthy information is often repeated frequently across the input document(s). Radev et al. (2000) propose the MEAD summarization system, which works by clustering sentences and using cluster centroids as candidates for inclusion in the summary. Mihalcea and Tarau (2004) propose TextRank, which constructs a sentence graph where individual sentences share edges weighted by their similarity. They then use a modified variant of the PageRank algorithm (Page et al., 1999) to identify sentences with high centrality in the graph. Erkan and Radev (2004)

combine a similar centrality feature with additional indicative features and a redundancy heuristic. Zheng and Lapata (2019) later revisit the idea of graph-based sentence scoring by augmenting it with contextualized word embeddings. Alternatively, importance can also be learned directly from reference summaries. The most successful approaches here are based on directly exploiting the strong representation capabilities of neural language models (Cheng and Lapata, 2016; Nallapati et al., 2016a; Narayan et al., 2018b; Liu and Lapata, 2019). A core challenge for these models is to derive an extractive training signal from abstractive references. Proposed solutions range from heuristically deriving sentence labels using their similarity to the reference (Cheng and Lapata, 2016; Nallapati et al., 2016a) to using reinforcement learning to directly optimize summary/reference similarity (Narayan et al., 2018b).

While some, especially neural, summarizers directly select the highest-scoring sentences for the output (Cheng and Lapata, 2016; Nallapati et al., 2016a), creating a summary from individual sentence scores typically requires summarizers to deal with a relevance-redundancy trade-off. That is, sentences should not only be individually relevant but also not be redundant among each other. The maximum marginal relevance criterion (MMR) (Goldstein and Carbonell, 1998) provides a greedy approach for this. For each candidate sentence, it computes both the relevance of the sentence as well as its similarity to previously selected sentences. A linear combination of both is used to score each candidate sentence and the highest scoring sentence is selected for summary inclusion up to some length maximum. Liu and Lapata (2019) propose a simplified variant of this, where sentences with trigram overlap with previously selected sentences are discarded during inference. While this greedy approach has the advantage of efficiency, it is easy to construct counter examples where it does not yield an optimal selection. McDonald (2007) studies global optimization for summarization. He proposes to formulate the task as an integer linear program (ILP). ILPs allow summarization to be formulated as a constrained optimization problem, for which efficient solvers exist.¹ McDonald (2007) uses an MMR-inspired objective as the optimization target. ILP is also flexible enough to integrate additional terms into the objective. For example,

¹While solving an ILP is an NP-complete problem, in practice solvers exist which solve most problems in acceptable time (McDonald, 2007).

Parveen et al. (2015) add a coherence score to the objective function.

Alternatively, one can formulate the summarization objective in a way that allows it to be solved with a greedy algorithm with a theoretically grounded guarantee on the gap between the optimization results and the global optimum. Lin and Bilmes (2011) propose to express the summarization objective as a submodular function that allows them to give an upper bound on the optimality gap.

Finally, Zhong et al. (2020) propose to learn a neural scoring function over entire *summaries*, instead of individual sentences. The summarization system can then directly select the highest-scoring subset of sentences from the input as the summary. However, this must be combined with an approximate search procedure to avoid a combinatorial explosion over the possible subsets of sentences.

2.2.2 Abstractive Summarization

Early work on neural abstractive summarization initially focused on sentence summarization (Rush et al., 2015), due to the much smaller computational requirements and availability of sufficiently large training datasets. With the adoption of the CNN/DM dataset (Hermann et al., 2015; Nallapati et al., 2016b) for summarization, early neural single-document summarization systems became practical. Nallapati et al. (2016b) and See et al. (2017) both propose to enhance encoder/decoder recurrent neural networks (Bahdanau et al., 2015) with pointer mechanisms to allow models to explicitly copy parts of the input into the summary. This would often lead to highly extractive summaries, even though the models themselves were capable of fully abstractive generation (See et al., 2017).

Just as with many other tasks in NLP, the advent of large – in comparison to their predecessors – pretrained transformer language models has had a significant impact on the summarization landscape. Liu and Lapata (2019) fine-tune the encoder-only BERT (Devlin et al., 2019) for both extractive and abstractive summarization. Lewis et al. (2020) and Zhang et al. (2020a) both propose encoder-decoder transformer architectures for summarization. They use unsupervised pretraining tasks like gap infilling to create strong base models, which they then fine-tune on summarization datasets.

Finally, instruction-tuned large language models have shown remarkable zero-shot² summarization capabilities. Goyal et al. (2022) show that with simple prompting, GPT 3.5 can generate summaries that human raters prefer over summaries generated by purpose-built models. Adams et al. (2023) show that this can be further improved with specialized prompting methods that create summaries at different levels of detail.

2.3 Human Evaluation of Text Summarization

We now turn to the core concern of this thesis: The evaluation of summarization systems. The first important distinction we need to make in this regard is that between *extrinsic* and *intrinsic* evaluation.

2.3.1 Extrinsic Evaluation

Extrinsic evaluation focuses on summarization as an auxiliary step that should improve outcomes in some downstream task of interest. The TIPSTER SUMMAC evaluation (Mani et al., 1999), for example, considered two tasks which were, at the time of the study, routinely conducted by U.S. information analysts: A categorization task, where annotators were asked to assess the relevance of a given document for a topic, and a question answering task, where the document would serve as the source for writing a report on a given topic. To extrinsically evaluate a summarization system under this and similar settings, a study can compare accuracy and speed of annotators working with original documents with the accuracy and speed of annotators working with their summaries. A good summarization system should produce summaries that maintain task performance, while reducing human task completion times.

While extrinsic evaluation is attractive due to its ability to directly assess the utility of a system, there are complications that make it less suitable as a general evaluation framework. Firstly, extrinsic evaluation requires human workers trained

²In the sense that they are not explicitly trained on the task and do not require in-context examples. For non-open-source models it is unclear whether summarization demonstrations are a part of fine-tuning.

at some task of interest to devote time to conducting the evaluation study, which is expensive and often not feasible. Secondly, it requires a clear definition of a relevant task, which is not always natural. For example, with many summarization datasets such as XSum and CNN/DM, gold summaries are not designed to help users fulfill a well-defined task. Finally, a given set of tasks might not cover the full breadth of potential use cases.

This has led to only limited work on extrinsic evaluation. For the current generation of summarization systems, we are only aware of one extrinsic evaluation effort, that of Pu et al. (2024). They conduct a study on question answering, similarity judgement, and categorization on a set of recent summarization systems.

2.3.2 Intrinsic Evaluation

In **intrinsic evaluation**, summary quality is measured directly along some predefined quality dimensions. This requires researchers to carefully select appropriate evaluation dimensions, which are likely to predict the actual utility of the summary. Traditionally, these can be divided into two categories:

Content Dimensions consider how effectively the summary communicates relevant content to the reader.

Linguistic Quality Dimensions consider how well-formed the generated summaries are, independently of their content.

Content

Summary content can be measured either by overlap with a human reference summary (*reference-based* evaluation), or directly scored by human annotators (*reference-free*³ evaluation). Reference-free evaluation exists in a number of different variants: In query-based summarization, *responsiveness* measures how well the summary responds to the given question (Over and Yen, 2003; Over and Yen,

³The term *reference-free* is slightly misleading, since even reference-free evaluation can sometimes contain human-written references as an anchor or as a point of comparison. We consider an evaluation reference-free if it does not involve directly determining the similarity to a reference.

2004). In a similar vein, *usefulness* measures the utility of a summary for a hypothetical downstream task (Over and Yen, 2003). For generic summarization, this overall score is often referred to as informativeness or relevance (Grusky et al., 2018; Fabbri et al., 2021b).⁴ Clark et al. (2023) use a binary judgement they call “main points” that asks annotators whether they think the summary contains the most important information from the input. The most extreme variant here is to simply ask for user preference between a set of given summaries, without specifying any comparison dimension as done by Goyal et al. (2022).

For reference-based evaluation, the goal is usually to evaluate *coverage* of the information in the reference summary/summaries by the content in the system summary. Here, design decisions for human annotation usually focus on how to decompose summaries into individual content units to facilitate easy comparison and how to judge coverage of these units. An early effort here is the Summarization Evaluation Environment (SEE) (Lin, 2001). SEE splits both the reference and the system summary into elementary discourse units (EDUs) as a basic unit of comparison. Annotators then mark all system summary units that have some degree of overlap with the reference and finally give a percentage estimate of the overall overlap between system and reference summary. During application for the DUC 2001-2004 shared tasks, this approach proved to be difficult for annotators, leading to unstable system rankings (Lin and Hovy, 2002). As Over et al. (2007) note, comparison would also be limited to a single reference/summary pair, which is not ideal given the wide range of permissible summaries for any given topic. To make coverage annotation more reliable, Halteren and Teufel (2003) propose to use *factoids* as a finer-grained content representation than EDUs. Unlike EDUs, which are derived automatically for each summary in isolation, manual factoid annotation considers a set of summaries of the same input and identifies *atomic* factoids. A factoid is atomic if it only consists of information that always appears together in the given summaries. Overlap can then be computed by counting the number of factoids that are in both the reference and system summary. Nenkova and Passonneau (2004) propose a very similar system called Pyramid, which works

⁴These terms are, however, not always used synonymously. For example, Grusky et al. (2018) use both informativeness and relevance, where the former is meant to evaluate coverage of key points in the input and the latter consistency of details between source and summary.

with Summary Content Units (SCUs). SCUs are created in a process that is very similar to the factoid annotation. The namesake idea of Pyramid is that SCUs can be assigned importance scores by giving higher weight to SCUs that appear in multiple references. This can be conceptualized as SCUs forming a pyramid, where SCUs that appear across all reference summaries form the peak and less frequent and thus less relevant SCUs are at the bottom. A summary attains a high Pyramid score by covering as much of the upper part of the pyramid as possible. Later work in manual coverage annotation has largely followed the same ideas but focused on two areas:

1. Reducing annotator workload
2. Making the task suitable for non-expert annotators recruited via crowdsourcing platforms

Shapira et al. (2019) propose a light-weight variant of Pyramid called LitePyramid that removes the need for exhaustively collecting SCUs across multiple summaries. Instead, they ask annotators to write a fixed number of SCU-like statements per reference summary and then have a second set of annotators judge whether a random sample of these SCUs appears in a given system summary. Zhang and Bansal (2021) propose to further reduce workload by automatically identifying the presence of SCUs in a summary using a natural language inference (NLI) model. They also propose to automatically generate a subset of “easy” SCUs, leaving human annotators to create SCUs for more challenging parts of the summaries.

Liu et al. (2023b) propose a similar setup to LitePyramid, although their *atomic* content units (ACUs) are closer to factoids. Their annotation relies on a hybrid approach, where ACUs are created by experts and crowd workers conduct the matching procedure.

As we will show in a survey in Chapter 3, human reference-based evaluation has lost popularity in recent years compared to reference-free evaluation. We identify a number of contributing factors for this:

1. Reference-based evaluation is often expensive. Most annotation procedures require some kind of expert annotation and the process is often lengthy. For

example, Halteren and Teufel (2003) note that they require about 30 minutes per summary.

2. The number of reference summaries is limited to one in most benchmark datasets, as discussed in Section 2.1. This limits the utility of approaches like Pyramid, which rely on multiple reference summaries.
3. With the advent of large language models, the ad-hoc reference summaries in popular corpora like XSum and CNN/DM have been shown to be inferior to machine output (Goyal et al., 2022). This makes reference similarity unsuitable as an evaluation metric.

Another recent development in the evaluation of summary content is the advent of *faithfulness* as an important evaluation criterion (Maynez et al., 2020). A summary is faithful if all of its content is grounded in the input document(s). Unfaithful summaries introduce additional information into the summary, which may or may not be factual in relation to the real world. This has become particularly problematic with the advent of abstractive summarization systems. While extractive systems *can* introduce factual errors into a summary, for example due to referential errors (Zhang et al., 2023b), this is much more likely with abstractive generations. We defer a more thorough discussion of this phenomenon and related work to Chapter 5, where we will develop an automatic evaluation metric for faithfulness.

Linguistic Quality

Linguistic quality encompasses a wide range of potential quality dimensions. This is nicely exemplified by the progression of the DUC conferences and their evaluation of linguistic quality. The initial DUC conference in 2001 (Over, 2001) asked annotators to rate summaries from grammaticality, summary organization, and summary cohesion on a five point scale. In the DUC 2002 (Over and Liggett, 2002) and DUC 2003 (Over and Yen, 2003) iterations, this changed to a specific set of twelve error types: capitalization errors, word order, subject-verb agreement, missing subjects/verbs/objects, unrelated fragments, missing articles, incorrect use of pronouns, referential clarity for nouns, improper use of nouns instead of pronouns,

dangling conjunctions, unnecessary repetition, and lack of cohesion and coherence. For DUC 2004, this was again changed to coherence, conciseness, repetition, information status, and improper use of nouns instead of pronouns. Finally, for the remaining three DUC conference 2005, 2006, and 2007, evaluation included five dimensions: grammaticality, non-redundancy, referential clarity, focus and structure, and coherence; all rated on a 1-5 Likert scale (Dang, 2005; Dang, 2006; NIST, 2007).

With the shift to development of summarization systems outside of shared tasks, we observe less standardization in quality dimensions. Notable examples of larger-scale evaluation campaigns that include linguistic quality are Grusky et al. (2018) and Fabbri et al. (2021b), who use fluency and coherence for their respective studies. Clark et al. (2023) ask annotators to label a summary as comprehensible, repetition-free, grammatical, and concise. Zhang et al. (2024) exclusively use coherence as the only linguistic quality dimension in their comparison of human and LLM written summaries, reflecting the consistently high grammatical quality of LLM outputs. Following this, we are going to focus on coherence evaluation in Chapters 3 and 4.

2.4 Automatic Evaluation Metrics

The evaluation procedures we have discussed thus far have in common that they are dependent on human labor.⁵ The resulting large cost and time requirements raise a need for automatic evaluation metrics. Analogously to human evaluation, these can again be differentiated into *reference-based* and *reference-free* evaluation metrics.

2.4.1 Reference-based Metrics

Reference-based metrics compute the overlap between one or more human-written reference summaries and a system summary. The most well-known of these metrics is ROUGE (Lin, 2004b). ROUGE is a token-overlap-based evaluation metric,

⁵We note that for some of the automatic metrics here, human labor is still required, for example to create references. However, unlike human evaluation, this effort does not need to be repeated for every new set of summaries.

similar to BLEU (Papineni et al., 2002), which is commonly used in machine translation. Unlike its close relative, which measures precision and includes a brevity penalty to punish overly long sequences, ROUGE is a recall-oriented metric.

ROUGE exists in several variants, the most commonly used being ROUGE-N, which computes n-gram overlap between the references and the candidate summary. Let $\text{cnt}(w, T)$ be the number of occurrences of an n-gram w in a text T and let $\text{ngrams}(n, T)$ be the set of distinct n-grams of length n in T . Given a set of reference summaries \mathcal{R} and a generated summary S , ROUGE-N recall is defined as follows:

$$\text{ROUGE-n-rec}(\mathcal{R}, S) = \frac{\sum_{R \in \mathcal{R}} \sum_{w \in \text{ngrams}(n, R)} \min(\text{cnt}(w, R), \text{cnt}(w, S))}{\sum_{R \in \mathcal{R}} \sum_{w \in \text{ngrams}(n, R)} \text{cnt}(w, R)}. \quad (2.1)$$

Recall is an obvious choice for settings with strict length constraints but is problematic when the length is unconstrained during generation. Research on corpora without length constraints thus uses F1 score instead (Nallapati et al., 2016b; See et al., 2017):

$$\text{ROUGE-n-prec}(\mathcal{R}, S) = \frac{\sum_{R \in \mathcal{R}} \sum_{w \in \text{ngrams}(n, R)} \min(\text{cnt}(w, R), \text{cnt}(w, S))}{|\mathcal{R}| \sum_{w \in \text{ngrams}(n, S)} \text{cnt}(w, S)}, \quad (2.2)$$

$$\text{ROUGE-n-F1}(\mathcal{R}, S) = 2 \frac{\text{ROUGE-n-prec}(\mathcal{R}, S) \cdot \text{ROUGE-n-rec}(\mathcal{R}, S)}{\text{ROUGE-n-prec}(\mathcal{R}, S) + \text{ROUGE-n-rec}(\mathcal{R}, S)}. \quad (2.3)$$

This allows the comparison of summaries with different lengths, although Sun et al. (2019a) note that the resulting metric is sensitive to length variation in generated summaries.

Other, less frequently used variants of ROUGE include ROUGE-L, which computes the length of the longest common subsequence between each reference and the summary,⁶ ROUGE-W, which is a weighted variant of ROUGE-L, and ROUGE-S/ROUGE-SU, which are based on skip-gram statistics.

While ROUGE is by far the most popular evaluation method in text summarization, a number of attempts have been made to improve upon its formula with more sophisticated methods for computing overlap: Ng and Abrecht (2015)

⁶More precisely, the longest common subsequences are computed between pairs of sentences and merged.

use word-embeddings to allow for more flexible matching between n-grams; Zhang et al. (2020b) and Zhao et al. (2019) use the similarity of contextualized embeddings to compute a matching between reference and generated tokens; Clark et al. (2019) use both word and sentence embedding similarity to measure overlap. The common theme of these approaches is to allow for a more flexible matching between reference and generated summary by exploiting embedding-based similarity measures. This leads to measures that can more easily deal with paraphrasing and synonymy.

A number of non-matching-based approaches have also been proposed over time:

Yuan et al. (2021) propose to compute the probability of the generated summary under a conditional language model using the reference as input.

Chen and Eger (2023) propose to use natural language inference models, which predict whether a given hypothesis is entailed by, contradicted by, or logically neutral to a given premise. For reference-based evaluation, either the generation or the reference can be chosen as the hypothesis, with the other text forming the premise. A score can be derived by considering the entailment probability, possibly combined with the contradiction probability.

Gao et al. (2019) propose an automated variant of the Pyramid method discussed in Section 2.3.2 by automatically decomposing sentences into content units before clustering and then matching them. The semi-automated Pyramid method of Zhang and Bansal (2021), described in Section 2.3.2, can also be run in a fully automatic setting following a similar paradigm. Nawrath et al. (2024) systematically investigate how to best generate SCUs automatically for this purpose and find both generating them using LLMs and from abstract meaning representation (Banarescu et al., 2013) to be competitive.

Deutsch et al. (2021a) propose to use a question generation model to generate questions from a reference and then use an automatic question answering model to derive answers from the generated summary. A summary that is similar to the reference should generate the same answers for the questions. Conceptually, question generation takes the role of deriving content units and generating answers replaces the matching step.

2.4.2 Reference-free Metrics

Reference-free evaluation metrics compute summary quality directly from the summary and, possibly, the input. This avoids the need for reference summaries, which may be costly to create or of poor quality. We can broadly categorize approaches here into *unsupervised* and *supervised approaches*.⁷ We refer to any method that explicitly uses human evaluation data for training as *supervised*, whereas we consider all other methods *unsupervised*, even if they are trained on labeled data from other sources.

Unsupervised approaches encompass heuristics and methods learned on external tasks. Zhu and Bhat (2020) propose a set of heuristics to measure the DUC 05-07 linguistic quality dimensions (except referential-clarity). For content evaluation, Vasilyev et al. (2020) propose to predict the source document using a language model, providing the summary as an auxiliary input. The intuition here is that a good summary should lead to a large drop in perplexity when modelling the input. Gao et al. (2020) propose to construct pseudo-references which they then compare using Sentence-BERT (Reimers and Gurevych, 2019) embedding similarity. Darin et al. (2024) propose to measure the mutual information between summary and source. They justify this by showing that this is highly predictive of downstream task performance when replacing the full input with the summary.

More recently, LLM-as-a-judge approaches have become popular, both for summarization specifically and NLG in general (Chiang and Lee, 2023; Liu et al., 2023a; Shen et al., 2023). Here, a LLM is prompted to output a score for a quality dimension, often together with a short explanation of the reasoning for the score. While this shows promising results, the black-box nature of these scores makes them susceptible to hard-to-detect biases, such as a preference for generations made by the same LLM that is used as a judge (Panickssery et al., 2024; Koo et al., 2023), or confounding quality dimensions with text length (Koo et al., 2024).

⁷In contrast to reference-based metrics, which are all unsupervised. It is also possible to design supervised reference-based metrics. Such supervised reference-based metrics exist in machine translation (Yan et al., 2023). However, we are not aware of such approaches in summarization evaluation.

Additionally, there is a large body of work on measuring text coherence, which can be directly applied to evaluating summary coherence in linguistic quality evaluation. We defer an in-depth discussion of this to Chapter 4, where we will conduct a meta-evaluation of their capabilities.

Supervised metrics rely on human annotations to learn to score summary quality dimensions. While this is common in machine translation, where large datasets of human annotations are available (Rei et al., 2020), for text summarization, such data is scarce (Clark et al., 2023). Pitler et al. (2010) train a feature-based linguistic quality model on DUC 2006 and DUC 2007 annotations. In a similar vein, Xenouelas et al. (2019) use data from the DUC 05-07 conferences to train a BERT-based regressor capable of predicting human evaluation scores. However, their training data only contains pre-neural summarization systems, which can limit generalizability. For coherence, this will become apparent in our meta-evaluation in Chapter 4. Clark et al. (2023) gather a large-scale dataset of human annotations for summaries in multiple languages and use it to train a mT5-based (Xue et al., 2021) model for several quality dimensions.

Finally, we have purposefully left out the extensive work on faithfulness metrics in this discussion. We will discuss them in depth in Chapter 5 instead, where we will directly relate them to our work in this area.

2.5 Discussion

In this chapter, we have given an overview of the task of text summarization and its evaluation. We have constrained this introduction to be broad, leaving more specific background to the individual parts where we tackle these aspects of summarization evaluation. Two aspects of this overview will be particularly relevant for the remainder of this thesis:

- Just like most areas of NLP, text summarization has undergone large changes in recent years. Recent datasets are typically large and not specifically designed with the evaluation of automatic text summarization in mind. Recent summarization systems are learned and abstractive, as opposed to their extractive, often heuristic progenitors.

- The design space of evaluation for text summarization is huge, with a large number of quality dimensions and approaches to consider.

Both of these points highlight the need for a new, holistic perspective on summarization evaluation. This is what we seek to provide in this thesis.

Chapter 3

Human Evaluation of Summarization Systems

3.1 Motivation

As discussed in Chapter 2, the utility of a summarizer is ultimately determined by how well it supports a user’s information need. However, such extrinsic evaluations are difficult to conduct and can only ever cover a narrow set of potential applications. Generically evaluating summarizers thus typically requires intrinsic human evaluation. To design such studies, researchers need to answer a number of difficult questions:

1. What kind of rating should annotators be asked to provide (numerical, ranking, etc.)?
2. How can results be interpreted to avoid erroneous conclusions?
3. How should annotators be distributed across samples to maximize study power?

Howcroft et al. (2020) note that for natural language generation (NLG) evaluation in general, there is very little agreement on how to design human evaluation studies and details are typically underreported. Lee et al. (2019) mirror a similar sentiment. This problem extends, as we will show in this chapter, also to the evaluation of summarization systems.

While, as noted in Chapter 2, there was an early standardization in the form of shared tasks during the Document Understanding Conferences (DUC) (Dang, 2005), this standardization has not survived into current practice. A likely cause for this is the shift from extractive to abstractive summarization, along with a change in available resources: Whereas the NIST-organized DUC shared tasks could rely on a set of expert “assessors”, the smaller scale of most current evaluation campaigns requires researchers to either rely on locally recruited annotators, or on crowd-working platforms like Amazon Mechanical Turk¹ or Prolific.² Practices designed to have a set of expert annotators evaluate primarily extractive summaries do not necessarily translate to crowd-sourced evaluation of abstractive summaries. This lack of standardization not only leads to little comparability between individual studies but also carries the risk of less than optimal decisions leading to cost-inefficient designs, and, in the worst case, erroneous conclusions about summarizer performance.

In this chapter, we thus tackle the problem of designing intrinsic evaluation studies that are both reliable and cost-efficient. We will first survey the (at the time of this work) current practices in human evaluation for text summarization. We will then conduct a set of human evaluation studies with the goal of establishing best practices for human evaluation of summarization systems. In particular, we are going to focus on the three questions posed in the beginning of this section:

1. To identify the best choice of **annotation method**, we are going to compare the efficiency of the two most popular methods in current literature: Likert- and ranking-style evaluation.
2. With regard to **statistical analysis**, we are going to discuss how to properly analyze the resulting data. We will show how the presence of *grouping factors* in typical human evaluation data can lead to erroneous conclusions when analyzed with inappropriate statistical tools. In summarization evaluation, grouping factors arise whenever one annotator rates multiple summaries and when multiple summaries are generated for the same document.

¹www.mturk.com

²prolific.com

3. Finally, we will give recommendations for **study design**, particularly regarding the overall number of annotators and their distribution across samples.

In summary, this chapter contains the following contributions:

1. We conduct a comprehensive survey on the current practices in manual summary evaluation in Section 3.3. Often, important study parameters, such as the total number of annotators, are not reported. In addition, statistical significance is either not assessed at all or with tests (t-test or one-way ANOVA) that lead to inflated Type I³ error in the presence of grouping factors (Barr et al., 2013).
2. We carry out annotation experiments for coherence and repetition. We elicit both Likert- and ranking-style annotations on the output of four recent summarizers and on reference summaries. We show that ranking-style evaluations are more reliable and cost-efficient for coherence, supporting prior findings by Novikova et al. (2018) and Sakaguchi and Van Durme (2018). However, for evaluation of repetition, where many documents do not exhibit any problems, Likert outperforms ranking.
3. Based on our annotation data, we perform Monte-Carlo simulations to show the risk posed by ignoring grouping factors in statistical analysis and find up to eight-fold increases in Type I error rate when using standard significance tests. As an alternative, we propose to either use mixed-effects models (Barr et al., 2013) for analysis, or to design studies in such a manner that results can be aggregated into independent samples, amenable to simpler tools.
4. Finally, we show that the common practice of eliciting repeated judgements for the same summary leads to less reliable and powerful studies for system-level comparison when compared to studies with the same budget but only one judgement per summary.

The work presented in this chapter has previously been published as

³I.e. the probability of incorrectly rejecting the null hypothesis. We will discuss this concept more formally in Section 3.2.4.

Julius Steen and Katja Markert (2021). “How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo et al. Online: Association for Computational Linguistics, pp. 1861–1875. DOI: 10.18653/v1/2021.eacl-main.160. URL: <https://aclanthology.org/2021.eacl-main.160>.

3.2 Background and Related Work

Human evaluation in NLG in general and summarization in particular has a long history. We have already touched on the history of reference-based human evaluation in Chapter 2. Here, we are going to focus specifically on the design of reference-free evaluation studies. We split this discussion into four parts. First, we are going to discuss what it means for a human study to be reliable. We are then going to discuss how the use of crowd work, which is common in human evaluation studies, influences study design. We will then turn to which methods are typically used to elicit human judgements. Finally, we will give a brief introduction to null hypothesis significance testing, which is the most common framework for analyzing human evaluation results.

3.2.1 Reliability as a Proxy for Validity

When we conduct a human evaluation study, what we are ultimately interested in is that the judgements accurately reflect the true underlying quality of summarizer generations. This is also called the *validity* of the study. Validity, however, is difficult to establish in human evaluation since the unknown performance characteristics of a summarizer are what we seek to identify with our study. We can usually only qualitatively assess whether the questions, methods, and design in a study “make sense” for the quantity of interest. This is analogous to the concept of “face validity” in psychological research (Price et al., 2015).

In NLP, it is common practice to instead measure the *reliability* of a measure. A measure is reliable if it produces the same conclusions under different settings.

For example, in the context of human evaluation studies, exchanging annotators or documents with samples from the same population should not alter conclusions about system performance. While reliability does not imply validity, a lack of reliability is an indicator for an invalid measurement.

Agreement

Reliability of human annotations for human evaluation is often assessed using agreement measures, as is common practice in annotation experiments in computational linguistics (Carletta, 1996; Artstein and Poesio, 2008). Agreement measures measure the *inter-rater reliability*, i.e. the extent to which different annotators agree on the assessment of a single instance during evaluation.

The most straightforward measure of inter-rater reliability for two annotators is the *observed agreement*. Let C be the set of possible rating categories and n be the number of items in the study (i.e. summaries in our specific case). Let \hat{R} be the observation matrix where $\hat{R}_{i,j}$ is the score assigned by annotator j to item i . The *observed agreement*, i.e. the empirical probability that the two annotators agree computed on the given ratings, is

$$P_o = \frac{1}{n} \sum_{i=1}^n \mathbb{1} \left(\hat{R}_{i,1} = \hat{R}_{i,2} \right), \quad (3.1)$$

where $\mathbb{1}$ is the indicator function.

However, observed agreement has an important flaw in that agreement can also arise due to chance. This observation gives rise to so-called *chance-adjusted agreement measures*. Chance-adjusted agreement measures can generally be formulated as

$$\frac{P_o - P_e}{1 - P_e}, \quad (3.2)$$

where P_e is the expected agreement given that annotators randomly assign ratings.

For two annotator scenarios, Cohen’s κ (Cohen, 1960) is a popular agreement measure. It computes P_e from the marginal probabilities over C of the annotators:

$$P_e = \frac{1}{n^2} \sum_{c \in C} \text{cnt}(c, \hat{R}_{\bullet,1}) \cdot \text{cnt}(c, \hat{R}_{\bullet,2}), \quad (3.3)$$

where $\text{cnt}(c, \hat{R}_{\bullet,j})$ is the number of times category c is observed in the annotations of annotator j .

Scott's π is a similar agreement measure but instead computes the chance agreement from the joint probabilities of the two annotators:

$$p_c = \frac{\text{cnt}(c, \hat{R}_{\bullet,1}) + \text{cnt}(c, \hat{R}_{\bullet,2})}{2n} \quad \forall c \in C, \quad (3.4)$$

$$P_e = \sum_{c \in C} p_c^2. \quad (3.5)$$

For a set of multiple annotators, Fleiss' K is commonly used as a generalization of Scott's π . Here P_o and P_e are computed as follows:

$$P_o = \frac{1}{na(a-1)} \left[\left(\sum_{i=1}^n \sum_{c \in C} \text{cnt}(c, \hat{R}_{i,\bullet})^2 \right) - na \right], \quad (3.6)$$

$$P_e = \sum_{c \in C} \left(\frac{1}{na} \sum_{i=1}^n \text{cnt}(c, \hat{R}_{i,\bullet}) \right)^2, \quad (3.7)$$

where $\text{cnt}(c, \hat{R}_{i,\bullet})$ is the number of times category c is observed in the annotations of item i and a is the number of annotators.

Krippendorff's α (Krippendorff, 1970) can similarly be used for assessing agreement and has the additional advantage that it can be parameterized by a function δ that specifies the dissimilarity of two categories. This is particularly attractive for ratings, since they are typically on an ordinal scale, where a disagreement of 1 vs. 2 is intuitively much less severe than a disagreement of 1 vs. 5. Unlike Fleiss' K and Cohen's κ , it is expressed in terms of *disagreement*, although the core idea of comparing expected and observed disagreement remains the same:

$$\alpha = 1 - \frac{D_o}{D_e}. \quad (3.8)$$

Unlike the previously discussed coefficients, Krippendorff's α accounts for incomplete rating matrices, where some annotators only annotate a subset of the items. We thus introduce \tilde{R}_i as the multi-set of ratings for item i , i.e. the set of non-empty ratings in $\hat{R}_{i,\bullet}$. Krippendorff's α for a given \tilde{R} and δ can then be

computed with

$$D_e = \frac{1}{(M-1)M} \sum_{c_1 \in C} \sum_{c_2 \in C} \delta(c_1, c_2) \begin{cases} \text{cnt}(c_1, \hat{R})(\text{cnt}(c_1, \hat{R}) - 1) & \text{if } c_1 = c_2 \\ \text{cnt}(c_1, \hat{R}) \cdot \text{cnt}(c_2, \hat{R}) & \text{else} \end{cases}, \quad (3.9)$$

$$D_o = \frac{1}{M} \sum_{c_1 \in C} \sum_{c_2 \in C} \delta(c_1, c_2) \sum_{i=1}^n |\tilde{R}_i| \frac{\text{cnt}(c_1, c_2, \tilde{R}_i)}{(|\tilde{R}_i| - 1)|\tilde{R}_i|}, \quad (3.10)$$

$$M = \sum_{i=1}^n |\tilde{R}_i|, \quad (3.11)$$

where $\text{cnt}(c_1, c_2, \tilde{R}_i)$ is the number of pairs of ratings with values c_1, c_2 in \tilde{R}_i . $\text{cnt}(c, \hat{R})$ is the number of times category c appears in all ratings. An intuitive conceptualization of Krippendorff's alpha is to consider D_o as the average disagreement within each multi-set \tilde{R}_i weighted by the number of annotations in \tilde{R}_i and D_e as the disagreement within the multi-set of all ratings across all items (Honour, 2016).

For ordinal rating data, such as the responses on a Likert-scale, an appropriate δ is (Krippendorff, 2011)

$$\delta(c_1, c_2) = \sum_{c=c_1}^{c_2} \left(\text{cnt}(c, \hat{R}) - \frac{\text{cnt}(c_1, \hat{R}) - \text{cnt}(c_2, \hat{R})}{2} \right)^2. \quad (3.12)$$

Problems with Agreement

Agreement measures have two important flaws when applied to evaluation data. Firstly, their interpretation is generally difficult. While standardized scales exist (Landis and Koch, 1977), they may give conflicting advice. Especially in NLG evaluation, many studies have Krippendorff's α that is below the cutoff recommended by Krippendorff (1970) of 0.8 below which data should be considered unreliable (Amidei et al., 2018).

This leads to the second issue: Agreement is not necessarily expected in human ratings. Amidei et al. (2018) study the example of a question generation task where annotators were asked to rate the grammaticality and idiomaticity of a generation, as well as the appropriateness of the generated question for the input. In their

experiments, Amidei et al. find that – in spite of repeated refinement of annotation guidelines – agreement ultimately remains low. In a qualitative analysis, they find that this can often be explained by personal preferences and prior knowledge. In these cases, low agreement does not indicate a fault in the evaluation setup, but rather the diversity of human perception of question quality. Amidei et al. (2018) ultimately propose to improve reporting standards around agreement with confidence intervals and to designate boundaries within which annotator agreement should fall: Too low and data is insufficiently reliable, too high and the task might be insufficiently interesting.

In this chapter, we will propose another approach. While we will report Krippendorff’s α to allow for comparisons with other studies, we propose that measuring *consistency* is much more informative for human evaluation studies.

Consistency

The problems with agreement outlined above suggest that inter-rater reliability might not always be an informative quantity in evaluation studies. Another variant of reliability is to test for *consistency*. Internal consistency is a concept that originates from test theory and refers to the extent to which the different questions in a survey measure the same underlying latent concept in a subject (Tavakol and Dennick, 2011). Internal consistency can be measured with Split-Half Reliability (SHR). SHR splits a survey into two parts and computes the correlation of the resulting measures between both halves. In NLP, SHR has been previously used by Kiritchenko and Mohammad (2017) to compare the reliability of different annotation methods in sentiment intensity annotation.

For an evaluation study, the “subjects” are the systems, the latent concept we want to measure is the performance in human judgements, and the equivalent to “questions” in a survey is the evaluation on different documents and by different annotators. Care needs to be taken, however, to ensure the halves are independent. We will discuss the particularities of computing SHR for evaluation studies in Section 3.5.1.

3.2.2 Human Evaluation with Crowd Workers

The most critical ingredient for any human evaluation are, naturally, the human annotators themselves. Since we are going to recruit annotators from crowd-working platforms in this chapter, we will now discuss the particularities and trade-offs of choosing crowd workers over alternatives like locally recruited participants.

Crowd-working platforms like Amazon Mechanical Turk (MTurk) or Prolific allow researchers to create so-called “human intelligence tasks” (HITs).⁴ HITs consist of a task that should be completed by one or more human workers, as well as a compensation to be paid upon completion. Researchers can require additional qualifications like a location and a history of successfully completed HITs.

The simplicity of acquiring crowd workers has made this method of recruitment popular in human evaluation (Gehrmann et al., 2023), with some authors particularly highlighting the cost efficiency of the method (Callison-Burch, 2009). Additionally, if the demographic diversity of the annotator pool is relevant, which might be the case for subjective judgements, the pool of crowd workers is slightly more demographically diverse than typical alternatives, like working with college students (Buhrmester et al., 2011).

A concern when using non-expert annotators in general and crowd workers in particular is data quality. While Callison-Burch (2009) and Graham et al. (2017) report that crowd workers can be a replacement for experts in the evaluation of machine translation systems, most studies in text summarization find that agreement between crowd workers is too poor to lead to reliable summary-level ratings. Gillick and Liu (2010) compare crowd-sourced annotations to expert annotations from the Text Analysis Conference (TAC) 2009 shared summarization task (Dang and Owczarzak, 2009b). They find that crowd workers have a higher variance and are unreliable on the *summary* level, although their judgements are reliable on the *system* level. Fabbri et al. (2021b) elicit both expert and crowd worker judgements and find that they do not correlate at all on the summary level. Iskender et al. (2021) also conduct a comparison of expert and crowd worker judgements and

⁴The term is specific to MTurk, although the concept is the same in other platforms.

find that the typical number of three crowd workers per summary cannot provide reliable summary-level scores.

One source of variance in the annotation results is the presence of low-effort annotators in the worker pool, who essentially judge summaries at random. MTurk allows researchers to limit participation to workers with a certain number of HITs that have been accepted as completed by other HIT requesters, as well as an overall acceptance rate. Similar methods are typically available on other platforms. This can be used to limit the participation of potentially low-performing workers. For example, Fabbri et al. (2021b) use a limit of 10,000 HITs with a 97% approval rate for their experiments. To further mitigate risks to data quality, researchers often insert *attention checks* into their HITs. One common type of attention check, called instructional manipulation check (Oppenheimer et al., 2009), is to test whether annotators actually read instructions before submitting an answer to a HIT. For example, in a summary evaluation task, the summary might be replaced with an instruction to the annotator to assign a certain score. There are other variants of attention checks, such as requiring correct answers to predetermined distractor questions. If workers fail too many attention checks, their HITs will be discarded in further analysis and another response is elicited. While some platforms, like MTurk, allow study authors to withhold payment for HITs in these cases, other platforms are more restrictive due to the labour fairness implications of unpaid work. For example, the platform used in this chapter, Prolific, only allows rejection for at least two failed instructional manipulation checks.⁵ In these cases, rejections for other reasons lead to an increase in study cost.

An alternative approach is to recruit a set of annotators that have proven reliable in prior studies. Zhang et al. (2023a) propose a framework for identifying annotators using a series of qualifying tests. This allows them to build up a set of competent annotators which reduces noise in their data. However, this is mostly suitable for large or repeated evaluation studies due to the overhead in recruiting costs.

Another risk to data quality comes in the form of *sequence bias* (Mathur et al.,

⁵See <https://web.archive.org/web/20240107094551/https://researcher-help.Prolific.com/hc/en-gb/articles/360009223553-Prolific-s-Attention-and-Comprehension-Check-Policy>, accessed 14.05.2024, 19:57.

2017). Sequence bias arises when the distribution of annotator judgements differs depending on previously annotated instances. Mathur et al. conduct an analysis of multiple crowd-sourced tasks, including an analysis of adequacy judgements in machine translation. They find that annotators tend to assign higher scores to instances that follow other instances they had assigned a good score. The effect is particularly pronounced in the beginning of the annotation task. To avoid these biases effecting system scores, Mathur et al. recommend shuffling the order in which instances are presented.

Similar to the order of questions, Schoch et al. (2020) note that other subtle variations in the evaluation setup can have unexpected influence on the results. They particularly highlight framing effects that can exist in the question. They give the example of a pairwise comparison where annotators are asked how much *better* generation A is than generation B, which positively frames generation A. They advocate for a thorough reporting, as well as standardization of design parameters in human studies.

More recently (although not at the time when the study in this chapter was conducted), the availability of LLMs has begun to impact crowd-working platforms. Veselovsky et al. (2023) observe that in text production tasks, there is strong evidence that a large subset of workers use LLMs to produce their results. While it is unclear whether this translates to rating tasks, it remains a possible concern for future evaluation tasks, in particular with improvements in the integration of tool usage with LLMs.

Finally, the choice of crowd-working platform also influences data quality. Douglas et al. (2023) compare several platforms, including Amazon Mechanical Turk and Prolific, which we use in our study. They find workers on Prolific are generally more likely to pass attention checks and follow instructions, leading to less noisy data.

3.2.3 Methods for Judgement Elicitation

We now turn to surveying *how* judgements are elicited from human annotators. While the question may seem straightforward at first glance, a number of different

methods have been suggested in NLG evaluation. We differentiate them into three categories: error counting, direct assessment, and relative assessment.

In **error counting** approaches, annotators are asked to either specifically mark issues in the generation, or to give a count of the total number of such errors they identified. This was, for example, the method of choice during DUC 2002 (Over and Liggett, 2002).

In **direct assessment**, annotators are asked to rate the generation along a given rating scale. As we will show in Section 3.3, the most common choice in direct assessment in summarization evaluation is the five-point rating scale, often referred to as a *Likert* scale. However, alternative techniques have been proposed for NLG evaluation. Siddharthan and Katsos (2012) use magnitude estimation (Bard et al., 1996) for acceptability judgements, which allows annotators to assign an unbounded number of distinct positive ratings. To anchor the rater, a reference is provided together with a standard rating and annotators are asked to express the difference in quality as a ratio relative to the standard. For example, if the reference has a rating of 100 and the instance is twice as good as the input, it should receive a value of 200. Belz and Kow (2011) propose the use of visual-analogue scales, which allows the annotator to select the rating on a continuous scale.

Finally, **relative assessment** is based on comparison of individual generations. This can be achieved with either full (Callison-Burch et al., 2007) or partial ranking of results, where the extreme is a pairwise comparison between generations from all systems. For multiple systems, there is a trade-off in the number of annotations that need be elicited. To rate n systems on m outputs, ranking requires m full rankings, whereas pairwise comparison requires mn^2 pairwise annotations. A compromise between both is achieved by so-called best-worst scaling (BWS) (Louviere et al., 2015). In BWS, annotators are asked to select the best and worst sample out of a set of s items with $s < n$. By repeating the process with different permutations, a full ranking can be recovered. BWS has seen limited application in summarization evaluation (see our survey in Section 3.3), but has been shown to be more reliable than rating scales in word sentiment annotation (Kiritchenko and Mohammad, 2017), where the large number of items make a full ranking infeasible. For pairwise evaluation, the number of required annotations can be reduced

by selecting only a subset of comparisons for annotation: Sakaguchi et al. (2014) propose to use the TrueSkill (Herbrich et al., 2006) matchmaking-algorithm to prioritize comparing systems that have similar performance.

Novikova et al. (2018) introduce a hybrid approach, that combines magnitude estimation with ranking they dub *RankMe*. Here, annotators are presented with both a reference, as well as n sets of generations, for which they need to give magnitude estimates. Sakaguchi and Van Durme (2018) propose a similar hybrid approach where annotators give continuous direct assessments, but instances are presented in direct comparison and a variant of TrueSkill is used to decide which pairs are presented to annotators.

Finally, the process of annotation can be improved by combining human judgements with automatic metrics. Chaganty et al. (2018) propose the use of control variates, which allow to use an automatic metric to reduce the variance of human annotations and thus the number of annotations required.

3.2.4 Null Hypothesis Significance Testing

After eliciting judgements, researchers must interpret the results of a human annotation study. When evaluating a newly proposed summarization system, we are typically interested in establishing whether any improvements over other summarization systems are due to the randomness inherent in the annotations, or caused by actual improvements in the quality dimension.

In most experiments in NLP, this is evaluated using null hypothesis significance testing (NHST) (Dror et al., 2018; Sadeqi Azer et al., 2020). The goal of NHST is to establish whether a given observation provides sufficient evidence to support a given hypothesis. In the context of (human) evaluation for summarization, our empirical evidence is a set of acquired ratings \hat{R} for a set of summarizers S and we usually seek to answer the question of whether a summarizer $s_1 \in S$ has a higher expected human score than another candidate $s_2 \in S$.

To conduct a null hypothesis significance test, we first formulate the namesake null hypothesis. In our evaluation scenario, the null hypothesis states that s_1 and s_2 perform equally well in human evaluation. We then define a sample statistic δ , which maps our empirical observations (i.e. the elicited ratings) to a scalar value.

We can now compute a p -value as the probability of the test statistic being at least as extreme the statistic observed in our sample under the null hypothesis H_0 :

$$p_{H_0}(\delta(R) \geq \delta(\hat{R})), \quad (3.13)$$

where R is a random variable over possible observation sets.

Researchers must then define a *significance level* α at which the null hypothesis is rejected. A typical choice is $\alpha < 0.05$.

There are two error types that can occur in the application of this procedure: Type I and Type II errors. A Type I error occurs if we reject the null hypothesis when it actually holds. In evaluation experiments, this is typically the more concerning error, since it leads us to erroneously assume that a summarizer performs better than another when it does not. A Type II error describes the opposite situation: We accept the null hypothesis when it does not hold. In terms of summarizer comparison, this would lead to us not finding sufficient evidence to establish the superiority of one summarizer over another.

If an appropriate test is selected, whose assumptions are satisfied by our sample, the Type I error is exactly the significance level α . The inverse of the Type II error rate is also referred to as the *power* of a test.

One of the most well-known tests in NLP research is the paired student's t-test. It is a *parametric* test, i.e. it assumes samples are drawn from a given distribution, so the distribution of δ is known. The t-test assumes both samples are drawn from a normal distribution. A paired test is used to account for the dependence of samples derived on the same document.

An alternative test, that also finds wide application in NLP, is the approximate randomization test (ART) (Noreen, 1989). The approximate randomization test is non-parametric. Instead, it uses resampling to approximate the distribution of the test statistic under the null hypothesis. Given our set of ratings \hat{R} , the test creates a random sample \bar{R} from the paired samples in \hat{R} by randomly swapping the elements of each pair.

The significance level is then computed by repeatedly resampling \bar{R} and computing the probability that the statistic δ on \bar{R} is at least as extreme as that on

\hat{R} :

$$p(\delta(\bar{R}) \geq \delta(\hat{R})) = \frac{1 + \sum_{\bar{R} \in \mathcal{R}} \mathbb{1}(\delta(\bar{R}) \geq \delta(\hat{R}))}{|\mathcal{R}| + 1}, \quad (3.14)$$

where \mathcal{R} is a set of k permutations drawn from \hat{R} . The higher the number of iterations k , the more accurate our estimate of the p-value.

An important assumption that is shared by both this test and the t -test is that the pairs of samples in \hat{R} are independent.⁶ As we will discuss in the remainder in this chapter, this is usually not true for human annotation data. This results in inflated Type I error rates for these tests.

Study Power

A common way to determine test power is via simulation: If we know the underlying data distribution, we can sample artificial data with a known effect size – in our case, the difference in expected score between two systems – from the distribution. We can then repeatedly draw new samples and compute the number of times the test detects a significant difference on the samples.

Card et al. (2020) conduct such Monte-Carlo simulations and find that in NLP, many studies are *under-powered* for typical effect sizes, i.e. they have an insufficient probability to reject the null hypothesis, even if it is true. They note that this problematic since if a low-powered test *does* detect a significant difference between systems, it is likely to either exaggerate the effect or even invert it compared to the true effect.

Bayesian Methods

While NHST is the most popular analysis method in NLP and it is also what we use in this chapter, it is not the only option. Sadeqi Azer et al. (2020) note that p -values are often subject to subtle misinterpretations by researchers. They propose to replace the frequentist paradigm of NHST with a Bayesian approach. The equivalent to NHST in their approach is the Bayes factor, which instead quantifies how much the evidence provided by human annotations alters the credibility of

⁶More precisely, ART requires samples to only be *exchangeable*. However, for the purposes of this chapter, this distinction is not important.

two given hypotheses. While operating in a Bayesian framework would alter the methods used in the statistical analysis, the points we are going to make regarding the proper modelling of the underlying statistical dependencies remain valid under either perspective.

3.3 Literature Survey of Evaluation Practices

To establish which evaluation practices are common in summarization evaluation, we survey all summarization papers in ACL, EACL, NAACL, ConLL, EMNLP, TACL, and the *Computational Linguistics* journal in the years 2017-2019. We chose this timeframe as we were interested in current practices in summarization evaluation at the time of the original publication of this work: 2017 marks the publication of the pointer generator network (See et al., 2017), which has been highly influential for neural summarization. We focus our analysis on papers that present *a novel system* for single- or multi-document summarization and take a single or multiple full texts as input and also output text (SDS/MDS).⁷

Out of the resulting **105** SDS/MDS system papers, we identify all papers that conduct at least one new comparative system evaluation with human annotators for further analysis, leading to **58** papers in the survey. The fact that this is only about half of all papers is troubling given that it has been demonstrated that current automatic evaluation measures such as ROUGE (Lin, 2004b) are not good at predicting summary scores for modern systems (Schluter, 2017; Kryscinski et al., 2019; Peyrard, 2019).

We assess both *what* studies ask annotators to judge, as well as *how* they elicit and analyse judgements. Survey results are given in Table 3.1. Further details about the choices made in the survey, including category groupings/definitions and what is included under *Other*, can be found in Appendix B. As many papers conduct more than one human evaluation (for example on different corpora), we also list individual annotation studies (a total of 95).

⁷Excluded from the analysis are sentence summarization or headline generation papers, although most of the points we make hold for their evaluation campaigns as well. Summarization evaluation papers that do not present a new system but concentrate on sometimes large-scale system comparisons are also excluded.

	Category	Pa.	St.
Evaluation Questions	Overall	17	23
	Content	45	65
	Fluency	29	34
	Coherence	10	11
	Repetition	14	17
	Faithfulness	6	8
	Ref. Clarity	2	2
	Other	8	9
Evaluation Method	Likert	32	43
	Pairwise	10	14
	Rank	9	9
	BWS	6	9
	QA	9	14
	Binary	4	4
	Other	2	2
Number of Documents in Evaluation	< 20	6	10
	20-34	22	41
	35-49	3	4
	50-99	14	21
	100	11	14
	> 100	4	4
	<i>not given</i>	1	1
Number of Systems considered	< 3	13	20
	3	17	23
	4	16	23
	5	6	10
	> 5	12	19
	w/ Reference	16	25
	w/o Reference	45	70

	Category	Pa.	St.
Number of Annotations per Summary	1	2	5
	2-3	20	30
	4-5	12	27
	6-10	3	5
	<i>not given</i>	23	28
Overall Number of Annotators	1-5	19	25
	6-10	3	3
	> 10	5	9
Annotator Recruitment	<i>not given</i>	32	58
	Crowd	25	49
Statistical Evaluation	Other	35	46
	t-test	9	16
	ANOVA	9	18
	CI	4	6
	Other/unspec.	7	8
	None	32	47

TABLE 3.1: Our survey of 58 system papers with 95 manual evaluation studies (2017-2019). We show numbers both for individual studies and per paper. As a paper may contain several studies with different parameters, counts in the paper column do not always add up to 58.

Of the systems that do have human evaluation, many focus on *content*, including informativeness, coverage, focus, and relevance. Where linguistic quality is evaluated, most focus on general questions about fluency/readability, with a smaller number of papers evaluating coherence and repetition.

In the remainder of this section, we focus on the three aspects of evaluation we cover in this chapter: How to elicit judgements, how these judgements are analysed statistically, and how studies are designed.

3.3.1 Methods

The majority of evaluations is conducted using Likert-style judgements, with the second most frequent method being ranking-based annotations, including pairwise comparison and best-worst scaling. In QA evaluation (Narayan et al., 2018b), annotators must answer questions about the document from the given summary. This is naturally limited to evaluating content. This motivates us to compare both Likert and ranking annotations in Section 3.5.1.

3.3.2 Statistical Analysis

If a significance test is conducted, most papers analyse their data either using ANOVA or a sequence of paired t-tests. As already mentioned in Section 3.2.4, both tests are based on the assumption that judgements (or pairs of judgements, in case of the paired t-test) are sampled *independently* from each other. However, in almost all studies, annotators give judgements on more than one summary from the same system. Thus the resulting judgements are only independent if we assume that all annotators behave identically. Given that prior work (Gillick and Liu, 2010; Amidei et al., 2018), as well as our own reliability analysis in Section 3.5.1, show that especially crowd-workers tend to disagree about judgements, this assumption does not seem warranted. As a consequence, traditional significance tests are at high risk of inflated Type I error rates. This is well known in the broader field of linguistics (Barr et al., 2013), but is disregarded in summarization evaluation. We show in Section 3.6 that this is a substantial problem for current summarization evaluations and suggest alternative analysis methods.

3.3.3 Design

Most papers only report the number of documents in the evaluation and the number of judgements *per summary*. This, however, is not sufficient to describe the design of a study, lacking any indication about the overall number of annotators that made these judgements. A study with 100 summaries and 3 annotations per summary can mean 3 annotators did all judgements in one extreme, or a study with 300 distinct annotators in the other. Only 26 of the 95 studies describe their annotation design in full, almost all of which use designs in which a small number of annotators judge all summaries. Only **6** of **49** crowdsourced studies report the full design.

We show in Section 3.6 that a low total number of annotators aggravates Type I error rates with improper statistical analysis. In Section 3.7, we further show that, with proper analysis, a low total number of annotators leads to less powerful experiments. Almost all analysed papers choose designs with multiple judgements per summary. However, we show in Section 3.7.2 that this — for the purpose of system ranking — leads to loss of power when compared to a study with the same budget and only one annotation per summary.

3.4 Coherence and Repetition Annotation

To elicit summary judgements for analysis, we conduct studies on two linguistic quality dimensions. In the first, we ask annotators to judge the *Coherence* of the summaries, while in the second, we ask for the *Repetitiveness* of the summary. We select these two dimensions over the more frequent *Fluency* dimension as we found in preliminary investigations that many recent summarization systems already produce highly fluent text, making them hard to differentiate. We do not evaluate *Overall* and *Content* as both require access to the input document, which differentiates them from the linguistic quality dimensions.

For both quality dimensions, we conduct one study using a seven-point Likert-scale (**Likert**) and another using a ranking-based annotation method (**Rank**), where annotators rank summaries for the same document from best to worst.

Overview

In the following form you will be presented with a set of 25 summaries. Your task is to rate the coherence of each summary. Coherence is rated on a seven point scale where 7 means perfect coherence and 1 very poor coherence. Some summaries may cover the same or similar content. In these cases, please do not cross-reference between summaries and instead evaluate each summary on its own merits.

Coherent texts make sense, information is presented and organized in a logical order, and entities and events can be clearly identified. For example, a text in which it is unclear who or what noun phrases or pronouns refer to is probably less coherent than a text where all references are clear. Similarly, a text in which information is conveyed in a seemingly random order and/or a repetitive manner will have lower coherence than a well structured one.

[Start the study](#)

☐ I'm not a robot



(A) Likert - Coherence


Overview

In the following form you will be presented with a set of summaries about the same topic. Your task is to rank the summaries based on their coherence. You should rank the most coherent summary first and the least coherent summary last. You must assign every rank once. Ties are allowed.

Coherent texts make sense, information is presented and organized in a logical order, and entities and events can be clearly identified. For example, a text in which it is unclear who or what noun phrases or pronouns refer to is probably less coherent than a text where all references are clear. Similarly, a text in which information is conveyed in a seemingly random order and/or a repetitive manner will have lower coherence than a well structured one.

[Start the study](#)

☐ I'm not a robot



(B) Rank - Coherence

Overview


In the following form you will be presented with a series of summaries. Your task is to rate how well each summary avoids unnecessary repetition.

Problems with repetition can arise both by repeating full sentence or just nuggets of information. For example, in the following text the information about the date of the event is repeated twice: "The final concert of Justin Bieber's tour will take place on May 10th. The May 10th concert will take place in New York."

Please only judge the presence of unnecessary repetition and do not include any other aspects of the summaries in your rating, such as their overall quality.

[Start the study](#)

☐ I'm not a robot



(C) Likert - Repetition

Overview


In the following form you will be presented with a series of summaries about the same topic. Your task is to rank these summaries according to how well each summary avoids unnecessary repetition. You will be asked to rank the summary that best avoids unnecessary repetition first.

Problems with repetition can arise both by repeating full sentence or just nuggets of information. For example, in the following text the information about the date of the event is repeated twice: "The final concert of Justin Bieber's tour will take place on May 10th. The May 10th concert will take place in New York."

Please only judge the presence of unnecessary repetition and do not include any other aspects of the summaries in your ranking, such as their overall quality.

[Start the study](#)

☐ I'm not a robot



(D) Rank - Repetition

FIGURE 3.1: Screenshots of the annotator instructions.

We show screenshots of the instructions for both annotation methods and quality dimensions in Figure 3.1 and interfaces in Figure 3.2.

Corpus and Systems

Mirroring a common setup (see Section 3.3), we select four abstractive summarization systems and the reference summaries (**ref**) for analysis: The **pointer generator** model, the **abstractive sentence rewriter** model, **Seneca**, and **BART**.

The **pointer generator** model (PG) (See et al., 2017) enhances a standard encoder-decoder architecture (Bahdanau et al., 2015) with a *pointer-mechanism*. The pointer-mechanism allows the model to dynamically combine the usual prediction over the vocabulary P_{vocab} with the attention distribution A over the input tokens. This allows the model to “copy” any token in the input directly to the output, even if it is not part of the vocabulary.⁸ The combination is governed by a copy gate p_{copy} , which is computed from the hidden state. The final token distribution P is then

$$P = (1 - p_{\text{copy}})P_{\text{vocab}} + p_{\text{copy}}A. \quad (3.15)$$

We opt to include PG since it was, at the time of the study, a frequently used baseline model for abstractive summarization.

⁸Unlike the current state of the art for language modelling, PG uses a non-subword vocabulary.

[illegible]

<p>Summary 1/5</p>	
<p>Please make up following sentences and sort them in descending order of coherence in the list to the right.</p>	<p>Most coherent</p>
<p>It is announced this week, that England are pulling out of the event with immediate effect in order to achieve a more united France bid, including more foreign support, - England have pulled out of the home nations international under 16 tournament - they say their communications, including discussing the three nations games, would have harmed the profile of an historic competition that took place in 1905.</p>	<p>Least coherent</p>
<p>they sports' athletic cost-cutting across the board after paying 4 England is meant to retain premier league rights is being blamed for the demise of the victory stand, England have pulled out of the home nations international under 16 tournament, 4 sport are to broadcast the inaugural european games in June, having finally agreed terms.</p>	
<p>they allegedly withdrew their title sponsorship of under 16 tournament is sport are to broadcast the inaugural european games in June in new section -England join in due to location of stadium for an all-weather.</p>	
<p>they sports' athletic cost-cutting across the board after paying 4 England is meant to retain premier league rights is being blamed for the demise of the victory stand, the home nations under 16 tournament - England are pulling out of the event with immediate effect in order to achieve a more united France bid, - England have pulled out of a home nations international under 16 - the - and.</p>	
<p>they sports' athletic cost-cutting across the board after paying 4 England is meant to retain premier league rights is being blamed for the demise of the victory stand, 4 sports' have taken part of the programming since almost breaking the bank by committing 4.5 B, - England have pulled out of home nations under 16 -</p>	
<p>Comments</p>	

(A) Likert - Coherence

Summary 1/25

Please read the following summary

amanda berringer asked her brother broad fraser to make a toast at her wedding at eagle bay , south of perth. far from a conventional toast, mr fraser performed a song which posied fun at the borders of marriage, the crowd erupted into a standing ovation at the end of the performance, the song included jokes about the new husband needing to take the knee out and sailing the dog for a run.

How well does the summary avoid unnecessary repetition?

Very badly	Very well
1	2
3	4
5	6
7	8

Comments

Next

Show instructions

- The text repeats the same facts over and over, often using the same words.
- There is some repetition in the text, including repeated statements, but it is not too repetitive.
- There are no unnecessary repetitions in the text at all.

(B) Rank - Coherence

Summary 1/5	
Phase ran the following summaries into the list to the right so that the summary with the least amount of unnecessary repetition is first and the one with the most unnecessary repetition is last.	Least unnecessary repetition
bandrich was the highest-paid model in 2014, according to forbes magazine , with a total \$ 84 million in contracts. she is the face of chanel and carolina herrera has her own line of lingerie . the , and ,	
tom brady to glsie bandrich : " you inspire me every day " bandrich had last runway show wednesday she 'll be focusing more on family , " special projects "	Most unnecessary repetition
tom brady 's love for his wife will never go out of fashion. bandrich was the highest-paid model in 2014. bandrich announced her retirement from the catwalk last weekend. bandrich walked the runway for the last time wednesday and for new england patriots quarterback was n't there to support her in person .	
glsie bandrich, 34, announced her retirement from the catwalk last weekend. she was the highest-paid model in 2014, according to forbes magazine. she is the face of chanel and carolina herrera and has her own line of lingerie .	
tom brady 's love for his wife, model glsie bandrich, will never go out of fashion. bandrich, 34, announced her retirement from the catwalk last weekend. she is the face of chanel and carolina herrera and has her own line of lingerie .	
Comments	

(c) Likert - Repetition

(D) Rank - Repetition

FIGURE 3.2: Screenshots of the annotation interfaces.

The abstractive sentence rewriter (ASR) (Chen and Bansal, 2018) is a combined extractive and abstractive system that can be trained end2end using reinforcement learning. The extractor is a pointer network (Vinyals et al., 2015) that selects a set of input sentences from the input. The sentences are then rewritten using an *abstractor* network, which is itself implemented as a pointer-generator network. We select it for our study since it is a strong model that does not rely on external pretraining data.

Seneca (Sharma et al., 2019a) is a summarization model that is specifically designed to improve summary coherence. It extends the classical encoder-decoder architecture with a special *entity encoder* that encodes mention clusters for each entity in the document. Additionally, it is trained with a coherence reward using reinforcement learning, where the reward is provided by an external coherence measure. We include this model since it specifically aims to improve summary coherence, which makes it an interesting candidate model in our coherence experiments.

Finally, **BART** (Lewis et al., 2020), is a transformer (Vaswani et al., 2017) encoder-decoder network. Unlike the previous models, it is pretrained on a large corpus of unstructured text on a set of denoising tasks and then fine-tuned on a summarization dataset.

For all models, we use their variants trained on the CNN/DM corpus (Hermann et al., 2015). We randomly sample 100 documents from the CNN/DM test set and obtain the corresponding summaries from all summarizers to form the item set for all our studies.

Study design

To ensure a sufficient total number of annotators, we use a block design. We separated our corpus into 20 blocks of 5 documents and included all 5 summaries for each document in the same block. This results in $5 \times 5 = 25$ summaries per block.

All items in a block were judged by the same set of three annotators. No annotator was allowed to judge more than one block. This results in a total of $3 \times 20 = 60$ annotators and 1500 judgements per study. Figure 3.3 shows a schematic

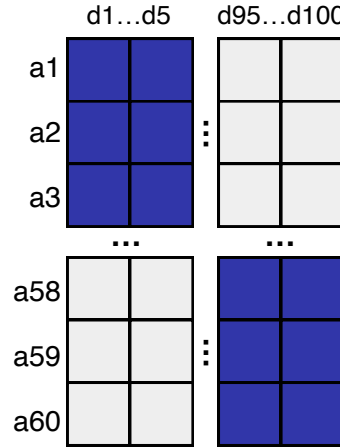


FIGURE 3.3: Schematic representation of our study design. Rows represent annotators, columns documents. Each blue square corresponds to a judgement of the summaries of all five systems for a document. Every rectangular group of blue squares forms one block.

overview of our design, which balances the need for a large enough annotator pool with a sufficient task size to be worthwhile to annotators.

We recruited native English speakers from the crowdsourcing platform Prolific⁹ and carefully adjusted the reward to be no lower than £7.50 per hour based on pilot studies. Summaries (or sets of summaries for **Rank**) within a block were presented in random order.

3.5 Ranking vs. Likert

Table 3.2 shows the average Likert scores and the average rank for all systems, quality dimensions, and annotation methods. We use mixed-effects ordinal regression to identify significant score differences. We describe the method and our reasoning for selecting it in detail in Section 3.6. Both annotation methods provide compatible system rankings for the two quality dimensions, in the sense that there are no statistically significant differences in opposite directions. However, we find that for repetition both methods struggle to differentiate between systems. If we were interested in the true ranking, we could conduct a power analysis given

⁹prolific.com

System	Likert (Coh)	Rank (Coh)	Likert (Rep)	Rank (Rep)
BART	5.25 ⁽¹⁾	1.73 ⁽¹⁾	5.85 ^(2/3)	2.88 ^(2/3/4)
ref	4.33 ^(3/4)	3.31 ^(3/4)	6.14 ^(1/2)	2.41 ^(1/2)
ASR	4.17 ^(3/4)	3.17 ^(3/4)	4.88 ^(4/5)	3.51 ^(4/5)
PG	4.81 ⁽²⁾	2.68 ⁽²⁾	5.63 ⁽³⁾	2.92 ^(3/4)
seneca	3.52 ⁽⁵⁾	4.11 ⁽⁵⁾	5.16 ^(4/5)	3.27 ^(3/4/5)

TABLE 3.2: Results of our annotation experiment. Numbers in brackets indicate the rank of a system for a given annotation method. Multiple ranks in the brackets indicate systems at these ranks are not statistically significantly different ($p \geq 0.05$, mixed-effects ordinal regression).

System	α	SHR
Coh: Likert	0.22	0.96
Coh: Rank	0.43	0.98
Rep: Likert	0.27	0.95
Rep: Rank	0.18	0.91

TABLE 3.3: Krippendorff’s α with ordinal level of measurement and Split-Half Reliability for both annotation methods on the two quality dimensions.

some effect size of interest and elicit additional judgements to improve the ranking. However, as we are concerned with the *process* of system evaluation and not with the system ranking itself, we do not conduct any further analysis. In the remainder of this section, we thus focus on the reliability of the two methods as well as their cost efficiency.

3.5.1 Reliability

As discussed in Section 3.2.1, while reliability is often computed by chance-adjusted agreement on individual instances, it is not necessarily a useful metric for subjective evaluation tasks, especially when we are not interested in individual *summary scores*, but in whether independent runs of the same study would result in consistent *system scores*. In Table 3.3 we thus report Split-Half Reliability (SHR) in addition to Krippendorff’s α . To compute SHR, we randomly divide judgements into two groups that share neither annotators nor documents, i.e. two independent

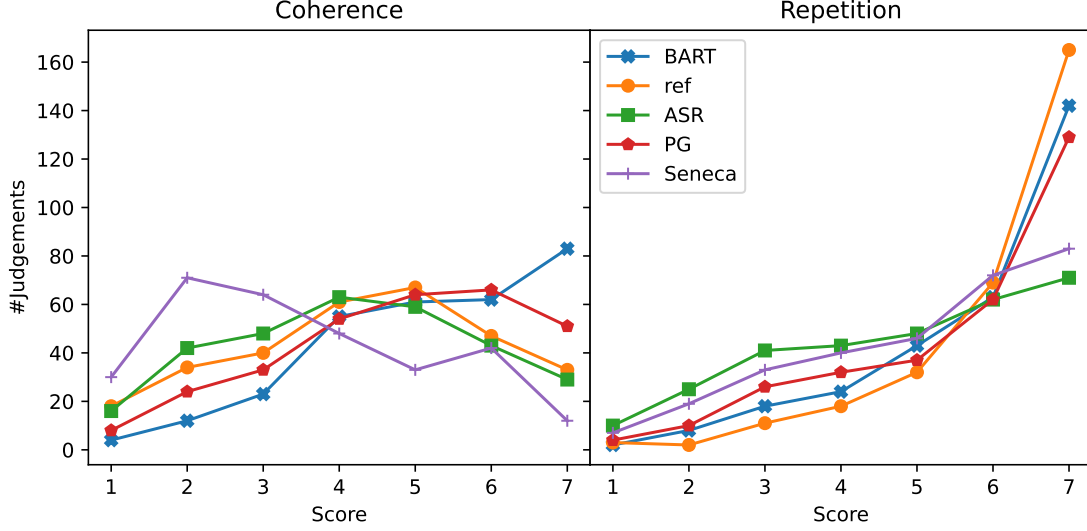


FIGURE 3.4: Score distribution of **Likert** for both quality dimensions. Each data point shows the number of times a particular score was assigned to each system.

runs of the study. We then compute the Pearson correlation¹⁰ between the system scores in both halves:

$$\rho(H_1, H_2) = \frac{\text{Cov}(H_1, H_2)}{\sqrt{\text{Var}(H_1)}\sqrt{\text{Var}(H_2)}}, \quad (3.16)$$

where $\text{Cov}(H_1, H_2)$ is the covariance of the human ratings in randomly sampled halves H_1, H_2 and $\text{Var}(H_1), \text{Var}(H_2)$ are their respective variances.

The final score is the average correlation after 1000 trials.

Though agreement on individual summaries is relatively low for all annotation methods, our studies still arrive at consistent system scores when we average over many annotators, as demonstrated by the SHR. This reflects similar observations made by Gillick and Liu (2010), who also find that non-expert annotators are unreliable on the summary level but produce similar overall rankings for linguistic quality judgements.

We find that on coherence, **Rank** is more reliable than **Likert**, though not on

¹⁰We use the Pearson correlation implementation of scipy (Virtanen et al., 2020).

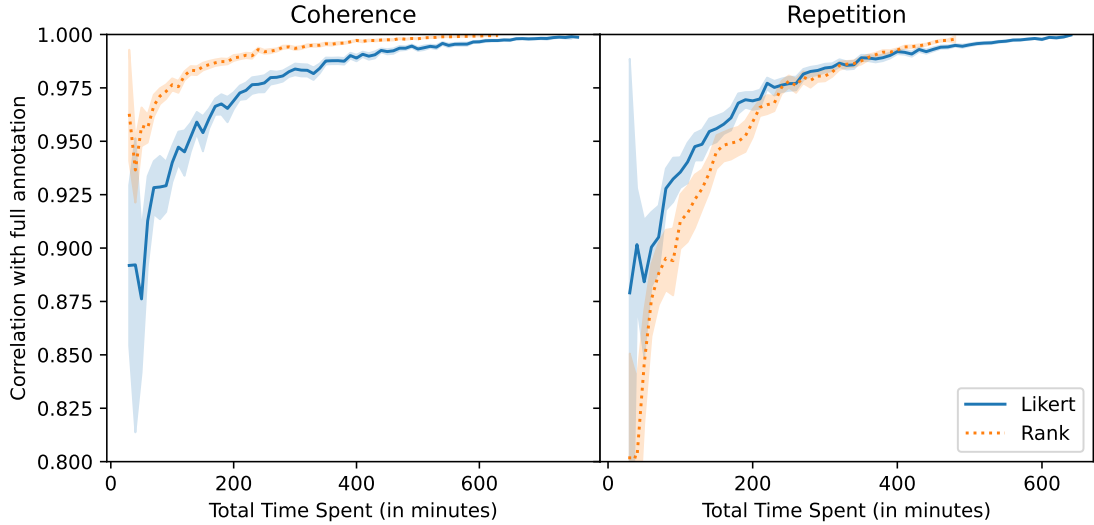


FIGURE 3.5: Time spent on annotation (in minutes) vs. correlation with the full-sized score. We gather annotation times in buckets with a width of ten minutes and show the 95% confidence interval for each bucket.

repetition. An investigation of the **Likert** score distributions for both quality dimensions in Figure 3.4 shows that coherence scores are relatively well differentiated whereas a majority of repetition judgements give the highest score of 7, indicating no repetition at all in most summaries. We speculate overall agreement suffers because ranking summaries with similarly low level of repetition (and not allowing ties) is potentially arbitrary.¹¹

3.5.2 Cost Efficiency

Computing SHR as above compares reliability for studies with an equal number of instances. However, the cost of a study is ultimately determined not by the number of instances but by the total time annotators need to spend on the task for a given reliability level. To enable this evaluation, we gather timing information during the completion of the annotations.

¹¹This is supported by feedback we received from annotators that the summaries were difficult to rank as they mostly avoided repetition well.

To fairly compare **Rank** and **Likert** annotations for time- and thus cost-efficiency, we randomly sample between 2 and 19 blocks from our annotations and compute the total time annotators spent to complete each set of annotations. We then compute the Pearson correlation of the system scores in each sample with the scores on the full annotation set. We relate time spent to similarity between sample and full score in Figure 3.5.

For coherence, **Rank** approaches near-perfect correlation faster than **Likert** in terms of overall time spent. On repetition, the lower SHR of both methods is also reflected in lower efficiency. In particular for **Rank** the lower overall SHR is reflected in much weaker correlation for small number of annotations and thus time spent. However, with additional annotation effort, reliability becomes on par with **Likert**. This is a consequence of the overall faster completion time for **Rank**.

3.6 Statistical Analysis and Type I Errors

Regardless of which of the two methods we choose for annotation, we are typically interested in establishing whether any differences we observe in system ratings are statistically significant. The two most common significance tests in summarization studies, ANOVA and t-test (see Table 3.1), both assume judgements (or pairs of judgements, in the case of t-test) are independently sampled. This is, however, not true for most study setups as a single annotator typically judges multiple summaries and multiple summaries are generated from the same input document. Both documents and annotators are thus grouping factors in a study that must be taken into account by the statistical analysis.

Generalized mixed-effects models (Barr et al., 2013) offer a solution but have, to the best of our knowledge, not been used in summarization evaluation at all. We choose a mixed-effects ordered logit model to analyse our data for both quality dimensions and annotation methods. We will show that traditional analysis methods have a substantially elevated risk of Type I errors, i.e. differences between systems might be overstated.

3.6.1 Mixed-Effects Models

To better understand mixed-effects models, we first repeat the definition of a linear regression model:

$$y = X\beta + \epsilon, \quad (3.17)$$

where $y \in \mathbb{R}^n$ is the target vector, $X \in \mathbb{R}^{n \times k}$ and $\beta \in \mathbb{R}^k$ are the feature matrix for a dataset of size n and the unknown parameter vector for k features respectively. $\epsilon \in \mathbb{R}^n$, $\epsilon_i \sim N(0, \sigma^2)$ are the residuals, which are drawn from a normal distribution with mean 0 and variance σ^2 . As is apparent from the above definition, this assumes that errors are independently, identically distributed across samples in the dataset. However, as discussed in the previous section, this assumption does not hold for many real-world datasets.

Linear mixed-effects models tackle this problem by introducing an additional structured error term into the equation:

$$y = X\beta + Zu + \epsilon. \quad (3.18)$$

$Z \in \mathbb{R}^{n \times m}$ is again a design matrix, similar to X , relating m “features” in u to samples in the dataset. However, unlike β , $u \sim N(0, \psi)$ is not a vector of observations but instead itself a random variable drawn from a multivariate normal distribution with zero mean and covariance matrix ψ . This allows us to express dependencies between the error terms of individual instances in the dataset by specifying a corresponding design matrix Z .

To better illustrate the purpose of the random effects, consider the unconditional distribution of y :

$$y \sim N(X\beta, Z\psi Z^T + I\epsilon), \quad (3.19)$$

where I is the $n \times n$ identity matrix.

Here, the role of the random effects becomes easily apparent: They allow us to model the complex covariance structures that arise from non-independent samples (Riezler and Hagmann, 2024).

3.6.2 Ordinal Regression

While the above formulation assumes a linear relationship between features and the response variable, this is not an appropriate assumption when working with Likert-style or ranking data,¹² such as in our experiments. We thus turn to ordinal regression.

Ordinal regression can be realized using a *generalized* linear model (McCullagh and Nelder, 2019). In generalized linear models, the output of the linear model is linked to the response variable using a *link function* g , so that the expected value of the response y given the features X is the result of applying the inverse function of g , g^{-1} to the output of the linear model:

$$E[y|X] = g^{-1}(X\beta), \quad (3.20)$$

where g^{-1} is applied element-wise.

This can naturally be combined with the mixed-effects model formulation above, which gives us the following formulation for a generalized mixed model:

$$E[y|X, u] = g^{-1}(X\beta + Zu). \quad (3.21)$$

We can now formulate ordinal regression within this framework. Given a set of K levels, ordinal regression divides the real number line into K different segments using $K - 1$ threshold values μ_1, \dots, μ_{K-1} . The probability of a response y being smaller than or equal to the i -th level of our response scale is then given by

$$P(y \leq i) = g^{-1}(\mu_i - (X\beta + Zu)), \quad (3.22)$$

where the choice of inverse link function g^{-1} determines which variant of ordinal regression we use. We chose g^{-1} as the logistic function $\sigma(x) = (1 + e^{-x})^{-1}$, which gives us the commonly used ordered logit model.

¹²Strictly speaking, the ordinal model is also not fully appropriate for ranking data, since it does not model the exclusivity in ranks within judgements for a single document. However, accounting for this would introduce additional complexity to the model and we found empirically that the model fits are of similar quality for Rank and Likert data.

3.6.3 Hypothesis Testing using Generalized Mixed Models

To compute the significance of score differences between instances from k different categories (e.g. systems), we first choose one category as the *reference* category and introduce $k - 1$ binary indicator variables, one for each remaining category. We then compute a maximum likelihood estimate of the model, yielding estimates $\hat{\beta}_1 \dots \hat{\beta}_{k-1}$ for the difference between the reference and each remaining category on the latent scale, as well as their $(k - 1) \times (k - 1)$ covariance matrix CV .

To test for significance between the reference category and any category i , we read out the variance $Var(\hat{\beta}_i)$ of the estimate from the diagonal of CV and compute a z-score $z_{0,i}$:

$$z_{0,i} = \frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}}. \quad (3.23)$$

To compare two non-reference categories i, j , we instead compute the z-score of the pairwise contrast:

$$z_{i,j} = \frac{\hat{\beta}_i - \hat{\beta}_j}{\sqrt{Var(\hat{\beta}_i) + Var(\hat{\beta}_j) - 2Cov(\hat{\beta}_i, \hat{\beta}_j)}}, \quad (3.24)$$

where we can again read out $Cov(\hat{\beta}_i, \hat{\beta}_j)$ from the corresponding cell in CV .

Since our experiments all involve multiple comparisons, one for each pair of systems, we compute p-values from the z-scores on the studentized range distribution with k groups and infinite degrees of freedom to adjust for multiplicity.¹³

3.6.4 Modelling our Annotations

We can now describe the model we use for analyzing our data. We choose the human-written reference summaries as the reference level and introduce one fixed effect per summarizer.

To specify the random effects structure, we follow common advice (Barr et al., 2013) in specifying a *maximal* random effects structure, that is to account for all potential grouping factors in the model. In our case, we specify both slopes

¹³This corresponds to the *tukey*-adjustment option in `emmeans`.

and intercepts for documents and annotators. In practical terms, this means that we allow for documents to be both easier or harder in general (i.e. summaries receive more or less favorable ratings overall) and to be easier or harder for each individual summarizer. Similarly, annotators can both be more or less generous in their scores overall and have individual annotator preferences.

The linear part of the model (i.e. before the inverse link function) is

$$X\beta + Z_a u_a + Z_d u_d. \quad (3.25)$$

X is a $n \times s - 1$ matrix, where s is the number of summarizers (including the reference summaries) in our study and n is the number of judgements. $X_{ij} = 1$ if the i -th judgements was given to a summary generated by the j -th summarizer, where all reference summaries have a zero row vector.

Z_a and Z_d specify our random effects structure. Z_a is a $n \times as$ matrix that specifies the random effects for the a annotators. For each annotator, the matrix contains s parameters: $s - 1$ slopes for each non-reference summarizer and one intercept. The matrix is highly sparse: Each row contains a value of 1 in the intercept column corresponding to the annotator who made the judgement. Additionally, for every summarizer except the reference summaries, the row contains one additional entry with value 1 for the corresponding summarizer slope column. All other entries are zero. Z_d is a $n \times ds$ matrix that specifies the random effects for the d documents. It is structured similarly to Z_a , except its non-zero entries correspond to the *document* for which the summary was generated.

$u_a \sim N(0, \psi_a)$, $u_d \sim N(0, \psi_d)$ are the random effect vectors. They are drawn from a normal distribution with zero mean and covariance matrices ψ_a for the annotator random effects and ψ_d for the document random effects. ψ_a and ψ_d are estimated together with β during model fitting.

We fit all models using the **ordinal** R-package (Christensen, 2019) and compute pairwise contrasts between the parameters estimated for each system using the **emmeans**-package (Lenth et al., 2018).

3.6.5 Demonstrating the Dangers of Ignoring Grouping Factors

Our analysis method introduces a large degree of additional complexity when compared to applying a simple t-test. To demonstrate the necessity of modelling the grouping factors, we can conduct a simulation experiment where we drop all fixed effects from the model. We can then sample from the model and analyse it with inappropriate tests. Since in a model without fixed effects all summaries have the same expected score, regardless of summarizer, we would expect a well-calibrated test to reject the null hypothesis that any pair of summarizers has scores drawn from different distributions at a rate of exactly the significance level α . This Monte-Carlo simulation is similar to the more general analysis of Barr et al. (2013).

We thus set β to $\vec{0}$ leaving only the following (linear) part of our model:

$$Z_a u_a + Z_d u_d. \quad (3.26)$$

Since u_a, u_d both have zero mean, it is easy to see that the resulting latent values (and thus overall scores) must have the same expected value.

We then repeatedly apply both the t-test and the approximate randomization test (ART) (Noreen, 1989) to samples drawn from the model and determine the Type I error rate at $p < 0.05$. We set the number of documents to 100 and demand 3 judgements per summary to mirror a common setup in manual evaluation. We then vary the total number of annotators between 3 and 300 by changing how many summaries a single annotator judges.

Results

We report results given the model estimated for **Likert** in Figure 3.6.¹⁴ Ignoring the dependencies between samples leads to inflated Type I error rates, whether we use the t-test or the ART. This is especially severe when only few annotators judge the whole corpus. In the extreme case with only three annotators in total, the null hypothesis is rejected in about 40% of trials at a significance level of 0.05

¹⁴We do not include **Rank** data in this and the following simulation experiments, as the ordinal regression model does not generate ranks.

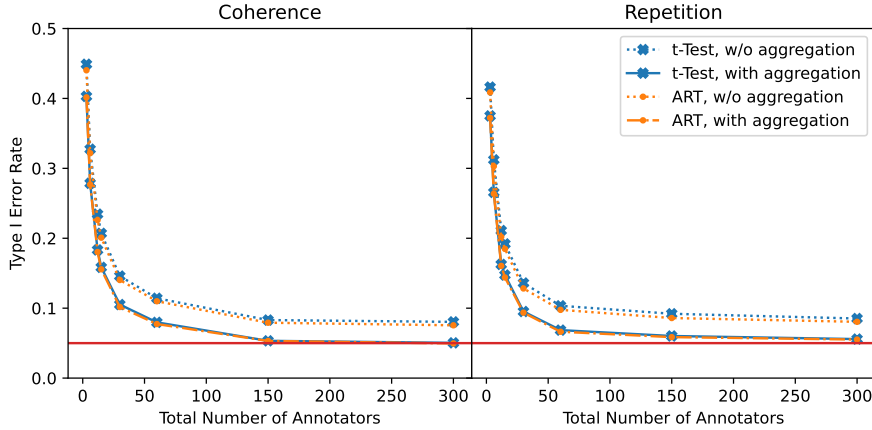


FIGURE 3.6: Relation of Type I error rates at $p < 0.05$ to the total number of annotators for different designs, all with 100 documents and 3 judgements per summary. We conduct the experiment with both the t-test and approximate randomization test (ART). We show results both with averaging results per document and without any aggregation. We run 2000 trials per design. The red line marks the nominal error rate of 0.05.

in both quality dimensions. Even our original design with 60 annotators still sees an increase of the Type I error rate by about 3 percentage points. Only if every annotator judges a single document and annotations are averaged per document, samples are independent and thus the real error is at the nominal 0.05 level. This design, however, is unrealistic given that annotators must be recruited and instructed.

We suggest two solutions to this problem:

1. Use mixed-effects models and fully specify the random effects structure to capture all dependencies in the data.
2. Aggregate the judgements so samples become independent.

The latter approach allows the assumptions of simpler tools such as ART to be met. In our study, we could average judgements in every block to receive independent samples. This is only possible, however, if the design of the study considers this problem in advance: a crowd-sourcing study that allows annotators to judge as many samples as they like is unlikely to result in a design with a sufficient number of independent samples.

3.7 Study Design and Study Power

When conducting studies for system comparison, we are interested in maximizing their power to detect differences between systems. For traditional analysis, the power is ultimately determined by the number of documents (or judgements, when no aggregation takes place) in the study. However, when analysis takes into account individual annotators, power becomes additionally dependent on the total number of annotators and how evenly they participated in the study. This gives additional importance to the design of evaluation studies. In this section, we thus focus on how to optimize studies for power and reliability.

We first show that for well-powered experiments, we need to ensure that a sufficient total number of annotators participates in a study. In the second part of this section, we will then demonstrate studies can improve their power by not eliciting multiple judgements per summary.

3.7.1 Overall Number of Annotators

To demonstrate the difference in power caused by varying the total number of annotators in a study, we determine the power for a design with the same total number of documents and judgements *per summary* but different *total* numbers of annotators.

We run the experiment both with regression and ART with proper aggregation of dependent samples as described in Section 3.6. We refer to the latter as **ARTagg** to differentiate it from normal ART.

For each design, we repeatedly sample artificial data from the **Likert** model and apply both tests to the data. The process is the same as in Section 3.6 except we do not set β to zero and count acceptances of the null hypothesis.¹⁵

We again set the number of documents to 100 and the number of repeat judgements to 3 and vary the total number of annotators between 3 and 75 by varying the number of blocks between 1 and 25. We test for power at a significance level of 0.05.

¹⁵As this is an observed (or *post-hoc*) power analysis, it probably overestimates the power of our analysis for the true effect. The analysis is thus only useful to compare designs under our best estimate of actual effect sizes.

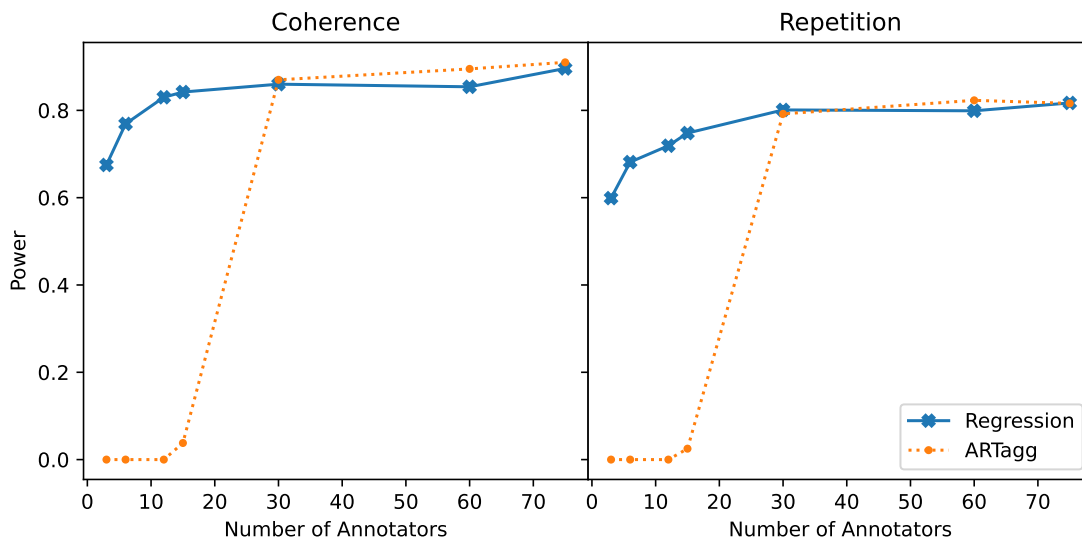


FIGURE 3.7: Power for 100 documents and 3 judgements per summary with different number of total annotators.

Figure 3.7 shows how power drops sharply when only few annotators take part in the study. This is in line with the theoretical analysis of Judd et al. (2017) that shows that the number of participants is crucial for power when analysing studies with mixed-effects models. **ARTagg** is especially sensitive to the number of annotators as fewer annotators mean fewer independent blocks and thus a lack of data points for the analysis. In the extreme case with three annotators judging the entire dataset, we only have a single data point, making analysis impossible. The mixed-effects model approach, on the other hand, performs better with a smaller number of annotators, at the expense of additional modelling complexity and compute intensity.

3.7.2 Annotator Distribution

Most studies elicit multiple judgements per summary, following best practices in NLP for corpus design (Carletta, 1996). While this leads to better judgements per *summary*, the goal of many summarization evaluations is a per *system* judgement. This mirrors our argument regarding SHR and agreement for reliability in Section 3.5.1: Our discussion in this chapter assumes we are interested in the global

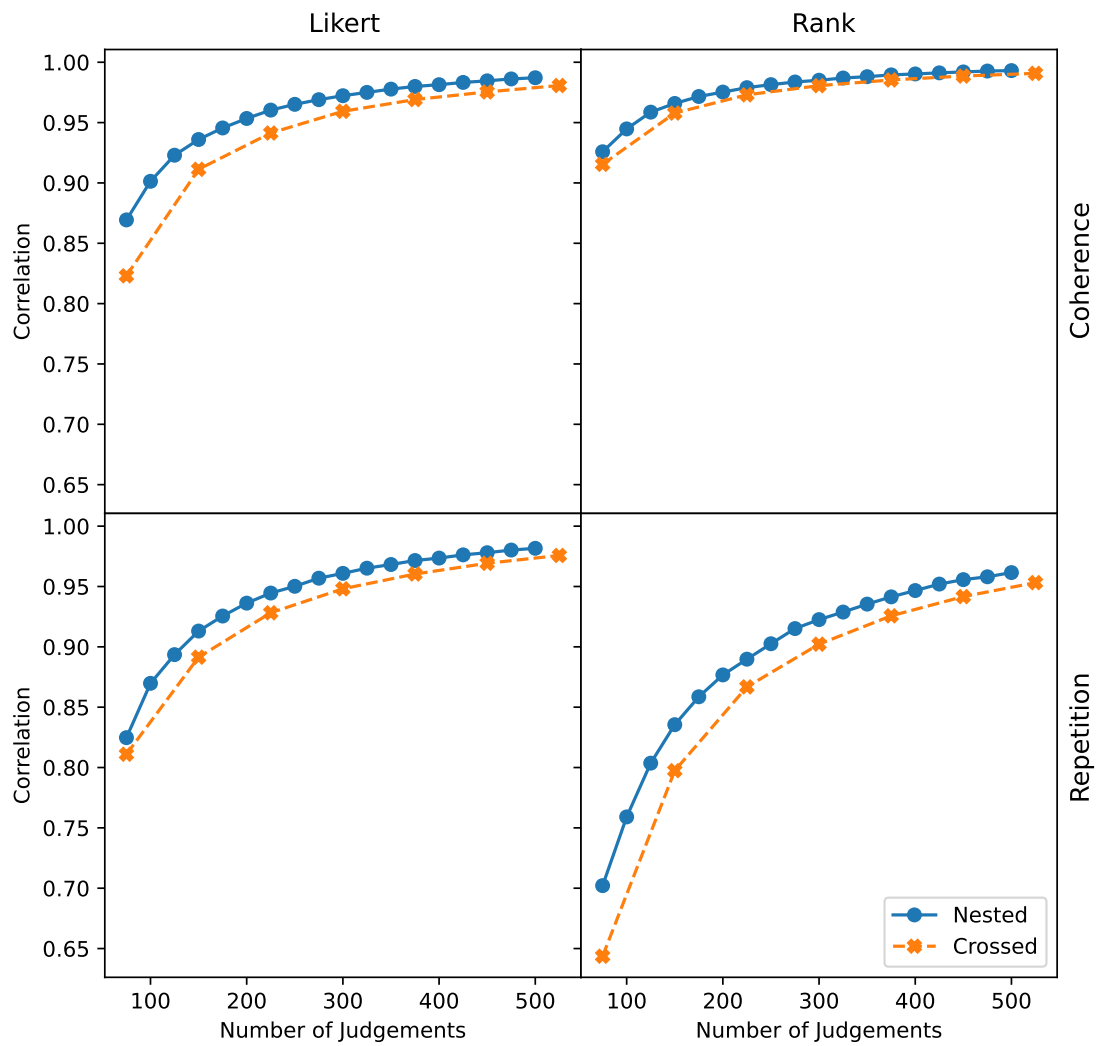


FIGURE 3.8: Reliabilities of nested vs. crossed designs for Rank and Likert for both quality dimensions.

performance characteristics of the model, not an accurate estimate of the quality of each individual summary.

For this kind of study, Judd et al. (2017) show that for mixed models that include both annotator and target (in our case, input document) effects, a design where targets are *nested* within annotators, i.e. every annotator has its own set of documents, is always more powerful than one where they are (partially) *crossed* with annotators, i.e. a study with multiple annotations per summary, *given the same total number of judgements*. In fact, power could be maximized by having each annotator judge the summaries for only a single, unique document. However, this is usually not realistic due to the fixed costs of annotator recruitment and instruction. To help conceptualize why this is the case, consider that forgoing multiple annotations per documents allows the study to obtain judgements across a wider variety of input documents. This in turn allows the study to better capture the variance of system performance across more diverse input documents, leading to a better estimate of model performance on the dataset.

We can demonstrate on our dataset how both reliability and power are affected by nested vs. crossed design.

To compare reliability, we randomly sample both nested and crossed designs from our full study and then compute the Pearson correlation of the system scores given by this smaller annotation set with the system scores given by the full study. As shown in Figure 3.8, nested samples are always at least as good and usually better at approximating the results of the full annotation compared to a crossed sample with the same annotation effort.

Alternatively, we can also simulate the effect of the design on study power. We conduct a power analysis for regression and **ARTagg** comparing nested and crossed designs. We again turn to Monte-Carlo simulation on the **Likert** models and sample nested and crossed designs with the same total number of judgements (i.e. the same cost). We keep the block size constant at 5 and vary the number of annotators between 3 and 60. For nested designs, we drop the document-level random effects from the ordinal regression, as the input document is no longer a grouping factor in nested designs.

Figure 3.9 shows that nested designs always have a power advantage over crossed designs, especially when few judgements are elicited. We also find that

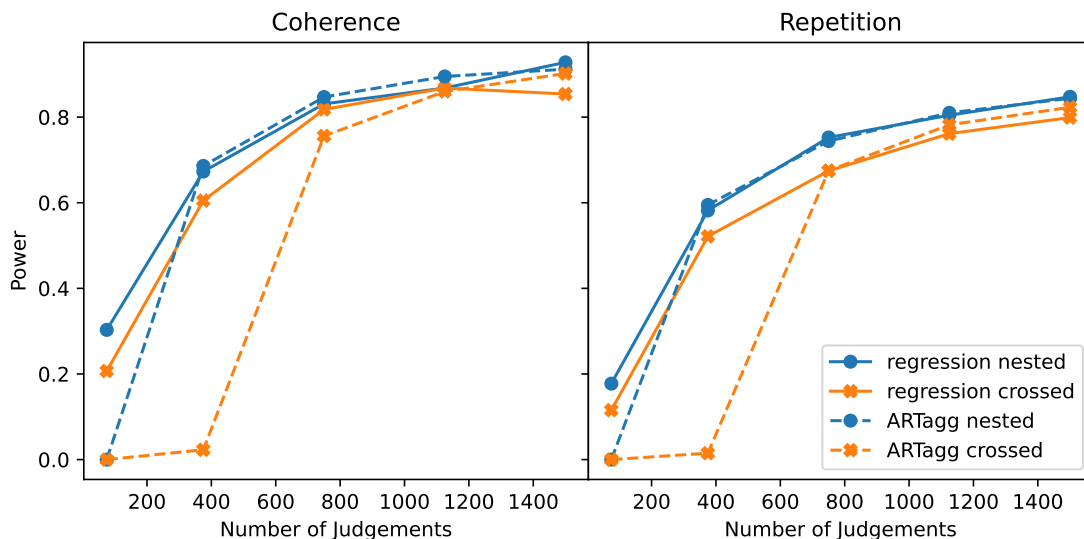


FIGURE 3.9: Power for $p < 0.05$ of nested and crossed designs for ARTagg and regression. X-axis shows the number of judgements elicited, Y-axis the power level.

ART can be used to analyse data without loss of power when there are enough independent blocks. This might be attractive as ART is less computationally expensive than ordinal regression.

3.8 Five Years Later: Have Practices Changed?

The survey in Section 3.3 was conducted for the original publication of the work underlying this chapter. Have practices improved in the intervening five years?

To answer this question, we repeat a smaller-scale version of our survey. We select two recent *ACL conferences at the time of this writing: NAACL 2024 and EACL 2024. We also include papers from the Findings of the conferences. Findings is an additional acceptance category that has been introduced to *ACL conferences starting from 2020.¹⁶ We use the same criteria as in our original survey. Details on the new survey can be found in Appendix B.3.

¹⁶See <https://web.archive.org/web/20240404125546/https://2020.emnlp.org/blog/2020-04-19-findings-of-emnlp>.

	Category	Pa.	St.
Evaluation Questions	Overall	1	2
	Content	6	9
	Fluency	4	5
	Coherence	4	6
	Repetition	2	2
	Faithfulness	3	3
	Other	1	1
Evaluation Method	Likert	3	4
	Pairwise	3	4
	Rank	1	1
	BWS	1	1
	Other	1	1
Number of Documents in Evaluation	< 20	1	2
	50-99	2	2
	100	4	5
	> 100	1	1
	<i>not given</i>	1	1
Number of Systems considered	< 3	2	2
	3	1	1
	4	4	7
	> 5	1	1
	w/ Reference	4	5
	w/o Reference	4	6

	Category	Pa.	St.
Num. of Ann. per Sum.	2-3	3	4
	4-5	1	1
	<i>not given</i>	4	6
Overall	1-5	6	8
Number of Annotators	> 10	1	1
	<i>not given</i>	2	2
Annotator Recruitment	Crowd	3	3
	Other	6	8
Statistical Evaluation	Other/unspec.	1	1
	None	7	10

TABLE 3.4: Results of our new survey of 8 papers with 11 studies from EACL 2024 and NACL 2024. For better comparability, we reproduce the structure from Table 3.1. We remove categories which do not apply to any of the covered studies.

We find a total of 23 papers, 8 (i.e. 34%) of which contain human evaluation. This is an unfortunate decrease from the 55% in our original survey. While the exact causes are difficult to establish, we find some contributing factors during our survey:

- The advent of larger contexts in models has allowed for the summarization longer documents, which increases the difficulty of human evaluation. For example, Saxena and Keller (2024) point this out as a justification for not conducting human evaluation.
- Some of the surveyed papers optimize specific aspects of the summaries, such as their faithfulness (Elhady et al., 2024; Shi et al., 2024) and exclusively rely on automatic metrics to assess improvements in this dimensions.

However, we argue that in both cases, forgoing human judgements is risky, since automatic evaluation often remains unreliable, in spite of recent improvements (Chen et al., 2024; Koo et al., 2024; Panickssery et al., 2024).

For the 8 papers with human evaluation studies, we conduct the same detailed survey as before and report results in Table 3.4. We find that study sizes have remained largely similar. Regarding criteria, we find that in addition to content, which is the dominant category, coherence remains an important quality dimension even with newer models. We find the definition for repetition has shifted, with a focus on *conciseness* instead of verbatim repetition of sentences. Likert remains a popular choice of annotation method, although ranking-based approaches have become relatively more frequent in our sample.

Discouragingly, the under-reporting of experimental details is still frequent in the surveyed papers. Only three studies give sufficient information to reproduce the full design because all annotators annotate all instances. The number of annotators – where reported – is also low. As our results show, if the task is subjective, this is unlikely to represent the true population preferences.

A more worrisome development can be observed in the use of statistical tests: Only a single paper conducts a statistical test at all, but does not account for grouping factors. The absence of good tools for properly interpreting numeric results increases the risk of misleading conclusions.

While our sample is small due to the low number of human studies overall, this survey suggests that most of the problems reported in this chapter remain unfortunately unaddressed in spite of our suggestions.

3.9 Discussion

In this chapter, we have presented both a survey of the state of manual summary evaluation, as well as our investigation of methods, statistical analysis, and design of such studies. We can distill our findings to a set of recommendations for manual summary quality evaluation:

Method. Both ranking and Likert-style annotations are valid choices for quality judgements. However, we present preliminary evidence that the optimal choice of method is dependent on task characteristics: If many summaries are similar for a given quality dimension, Likert may be the better option.

Analysis. Analysis of elicited data should take into account variance in annotator preferences to avoid inflating Type I error rates. We suggest the use of mixed-effects models for analysis that can explicitly take into account grouping factors in studies. Alternatively, traditional tests can be used with proper study design and aggregation.

Study Design. Study designers should control the number of annotators and how many summaries each individual annotator judges to ensure sufficient study power. Additionally, to ensure reliability of results, studies should report the design and the total number of annotators in addition to the number of documents and repeat judgements. Studies with repeat judgements on the same summary do not provide any advantage for system comparison and are less powerful than nested studies of the same size.

These recommendations are designed to ensure researchers can conduct reliable annotation studies that are also cost-efficient. However, our survey shows that they are not widely followed in current evaluation practices. This is unfortunate as

particularly the improper analysis of elicited data can lead to incorrect conclusions about system performance.

An important caveat in these recommendations is that they are made under the assumption that the study in question is designed to derive a *system ranking*. Thus, all our evaluations focus primarily on the reliability of the resulting system scores. This is particularly relevant for our recommendations with regard to study design: Forgoing repeat judgements naturally leads to unreliable results on the level of individual summaries. For certain studies, e.g. those that are designed for metric meta-evaluation, a topic we will cover in the following chapter, this is highly undesirable.

With these recommendations, we have completed the first component of our holistic evaluation framework. While we have conducted our experiments on coherence and repetition, our findings with regard to study design and analysis are applicable to any reference-free quality dimension, including both linguistic quality and content dimensions. Finally, our findings suggest that different quality dimensions might require individual evaluations to find the optimal annotation method.

Chapter 4

Meta-Evaluation: A Case Study in Summary Coherence

4.1 Motivation

While the insights into human evaluation from the previous chapter allow us to design cost-efficient studies, this cost-efficiency is relative only to other manual evaluation studies. Even well-designed human studies are still cost-prohibitive where frequent evaluations are needed. Such evaluations are, however, critical to provide feedback for model development. We will thus now turn to *automatic* evaluation metrics, beginning with a discussion of *meta*-evaluation, i.e. the evaluation of evaluation metrics. Whereas in human evaluation we are limited to proxy measures of validity, we can validate automatic metrics by comparing them to human preferences. This is critical to ensure that automatic metrics are useful proxies for manual evaluation and is usually achieved by computing correlation between human and predicted scores on a set of system outputs (Lin, 2004b; Papineni et al., 2002; Zhao et al., 2019; Zhang et al., 2020b, among others).

While this procedure is intuitive, a fundamental shortcoming is that the estimates of correlation with human judgements must, in practice, always be computed using the outputs of a limited set of summarizers which have been rated by human annotators. This causes similar issues of dependence between individual samples as those discussed in Chapter 3 (see Deutsch et al., 2021b, for a discussion). Additionally, however, the generalizability of meta-evaluation poses a

particular challenge. Since one primary purpose of evaluation metrics is to evaluate the performance of future summarizers, the distribution of summarizers in the meta-evaluation dataset does not necessarily reflect the distribution of summarizers the metric is expected to work on. Correlation with scores on typical summaries at the time of evaluation is not necessarily predictive of future utility.

To some degree, this issue is unavoidable since solving it would require us to make predictions about future developments. However, in this chapter, we argue that we can increase the chance of a meta-evaluation yielding generalizable results by designing it to avoid what we will refer to as *system-level confounders*. We define system-level confounders to be any features of summarizer outputs that are not related to the modelled quality dimension but instead happen to identify good summarization systems in a particular dataset. To illustrate, consider a dataset for summary relevance which only contains summaries generated from strong summarizers, that are both relevant and grammatical, and weak summarizers, that are neither. A metric that only measures grammaticality could easily achieve strong correlation with relevance judgements on this dataset but would not generalize to a scenario with multiple summarizers that are all grammatical. Accounting for system-level confounders increases the chance that any observed correlation is due to a metric genuinely capturing some aspects of the relevant quality dimension. It can, however, naturally not *guarantee* that these aspects are also where future systems will differ.

In this chapter, we will thus introduce a new meta-evaluation metric and a new analysis method that can be used for reliable meta-evaluation. To avoid confusion, we are going to refer to meta-evaluation metrics as *evaluation metrics* (**EMs**) in the remainder of this chapter. To contrast, we are going to call the automatic metrics we seek to evaluate *measures*. Supported by our insights on meta-evaluation, we are then going to conduct a case study on the meta-evaluation of coherence measures (**CMs**).

We choose coherence as the quality dimension in our study, since there is very little agreement on which coherence measures are most suitable for practical application. While many CMs have been suggested for automatically assigning a coherence score to a summary, there is, to the best of our knowledge, little in the way of systematic comparison. Proposed methods include learning from human

judgements (Barzilay and Lapata, 2008; Tien Nguyen and Joty, 2017; Xenouleas et al., 2019; Mesgar et al., 2021), learning from the *shuffle task* (Mohiuddin et al., 2021; Jwalapuram et al., 2022), where models are trained to discriminate original documents from documents with randomized sentence order (Barzilay and Lapata, 2008), and using next sentence prediction as a proxy task (Koto et al., 2022). Finally, unsupervised measures that exploit heuristics (Pitler et al., 2010; Zhu and Bhat, 2020) or large-scale LMs (Yuan et al., 2021) have also been suggested.

However, evaluation is often conducted on disparate datasets, which makes scores incomparable. Meta-evaluation also often uses system outputs from DUC conferences (Barzilay and Lapata, 2008; Tien Nguyen and Joty, 2017; Xenouleas et al., 2019; Mesgar et al., 2021), which do not represent recent advances in text summarization. In addition, there is no agreement on *how* the CM scores should be compared to human scores. System-level correlation (Xenouleas et al., 2019; Fabbri et al., 2021b), pairwise ranking accuracy (Barzilay and Lapata, 2008; Tien Nguyen and Joty, 2017; Mesgar et al., 2021), and summary-level correlation (Yuan et al., 2021) have all been suggested as EMs. This makes it hard to ascertain the state of summary coherence modelling and to identify promising directions for future research.

We will make the following contributions with regard to meta-evaluation in this chapter:

1. We show that current EMs provide an incomplete picture of measure performance as they focus on comparing summaries generated by different summarizers. In case of popular summary coherence datasets, this often includes many easy decisions due to the large performance gaps between them. EMs are also vulnerable to CMs exploiting confounding system properties to correctly rank systems without modelling coherence.
2. We introduce a new EM, *intra-system correlation*, that measures performance within the summaries generated by a single summarizer and is both more challenging and more resilient against system-level confounders.
3. We introduce *bias matrices* as a novel analysis tool that allow to easily detect when measures are biased towards specific summarizers.

Using these insights, we conduct a large-scale comparison of CMs on the *SummEval* dataset (Fabbri et al., 2021b). We show that:

1. All investigated CMs exhibit significant weaknesses under evaluation regimes other than system-level correlation.
2. Even relatively strong CMs are biased towards outputs of certain summarizers, which raises concern about their generalizability.
3. SummEval is not conducive to entity-based modelling, which has been successful on many other coherence tasks (Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Tien Nguyen and Joty, 2017; Mesgar et al., 2021).
4. While most of the shuffle-based models transfer poorly to summaries, which is in line with prior results by Mohiuddin et al. (2021), the most promising performance is achieved by fine-tuning a masked language model (MLM) on the shuffle task as a *classifier*. We present evidence that this allows the model to adapt more easily to comparing summaries of different content and lengths, highlighting a possible avenue for future work.

The work presented in this chapter has previously been published as

Julius Steen and Katja Markert (2022). “How to Find Strong Summary Coherence Measures? A Toolbox and a Comparative Study for Summary Coherence Measure Evaluation”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6035–6049. URL: <https://aclanthology.org/2022.coling-1.527>.

4.2 Background

4.2.1 Meta-Evaluation

Most meta-evaluation for summarization starts by gathering human judgements $H = \{H_{(d,s)} | d \in D, s \in S\}$, where D is the set of input documents used to elicit

summaries from the summarizers and S is the set of summarizers in the meta-evaluation dataset. Human judgements are typically the result of an average over the judgements of multiple annotators to increase reliability of the summary-level scores. We can then compute the automatic scores of the measure in question on the same set of documents $P = \{P_{(d,s)} | d \in D, s \in S\}$.

We can directly define a meta-evaluation metric by computing some correlation function $f(H, P)$ between human ratings H and predicted scores P . This is also referred to as *summary-level* correlation.

If we additionally aggregate over scores for each summarizer, we arrive at the also commonly used *system-level* correlation:

$$H_{s_j}^{(\text{sys})} = \frac{1}{|D|} \sum_{d_i \in D} H_{(d_i, s_j)}, \quad (4.1)$$

$$P_{s_j}^{(\text{sys})} = \frac{1}{|D|} \sum_{d_i \in D} P_{(d_i, s_j)}. \quad (4.2)$$

The meta-evaluation score for the predicted scores P is then $f(H^{(\text{sys})}, P^{(\text{sys})})$.

The correlation function f is typically instantiated as one of three commonly accepted correlation coefficients: Pearson's ρ , Spearman's ρ , or Kendall's τ . We will briefly introduce each of these. For clarity of exposition, we will assume we are working with *summary-level* aggregations H, P .

Pearson's correlation coefficient tests the strength of the linear correlation between H and P . We have already introduced this correlation more generally in Section 3.5.1 but restate it here for the specific case of meta-evaluation. Pearson's correlation takes a value of 1 if P and H are perfectly linearly correlated and a value of 0 if there is no linear correlation between P and H . A value of -1 would indicate perfect *inverse* correlation between predicted and human scores, although this is naturally rare in meta-evaluation. Pearson's correlation coefficient is computed as sample co-variance, normalized by the product of the standard deviations of P and H :

$$\rho_{\text{Pear}}(H, P) = \frac{\sum_{d \in D} \sum_{s \in S} (P_{(d,s)} - \bar{P}) (H_{(d,s)} - \bar{H})}{\sqrt{\sum_{d \in D} \sum_{s \in S} (P_{(d,s)} - \bar{P})^2} \sqrt{\sum_{d \in D} \sum_{s \in S} (H_{(d,s)} - \bar{H})^2}}, \quad (4.3)$$

where \bar{P}, \bar{H} are the mean predicted and human scores, respectively. For system-level correlation, we simply use $H^{(\text{sys})}, P^{(\text{sys})}$ and drop the sums over the documents. The focus on measuring the linear relationship with human scores means that a metric can have low Pearson correlation even if it ranks individual summaries perfectly. A second issue is that Pearson's ρ is well known to be highly vulnerable to outliers. In the context of meta-evaluation for machine-translation metrics, Mathur et al. (2020) show that this can confound meta-evaluation results.

Ranking-based coefficients instead discard the numerical information in P and H and only focus on the correctness of the *ranking* of the individual instances. Both **Spearman's** ρ and **Kendall's** τ are ranking-based coefficients.

Spearman's ρ is computed as the Pearson's correlation over the ranks of instances in P and H . While it has found some use in metric meta-evaluation (Lin, 2004a), a problem with Spearman's ρ is that it does not work well in the presence of ties in the input data. If human data is derived using Likert scores, which is a common choice (see Chapter 3), ties are bound to be frequent. As a consequence, the more commonly used metric in meta-evaluation is Kendall's τ .

Kendall's τ is based on computing the number of *pairs* of instances that are ranked the same in H and P . Pairs that are ranked the same in H and P are called *concordant* pairs, whereas pairs that are ranked incorrectly are called *discordant*. In its simplest form, it can be computed as

$$\tau(H, P) = \frac{n_c(H, P) - n_d(H, P)}{\text{pairs}(n)}, \quad (4.4)$$

$$\text{pairs}(n) = \frac{n(n-1)}{2}, \quad (4.5)$$

where $n_c(H, P), n_d(H, P)$ are the number of concordant and discordant pairs respectively and $n = |S||D|$ is the total number of samples. For system-level correlation, the computation proceeds in the same way, except for the number of samples, which becomes $n = |S|$.

The above formulation still does not account for tied data points. The τ_b variant deals with this problem by adjusting the denominator to account for the presence of ties:

$$\tau_b(H, P) = \frac{n_c(H, P) - n_d(H, P)}{\sqrt{(\text{pairs}(n) - \text{ties}(H))(\text{pairs}(n) - \text{ties}(P))}}, \quad (4.6)$$

where $\text{ties}(X)$ computes the number of tied scores in a set of ratings X . In the remainder of this work, we will use τ_b exclusively.

The use of τ_b for tie handling is not entirely uncontroversial. Deutsch et al. (2023) argue that many τ variants, including τ_b , are prone to exploitation by measures producing many ties. They propose to instead use pairwise accuracy, where a measure is rewarded for correctly predicting the ranking between two instances, *including ties*. This is naturally problematic for measures that assign continuous scores, since they struggle to exactly tie instances. For these cases, Deutsch et al. propose a tuning procedure, that determines some threshold ϵ so that the pairwise accuracy is maximized if instances with a score difference less than ϵ are treated as tied. We opt not to include this procedure in our analysis, since we investigate CMs that output continuous scores and are thus unlikely to exploit tie behaviour, whereas the proposed procedure would introduce additional complexity. Furthermore, Perrella et al. (2024) show that the calibration procedure can overestimate the performance of continuous metrics, since it leaks information about the number of ties in the input dataset.

Besides the pairwise score of Deutsch et al. (2023), two more variants of meta-evaluation metrics are worth mentioning in the context of this chapter.

First, we can consider the ranking of pairs, either of summaries generated on the same document or of average system performance, as a classification task and use **pairwise accuracy**. In some work (Dang and Owczarzak, 2009a; Dang and Owczarzak, 2009b; Owczarzak and Dang, 2010; Owczarzak and Dang, 2011; Owczarzak et al., 2012), pairwise accuracy is additionally limited to pairs that exhibit statistically significant differences according to some statistical test. This is usually conducted on system-level scores but could, with a sufficient number of per-summary annotations, also be done for summary-level scores.

Finally, if the target quality can be expressed as a per-instance binary classification task, classification metrics such as balanced accuracy or area under the receiver operating characteristic curve (ROC AUC) can be used as EMs. This is

particularly relevant for faithfulness (Honovich et al., 2022). Since the measures we consider in this chapter do not fall in this category, we defer a discussion of this to Chapter 5, where we use it to evaluate our own faithfulness measure.

Regardless of which EM we select, we usually want to provide a confidence interval (CI) around the point estimate for the EM score to account for uncertainty due to the limited size of our meta-evaluation dataset. A pitfall here are the dependencies between scores for summaries that have been generated from the same input document or by the same summarizer. Similar to the inflated Type I error rates for significance tests discussed for human evaluation in Chapter 3, this can lead us to underestimate the width of the CIs when not accounted for. To remedy this, Deutsch et al. (2021b) propose to compute CIs using a modified bootstrap resampling algorithm:

1. Sample with replacement a multi-set of documents \tilde{D} with size $|D|$.
2. Sample with replacement a multi-set of summarizers \tilde{S} with size $|S|$.
3. Construct \tilde{H}, \tilde{P} by selecting values corresponding to each possible pair of summarizers and documents $d \in \tilde{D}, s \in \tilde{S}$.
4. Compute the EM $f(\tilde{H}, \tilde{P})$.
5. Repeat steps 1-4 n times for some sufficiently large n to accumulate scores M .
6. Compute the confidence interval from the corresponding percentiles of M .

We employ their method in this chapter.

4.2.2 Measuring Coherence

While we have already used coherence operationally in the form of annotator instructions in Chapter 3, we will now take a closer look at how coherence can be modelled computationally. In the interest of brevity, we will focus on giving a broad overview of the field, leaving a more in-depth technical description of the CMs we actually evaluate to Section 4.6.

Coherence is the mechanism by which individual sentences form a unified text, as opposed to an unconnected assembly of sentences (Halliday and Hasan, 2013).

Coherence can be divided into *global* coherence (Mann and Thompson, 1988), which is the way individual segments in a text form the overall discourse, and *local* coherence (sometimes also called *cohesion*), which focuses on coherence within a single discourse segment (Poesio et al., 2004). For summarization, local coherence is of greater interest due to the short length and simple structure of most automatically generated summaries. Halliday and Hasan (2013) propose that local coherence arises from the presence of *ties* between individual sentences in a text. Ties arise when we need information from one sentence to interpret another sentence. Halliday and Hasan identify two categories of cohesion: grammatical and lexical. They further subdivide grammatical cohesion into four subcategories. We give a brief overview of each one:

Reference Reference arises when a sentence directly refers to a thing or entity.

A common cohesive element, that is also highly relevant in computational coherence modelling, are *anaphoric* references, such as *he* in the following example:

- Peter is busy. He is writing his thesis.

References can, however, also be *cataphoric* (i.e. to the following text). Halliday and Hasan also discuss *exophoric* references, which are references to context in which the text is produced.

Substitution Substitution is the replacement of one element in the text by another. This is best illustrated with an example:

- I have not started working on this chapter. I am still finishing the previous one.

Here, *one* is substituted for *chapter* in the second sentence. Halliday and Hasan differentiate nominal, verbal, and clausal substitution.

Ellipsis Ellipsis is the substitution of an element by nothing.

- I wrote two pages today. Tomorrow I will write two more.

Here, *two more* omits *pages*, which must be inferred from the first sentence.

Conjunction Conjunctions can also have a cohesive effect in a text, as in the following example:

- The discussion is not complete. However, the introduction is finished.

This is an example of an *adversarial* conjunction.

Lexical cohesion, on the other hand, arises from the repetition of words across sentences. This includes verbatim or (near-)synonymous repetition, as well repetition in the form of hypernyms or what Halliday and Hasan call *general nouns*. We illustrate this in the following text:

- I am writing my thesis. Soon the dissertation is complete. I can then submit the thing.

Here, *dissertation* is a synonymous repetition, followed by a reference via the general noun *thing*.

Another influential theory of local coherence, which has had great impact on computational coherence models, is *centering theory* (Grosz et al., 1995). Centering theory identifies the repeated mention of entities across sentences as a key contributor to local coherence. It posits that at any given utterance u_t , there is a set of so-called *forward-looking centers*, i.e. the entities mentioned in the utterance u_t , each of which can become the focus of discourse in the following utterance u_{t+1} . Forward-looking centers are ranked according their salience in the discourse. The most salient center is referred to as the *preferred center*. The most highly ranked forward-looking center that is realized in a subsequent utterance u_{t+1} is referred to as the *backward-looking center* of u_{t+1} (Poesio et al., 2004). The changes in preferred and backward-looking centers form a set of transitions: *Continuations*, where both the preferred and backward-looking centers remain the same across adjacent utterances; *Retains*, where the preferred center changes; and *Shifts*, where the backward-looking center changes. One key claim of centering theory is that sequences of continuations are preferred over sequences of retains, which are in turn preferred over shifts (Grosz et al., 1995). The frequency of the different transition types can be used to predict the coherence of a given text.

Having established some fundamental theory of coherence, we will now turn to how it can be modelled computationally. Summary coherence modelling is studied in two contexts. In the development of *general coherence models*, it is used as a downstream evaluation task (Barzilay and Lapata, 2008; Tien Nguyen and Joty, 2017; Mesgar et al., 2021), similar to other coherence-related tasks such as essay scoring (Jeon and Strube, 2020, among others) and readability assessment (Mesgar and Strube, 2015, among others). In this context, evaluation is often conducted on a DUC 2003-based dataset originally created by Barzilay and Lapata (2008). In *linguistic quality modelling*, coherence is modelled alongside other linguistic quality dimensions with the goal of creating practical evaluation measures. This leads to a divergence in evaluation between the two strands of research, which motivates our meta-evaluation. We will now give an overview of the work in both areas.

General Coherence Models

In general coherence modelling, an influential line of research derives from the so-called **Entity Grid** (Barzilay and Lapata, 2008). Motivated by the concepts of continuations, shifts, and retains from centering theory, the entity grid tracks local transitions in the mentions and grammatical roles of entities. These occurrence patterns can be derived from a text by first building a matrix where rows correspond to sentences and columns to entities (the namesake *grid*). The grid cells indicate for each entity in which sentences it appears and, optionally, in which grammatical role. Patterns are derived by choosing a window size and then counting the frequency of the different possible “ n -grams” in the columns. The resulting feature vector, along with labelled examples of coherent and incoherent texts, can be used as input to a machine learning algorithm to derive a coherence model. Barzilay and Lapata use a pairwise learning setup to learn to rank coherent and incoherent texts.

The entity-focused approach of Barzilay and Lapata has been refined in a number of ways. Elsner and Charniak (2011) extend the features derived from the entity grid with entity-specific features, such as if it is a proper noun and its named entity type. Feng and Hirst (2012) focus on improving the learning

scheme by extending the original pairwise learning approach of Barzilay and Lapata. Instead of using coherent and incoherent samples only, they propose to use the similarity of an incoherent sample to a coherent sample to rank incoherent samples among each other. For summary evaluation specifically, they propose to exploit the similarity to a human reference summary as a ranking signal. With the advent of neural methods, Tien Nguyen and Joty (2017) replace the feature-extraction-based approach of prior work with a convolutional neural network over the entity grid.

Beginning with Guinaudeau and Strube (2013), another branch of research casts the entity grid as an **entity graph**. This follows the observation that the entity grid can be considered as the incidence matrix of a bipartite graph, which can subsequently be projected into a graph of sentences, where individual sentences are connected if they share an entity. Guinaudeau and Strube use this to derive an unsupervised coherence score based on graph connectivity. Mesgar and Strube (2015) introduce the idea of modelling frequent subgraphs into this framework, which they dub *coherence patterns*. The number of occurrences of different coherence patterns can be used as a feature for learning approaches, similar to entity transitions.

Coherence measures can also operate entirely on the *lexical* level. Mesgar and Strube (2016) propose a lexicalized variant of the entity graph, where sentences are connected based on their maximum embedding similarity. Mesgar and Strube (2018) extend this similarity-based approach to contextualized embeddings. Joty et al. (2018) lexicalize the neural entity grid by enhancing entity roles with embeddings for each mention. Moon et al. (2019) propose a fully lexicalized coherence model combining sentence-level and global embeddings. Mesgar et al. (2021) combine the entity graph with lexical representations in a graph neural network.

With the exception of the entity graph and the entity grid variant of Elsner and Charniak (2011), which uses a generative model learned only from coherent documents, all models require examples of coherent and incoherent documents as training data. Where sufficient data is available, this can be done in a supervised fashion. The DUC 2003 coherence dataset of Barzilay and Lapata (2008),

for example, contains a training set of about 144 summaries with pairwise preference information.¹ However, this is often insufficient for recent neural models. Alternatively, CMs can be trained in a weakly supervised fashion on the so-called *shuffle task* (Barzilay and Lapata, 2008). In the shuffle task, an input document is divided into individual sentences, which are then randomly reordered to form a less coherent document. These pairs of original and shuffled documents can then be used as a training signal, as well as for evaluation (Barzilay and Lapata, 2008; Tien Nguyen and Joty, 2017; Mesgar et al., 2021; Moon et al., 2019). Laban et al. (2020) argue that the shuffle task is a bad proxy for evaluation, since it can be solved near-perfectly by a RoBERTa-based (Liu et al., 2019b) classifier trained on the shuffle task. They suggest to exclusively use the task in a zero-shot setting, as well as to shuffle blocks of sentences, instead of individual sentences, to increase task difficulty. We will test their hypothesis that the shuffle task leads to poor coherence models in this chapter.

Automatic Linguistic Quality Estimation

Pitler et al. (2010) note that the DUC 2003 derived data used in evaluation of general coherence models is potentially misleading for evaluation of summary coherence due to the mix of weak automatic and human summaries in the evaluation data. They develop a set of measures for the five linguistic quality dimensions in the DUC 05-07 shared tasks, previously discussed in Chapter 2. They investigate the use of several features as input to a machine learning setup trained on DUC 2006 data. For coherence, they propose counting the frequency of cohesive devices, as well as coreference, and word similarity features. Xenouleas et al. (2019) also train a supervised model on DUC data, although they replace feature engineering with BERT-based (Devlin et al., 2019) representations.

There are also entirely unsupervised approaches to coherence evaluation. Zhu and Bhat (2020) propose a weighted sum of heuristic measures to derive an overall summary score. Yuan et al. (2021) use the probability assigned to the output by a conditional language model conditioned on the input document for evaluation.

¹The original annotation process uses a seven point Likert scale to get individual scores for each summary, but this is converted into pairwise preferences.

Finally, summary coherence is sometimes modelled using measures that make use of human-written reference summaries. Fabbri et al. (2021b) benchmark several reference-based evaluation measures, including ROUGE, on coherence annotations in their SummEval dataset and find moderate system-level correlation for the top performers. Zhao et al. (2023) propose DiscoScore to integrate coherence into reference-based evaluation. DiscoScore is based on tracking discourse focus, i.e. the entities that hold the reader’s attention at any given point in the text. DiscoScore rewards a summary that has similar discourse focus to the reference. We do not include reference-based measures in our study, since their dependence on high-quality reference summaries makes them fundamentally less flexible than reference-free measures.

4.3 Related Work

With regard to meta-evaluation of coherence measures, we are only aware of a single comparable effort by Mohiuddin et al. (2021), who conduct a comparative study of five CMs. They study 10 summaries each from 4 recent summarizers as well as instances from the DUC 2003 dataset of Barzilay and Lapata (2008). Since – unlike our work – they reuse the outdated summaries in the DUC 2003 dataset, their evaluation does not necessarily generalize to recent summarization systems.

In concurrent work on meta-evaluation methods, Deutsch et al. (2022) propose to improve system-level correlation. They show that the variance in system-level scores can be reduced by computing average predicted scores P not only on documents with human rated instances D but instead on the entire corpus $D^* \supseteq D$. Since, after averaging, this leads to a more accurate estimate of the predicted score P_s for each summarizer $s \in S$, the resulting correlation has lower variance. Additionally, Deutsch et al. (2022) propose to only compute system-level correlation on pairs of summarizers s_1, s_2 where the difference between their respective human scores $|H_{s_1} - H_{s_2}|$ is below some threshold. This ensures measures are not rewarded for getting “easy” comparisons right. Their approaches are complementary to our analysis in that they look at informativeness instead of coherence and do not address the shortcomings of system-level correlation in the presence of system-level confounders.

Also concurrently, Pagnoni et al. (2021) propose *partial correlation* for meta-evaluation. For human and predicted scores H, P , partial correlation is computed by first determining residuals $\epsilon^{(H)}, \epsilon^{(P)}$ as follows:

$$\epsilon_{(d,s)}^{(H)} = H_{(d,s)} - \bar{H}_s, \quad (4.7)$$

$$\epsilon_{(d,s)}^{(P)} = P_{(d,s)} - \bar{P}_s, \quad (4.8)$$

where \bar{H}_s, \bar{P}_s are the mean human and predicted scores for system s .² The partial correlation is then determined by computing the correlation of the residuals ϵ_H and ϵ_P .

This tackles the problem of system-level confounders, similar to our intra-system correlation. Unlike our approach, however, it does not have an intuitive correspondence to a particular set of comparisons between instances and does not allow for detailed inspection of correlations for individual systems.

Also concurrently, Durmus et al. (2022) identify spurious correlates in faithfulness datasets. They find contemporary measures for faithfulness are outperformed by measures for the extractiveness of summaries, as well as a combination of length and the perplexity of a strong language model on the input. This suggests that our methods will be useful beyond coherence evaluation.

4.4 Selecting a Meta-Evaluation Dataset

As discussed in Section 4.1, data that is suitable for meta-evaluation is challenging to find. We identify three properties that a dataset requires to be a viable choice for our meta-evaluation study:

1. Reliable judgements at the summary level
2. A sufficient number of different summarization systems
3. Output from recent summarization systems

²More generally, the residuals $\epsilon^{(H)}, \epsilon^{(P)}$ are derived from a linear regression of H (P) and the confounding variable(s). We state the specific case of meta-evaluation for clarity.

In addition, for our specific use case of summary coherence evaluation, datasets must of course also cover our desired quality dimension.

To illustrate these challenges, consider that the dataset we have gathered in Chapter 3 fulfills neither the first nor the second desideratum: It contains only five summarization systems and is annotated by untrained crowd workers, who are unreliable on the summary level. As discussed previously, this is a deliberate choice, since it mirrors common setups of studies conducted for cost-efficient evaluation. Consequently, most human evaluation data is difficult to use for meta-evaluation. Datasets for meta-evaluation are thus usually derived either from shared tasks, where a large number of summarizers is evaluated using the same evaluation protocol, or alternatively from large-scale meta-evaluation studies.

For early summarization systems, the shared tasks of DUC-2005 to DUC-2007 (Dang, 2005; Dang, 2006; NIST, 2007) conferences are one source of such data, since they cover a large number of summarizers rated by trained assessors. They include both scores for content, as well as linguistic quality, including coherence. The successor Text Analysis Conferences (TAC) included a shared meta-evaluation task called *Automatically Evaluating Summaries of Peers* (AESOP) (Dang and Owczarzak, 2009b; Owczarzak and Dang, 2010; Owczarzak and Dang, 2011) from 2009 to 2011, also using professional summary ratings, although only for content and later for content and readability. The coherence evaluation dataset of Barzilay and Lapata (2008) uses outputs from five summarizers, plus reference summaries, from 16 input documents from DUC 2003 (Over and Yen, 2003) with human pairwise coherence judgements.

However, all of these datasets naturally cover only extractive summarization systems that were proposed before the advent of abstractive neural models. Given the shift in the methods employed in the field, performance on these datasets does not necessarily permit conclusions about the utility of a measure on summaries of future systems. For content measures, this problem is demonstrated by Peyrard (2019), who show using simulation methods that while commonly used (reference-based) evaluation measures, such as ROUGE (Lin, 2004b), agree with each other on weak summaries, they start disagreeing with each other on high-scoring ones. This suggests that studies conducted on weak summarizers, such as those present in the previously mentioned datasets, do not provide much information about

the ability to judge strong summarizers. In an empirical evaluation, Bhandari et al. (2020) compare meta-evaluation results on TAC data from 2008 and 2009 with results on a newly introduced dataset they dub RealSumm. RealSumm uses LitePyramid (Shapira et al., 2019), which we previously discussed in Chapter 2, on summaries generated by 25 summarizers on CNN/DM (Hermann et al., 2015). They find both datasets lead to different conclusions. While both studies focus exclusively on reference-based content evaluation, they demonstrate the risk of transferring meta-evaluation results from datasets using outdated summarizers. This motivates the third desideratum we give above.

In addition to the mentioned RealSumm, we are aware of two other datasets that are suitable for meta-evaluation in summarization: SEAHORSE (Clark et al., 2023) and SummEval (Fabbri et al., 2021b). SEAHORSE contains output from a total of eight summarizers, plus references, across 32,366 input documents. Unlike the other datasets mentioned here, SEAHORSE input documents are sampled from multiple datasets across different languages. However, while it is annotated by trained workers and contains linguistic quality dimensions, it does not include coherence judgements. In this work, we thus use the **SummEval** dataset. SummEval is based on 100 CNN/DM input documents, with summaries sampled from 17 summarizers.³ SummEval contains human annotations for four dimensions along a five-point Likert scale: Coherence, Consistency, Fluency, and Relevance. This makes it uniquely suitable for our investigation. SummEval contains both expert and non-expert judgements. For our work, we only use expert judgements to ensure reliability at the summary level.

For completeness, we note that Koto et al. (2022) also create a (crowd-sourced) dataset for, among other dimensions, summary coherence evaluation. However, their dataset only encompasses output from two models, which makes it less suitable for our purposes. Similarly, the work of Grusky et al. (2018) contains crowd-sourced annotations for 60 input documents summarized by seven summarizers for four categories, including coherence, again yielding data too small to be suitable for our meta-evaluation.

³There is a variant of SummEval with 16 summarizers, in which judgements from *Pegasus (dynamic mix)* are not included.

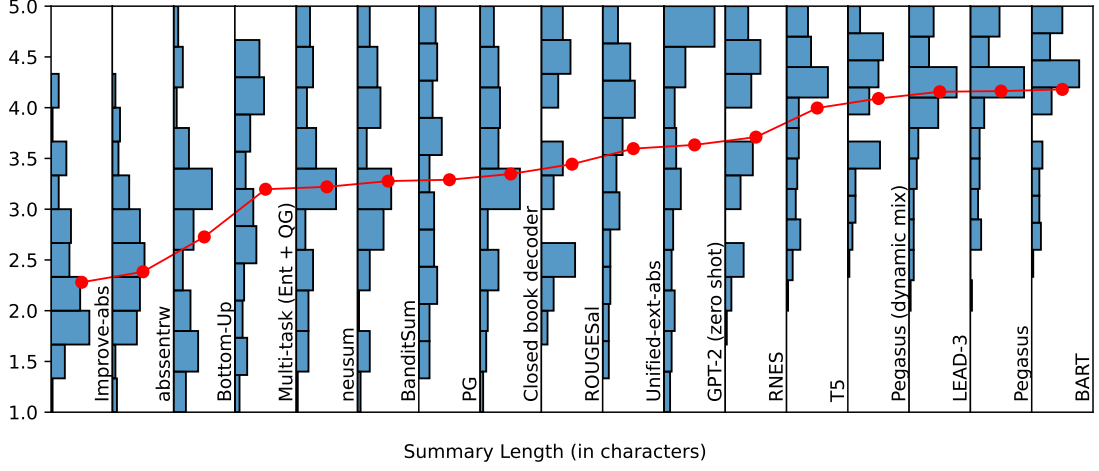


FIGURE 4.1: Distribution of human coherence scores for the 17 systems in the SummEval dataset. The red dots indicate the mean score of each system.

Figure 4.1 highlights two important properties of SummEval for our further analysis: Firstly, there is a large gap in average performance between different summarizers, and secondly, most summarizers exhibit considerable variance in scores.

4.5 A new Meta-Evaluation Metric

As discussed in Section 4.2.1, meta-evaluation is typically conducted on a set of summaries generated on document set D by a set of summarizers S using the correlation of predicted scores $P = \{P_{(d,s)} | d \in D, s \in S\}$ with human judgements $H = \{H_{(d,s)} | d \in D, s \in S\}$. There are three common variants:

System-level Correlation τ_{sys} assesses measure performance by correlating the *mean* human and *mean* measure scores of the individual summarizers.

Pairwise Accuracy Acc_{pair} assesses measure performance by comparing scores on outputs of two different systems on the same document.

Summary-level Correlation τ_{sum} compares scores on all generated summaries.

The correlation function used is usually Kendall’s τ_b , while the pairwise metric is usually reported as accuracy. It is easy to see from its definition that if there are no tied values and the set of pairwise comparisons is complete, Kendall’s τ and accuracy are equivalent with the only difference being the range shift from $[0, 1]$ to $[-1, 1]$. This allows us to also measure the pairwise accuracy metrics as Kendall’s τ for consistency.

We can now specify all three EMs in terms of the set of pairwise comparisons C they consider, where $C_{sys} \subset 2^{S \times S}$ considers comparisons between averaged system scores and $C_{pair}, C_{sum} \subset 2^{(D \times S) \times (D \times S)}$ consider comparisons between individual summary scores:

$$C_{sys} = \{(s_i, s_j) | s_i \neq s_j\}, \quad (4.9)$$

$$C_{pair} = \{((d, s_i), (d, s_j)) | s_i \neq s_j\}, \quad (4.10)$$

$$C_{sum} = \{((d_k, s_i), (d_l, s_j)) | (d_k, s_i) \neq (d_l, s_j)\}. \quad (4.11)$$

The EMs pose different demands to measures: system-level correlation requires a correct ranking of systems according to their average score. Pairwise accuracy requires correct ranking of summaries from different systems but only between summaries produced on the same document. Finally, summary-level correlation requires the correct ranking of any pair of summaries.

4.5.1 A new Evaluation Metric: Intra-System Correlation

All three EMs focus on comparisons between summaries generated by different summarizers. For system-level and pairwise evaluation this arises by construction, whereas for summary-level correlation it is contingent on the dataset structure: On SummEval, less than 6% of comparisons for τ_{sum} are between summaries of the same summarizer. We argue that this gives an incomplete view of CM performance for the following reasons:

1. SummEval covers summarizers with widely different performance levels (see Figure 4.1), leading current EMs to include many easy decisions. This is unlikely to reflect real-world evaluation of competitive summarizers.

2. While system-level evaluation is often the primary use case, measures can also be used in a reranking or ensembling context to select the highest quality summary from a set of candidates. In these situations, summaries are likely to be generated either by the same summarizer or a set of similarly (high) performing summarization systems. In these cases, system-level EMs offer only limited insight into likely measure performance, since they primarily measure the ability to discriminate between *different* systems with potentially large performance gaps.
3. EMs might not correlate with the target quality per se but instead with *system-level confounders* that are unlikely to generalize to new systems and settings. We elaborate on this in Section 4.5.2.

We thus suggest adding a new EM, **Intra-system Correlation** τ_{intra} , which we define on comparisons between summaries generated by the same system. This corresponds to considering the following pairs $C_{intra} \subset 2^{(D \times S) \times (D \times S)}$:

$$C_{intra} = \{((d_k, s), (d_l, s)) | d_k \neq d_l\}. \quad (4.12)$$

It neatly complements pairwise accuracy, as it is essentially the same computation but keeps the summarizer constant instead of the document. Intuitively, this measure both contains far fewer "easy" decisions and is much more resilient to any system-level confounders in the data. We use the average of the intra-system correlation of all systems as the EM.

4.5.2 System-level Confounders

We assess how EMs behave in the presence of system-level confounders on our coherence data. To this end, we investigate two summary features that are unlikely to be generalizable CMs but lead to surprisingly strong correlations on SummEval: capitalization and summarizer architecture.

For capitalization, we count the number of uppercase letters in each summary. This is a purely system-level heuristic, since only three of the 17 summarizers in SummEval produce capital letters.⁴ For architecture, we assign a score of 1 to each

⁴BART, GPT-2 (zero shot), and Pegasus (dynamic mix)

	Cap.	Cap. (r)	Arch.	Arch. (r)	UB	UB (r)
τ_{sys}	0.42	0.23	0.58	0.37	1.00	1.00
τ_{sum}	0.19	0.11	0.31	0.20	0.39	0.39
τ_{pair}	0.21	0.14	0.33	0.22	0.44	0.44
Acc_{pair}	0.23	0.57	0.34	0.62	0.73	0.73
τ_{intra}	-	-0.03	-	0.01	-	0.00

TABLE 4.1: Results for the confounders and upper bound. τ_{intra} for the non-random variants is undefined, as scores within each system are constant. Scores for the random variants (r) are averaged over 100 runs.

summary from one of the five summarizers that are derived from pretrained transformers in some fashion⁵ and 0 to all others. Neither of the two confounders can, by construction, be a reasonable and generalizable CM. Additionally, we compute an “upper bound” (UB) that assigns to each summary the mean human score of the system that produced the summary. It simulates perfect system ranking but no ability to correctly rank summaries within each system. Since these procedures result in many ties, we also compute a second variant of each confounder where we add small noise to each score. This prevents τ_b from profiting from these ties while preventing accuracy from unfairly suffering.

Table 4.1 shows the resulting correlations. We find all confounders achieve noticeable correlation with human scores under some EMs. In particular, system-level correlation comes close to or exceeds the best CM reported originally for SummEval (CHRF (Popović, 2017), 0.40). In contrast, using intra-system correlation, the problems of these pseudo-measures become easily apparent. These results show that, at least in the SummEval dataset, substantial correlation on both the system and summary level can be achieved by modelling proxies that are unlikely to actually correspond to coherence. In practical scenarios, system-level correlation might be a mix of modelling coherence and reliance on confounders. Intra-system evaluation is an important tool in this context as it is robust to system-level confounders, which increases the chance of it modelling generalizable information about the quality dimension in question.

⁵BART, Pegasus, Pegasus dynamic Mix, T5, and GPT-2

	EEG	EGR	NEG	UNF	GRA	CCL	SQE	GRU	BAS
Unsupervised	✓ ^(a)	✓						✓	✓
Shuffle			✓	✓	✓	✓			^(b)
Supervised (DUC 03)			✓		✓				
Supervised (DUC 05-07)							✓		

TABLE 4.2: Training settings for the CMs under investigation. (a) The extended entity grid estimates the multinomial distribution of an entity’s role given its prior occurrences. While this needs a dataset to estimate the distribution, it cannot be trained as a classifier. (b) The pretraining of BART includes a task where documents are corrupted by shuffling the sentence order and must be reconstructed correctly.

4.6 Coherence Measures

With the meta-evaluation metrics in place, we now turn to selecting coherence measures for our evaluation study.

In Section 4.2.2, we have discussed how, in prior work, there is a divide in research on general coherence modelling and linguistic quality modelling. Thus, to structure our selection of CMs and to allow for better comparison between both strands of research, we start by categorizing CMs based on their training settings. For both strands, we can divide models into **supervised CMs**, that are trained on human coherence ratings of summaries, and **unsupervised CMs**, which do not require human ratings. For general coherence models, supervision data is typically the pairwise DUC 2003 data of Barzilay and Lapata (2008), while for linguistic quality modelling, the typical setting is regression on DUC 05-07 ratings (Dang, 2005; Dang, 2006; NIST, 2007). Additionally, many general coherence models can be trained on the shuffle task. We refer to CMs trained in this manner as **self-supervised CMs**. To cover a wide range of diverse CMs, we include models from both strands of research across all applicable training settings. Table 4.2 gives a full overview of all CMs and their training settings.

For general coherence models, we select both entity-based models, that are derived from the entity grid representation, as well as lexical models, that directly take the summary as input. As representatives of entity-based general CMs, we select the Extended Entity Grid (EEG) (Elsner and Charniak, 2011) and the Entity

Graph (EGR) (Guinaudeau and Strube, 2013) as unsupervised, theoretically motivated measures. We also include the neural entity grid (NEG) (Tien Nguyen and Joty, 2017) as a more recent formulation of the entity grid. NEG can be trained both in the supervised setting on the pairwise DUC 2003 data and in the self-supervised setting using a ranking loss.

As lexical general CMs, we select the unified coherence model (UNF) (Moon et al., 2019) and the graph-based model neural coherence model of Mesgar et al. (2021) (GRA) as state-of-the-art (at the time of the original publication of this work) CMs. Both can be trained in the self-supervised setting with a ranking loss. Additionally, GRA permits training of a supervised model on DUC 2003 data.

Finally, we include a self-supervised RoBERTa (Liu et al., 2019b) model, that is fine-tuned on the shuffle task as a classifier (CCL). This follows the observation of Laban et al. (2021) that this can outperform more sophisticated coherence models at the shuffle task.

We train all self-supervised models on the WSJ corpus of newswire articles, which is frequently used in coherence modelling (Elsner and Charniak, 2011; Guinaudeau and Strube, 2013; Moon et al., 2019; Mohiuddin et al., 2021). We also train models using the same technique on reference summaries from the train portion of CNN/DM. For EEG, which uses a generative model, we also estimate parameters on both datasets. For WSJ, we follow the original implementations regarding the number of shuffled samples. For CNN/DM, we only use a single shuffled instance per summary, as it is larger by two orders of magnitude (WSJ: 1,400; CNN/DM: 287,113 documents before shuffling).

For the linguistic quality modelling strand of research, we select SumQE (SQE) (Xenouleas et al., 2019) as a supervised regression-based model, which is trained on regression data from the DUC 05-07 conferences. Finally, we include BARTScore (BAS) (Yuan et al., 2021) and GRUEN (GRN) (Zhu and Bhat, 2020) as the state-of-the-art unsupervised summary quality measures.

To anchor our scores, we include an upper and lower bound: RND assigns each summary a uniformly chosen score between 0 and 1, which establishes a lower bound for the CMs. To establish a realistic upper bound, we simulate what scores the human annotators would receive (HUM). We use the SummEval human annotations and select the annotator with the worst overall correlation to the remaining

annotators and use their scores as predictions.⁶

We will now give a technical description of the CMs. Detailed accounts of our experimental settings for each CM can be found in Appendix C.

4.6.1 Entity Grid and Extensions

To construct an entity grid of a given document consisting of sentences $S = [s_1, s_2, \dots, s_n]$, we first identify all entities $E = \{e_1, e_2, \dots, e_m\}$ in the document via coreference resolution. In practice, all entity-based models we consider in this chapter use the entity grid implementation of Elsner and Charniak (2011) (also known as the Brown Coherence Toolkit) to construct their grids. Here, coreference resolution is approximated with lexical overlap between noun phrases. The entity grid is then a matrix $M = |S| \times |E|$, where M_{ti} indicates the role of entity e_i in sentence s_t . The role is determined by the grammatical role of e_i in s_t : subject (S), object (O), other (X), or no occurrence (\emptyset).

Multiple approaches have been suggested to derive a coherence score from M . Barzilay and Lapata (2008) extract 3-grams from the columns of M to estimate role transition probabilities in the document. These can be used as features for a support vector machine (Cortes and Vapnik, 1995), optionally along with the salience of the entity as determined by its frequency in the document.

Extended Entity Grid (EEG)

In this work, we use a more recent formulation of Elsner and Charniak (2011), the extended entity grid, which instead uses a generative model. They compute the probability of an entity taking role r in the sentence at position t given its two preceding roles on a training corpus: $p(M_{ti} = r | M_{(t-1)i}, M_{(t-2)i})$. As an extension over the entity grid, they also include a number of entity-specific features: Whether the identified chain contains a proper noun mention, the named entity label of the entity, the number of modifiers in the chain, whether the chain contains a singular mention, and a number of features designed to correct coreference resolution errors. The coherence score of a summary is then the (log) probability of its entity grid

⁶We note that, unlike automatic measures, humans may only differentiate among five classes. We might thus underestimate actual human performance.

M , normalized by the grid size:

$$\text{score}(M) = \frac{1}{|S||E|} \sum_t \sum_i \log p(M_{ti} | M_{(t-1)i}, M_{(t-2)i}, F_{ti}), \quad (4.13)$$

where F_{ti} indicate the entity-specific features and entity salience for cell ti . In practice, this model can be learned using multinomial logistic regression on coherent texts. To keep the model comparable with those that are trained on the shuffle task, we use the coherent examples from WSJ and CNN/DM for training, respectively.

Entity Graph (EGR)

To build the entity graph (EGR) of Guinaudeau and Strube (2013), we treat M as the incidence matrix of a bipartite graph G between entities and the sentences they occur in, i.e. $G = (S \cup E, \{(s_t, e_i) | M_{ti} \neq \emptyset\})$. Instead of relying on a learned distribution of transitions, EGR first projects the bipartite graph G into a directed, weighted graph $G^* = (S, A^*, w)$ with arcs A^* connecting sentences S with weights given by a weight function w . The coherence score is derived by computing the average sentence centrality in the resulting graph, following the intuition that a more coherent document is one that is more tightly connected.

Guinaudeau and Strube propose three one-mode projections from G to G^* , which differ in the edge weight function w . All of them construct edges between all sentences that share at least one entity, with the edge direction following sentence order in the text. Let $\mathcal{N}_G(s_t)$ be the set of entities linked to s_t in G (i.e. its neighbours), then

$$A^* = \{(s_t, s_u) \mid |\mathcal{N}_G(s_t) \cap \mathcal{N}_G(s_u)| > 0, u > t\}. \quad (4.14)$$

The choice of weight function w allows us to influence the strength of association between linked sentences. w_U equally weights all edges in A^* :

$$w_U(s_t, s_u) = 1. \quad (4.15)$$

This ignores the amount of entity overlap between s_t and s_u , which motivates the introduction of w_W , which computes the edge weight based on the number of entities shared between s_t and s_u :

$$w_W(s_t, s_u) = |\mathcal{N}_G(s_t) \cap \mathcal{N}_G(s_u)|. \quad (4.16)$$

Finally, Guinaudeau and Strube find that integrating syntactic role information, as in the entity grid, leads to increased performance on a number of coherence modelling tasks. They thus introduce w_{Acc} , which additionally uses role information to increase the weight of edges involving subject or object roles:

$$w_{Acc}(s_t, s_u) = \sum_{e_i \in \mathcal{N}_G(s_t) \cap \mathcal{N}_G(s_u)} w_r(M_{ti}) \cdot w_r(M_{ui}), \quad (4.17)$$

where w_r is an auxiliary weight function for entity roles. We follow Guinaudeau and Strube by setting the weight of subjects to 3, of objects to 2 and all other roles to 1.

As described so far, the entity graph is permutation invariant. That is, any set of sentences with a given entity overlap would receive the same score. This is clearly not desirable for a coherence measure. Guinaudeau and Strube thus introduce a discounting factor that divides the weight of an arc between sentences s_t, s_u by the distance of their sentences in the text $u - t$.

The final weight function, which we also use in our implementation of **EGR** is then:

$$w_{Acc+Adj}(s_t, s_u) = \frac{1}{u - t} \sum_{e_i \in \mathcal{N}_G(s_t) \cap \mathcal{N}_G(s_u)} w_r(M_{ti}) \cdot w_r(M_{ui}). \quad (4.18)$$

Given G^* , the final score is then the average outdegree of each sentence in the graph:

$$score(S, A^*, w) = \frac{1}{|S|} \sum_{s_t \in S} \sum_{s_u | (s_t, s_u) \in A^*} w(s_t, s_u). \quad (4.19)$$

Neural Entity Grid (NEG)

The final variant of the entity grid we consider is the Neural Entity Grid (**NEG**) (Tien Nguyen and Joty, 2017). **NEG** feeds the entity grid matrix M directly into

a convolutional neural network, replacing manual feature extraction with learned filters.

The architecture associates each possible role in M with an embedding vector $\text{emb}(M_{ti})$ to construct an embedded entity grid $M^{(emb)} \in |S| \times |E| \times d$, where d is the embedding dimension.

NEG then computes the convolution of each column in $M^{(emb)}$ with filter weights $W \in \mathbb{R}^{n_f \times w \times d}$, where n_f is the number of filters and w is the window size. This results in a hidden representation $h_{ti} \in \mathbb{R}^{n_f}$ for each cell t, i in $M^{(emb)}$:⁷

$$h_{tij}(M^{(emb)}) = a \left(\sum_{k=1}^w \langle W_{jk}, M_{(t+k)i}^{(emb)} \rangle + b_j \right), \quad (4.20)$$

where $b_j \in \mathbb{R}$ is a trainable bias term for filter j and a is an activation function, which is set to the rectified linear unit (ReLU).

In the reference implementation, which we use in this chapter, this is implemented by concatenating all rows in $M^{(emb)}$ with $w - 1$ all-zero padding vectors to separate the embeddings for each row and then padding the resulting sequence to a predefined maximum length l_{\max} with more zero vectors.

The features are then max-pooled across time steps in windows of size p , to receive features $\tilde{h} \in \mathbb{R}^{\lfloor l_{\max}/p \rfloor \times d}$. Finally, the features are concatenated to a single $d \cdot \lfloor l_{\max}/p \rfloor$ -dimensional representation and fed into a scoring head:

$$\text{score}(M) = w^{(\text{score})T} [\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_{\lfloor l_{\max}/p \rfloor}] + b^{(\text{score})}, \quad (4.21)$$

where $w^{(\text{score})} \in \mathbb{R}^{d \lfloor l_{\max}/p \rfloor}$ and $b^{(\text{score})} \in \mathbb{R}$ are trainable parameters.

The model is trained using a pairwise ranking loss on instances of coherent and incoherent documents with corresponding matrices M^+, M^- :

$$L(M^+, M^-) = \max(0, \gamma - \text{score}(M^+) + \text{score}(M^-)), \quad (4.22)$$

where γ is a hyper-parameter for the margin between positive and negative examples.

⁷The original paper has b be a position-dependent parameter, which is not consistent with the implementation. We give the formula corresponding to the actual implementation here instead.

4.6.2 Lexical Coherence Models

All previous models have in common that they model coherence exclusively through entity occurrence patterns. This limits them in two important ways:

- They are reliant on external coreference resolution tools to function.
- They are unable to identify any other coherence devices, like lexical overlap.

The rise of efficient neural representations of text has opened the doors to solving this problem by instead directly modelling coherence from raw texts.

Unified Coherence Model (UNF)

The **Unified Coherence Model (UNF)** (Moon et al., 2019) implements this idea by combining both local and global neural representations of document sentences. Given a document with sentences $S = [s_1, s_2, \dots, s_n]$, they first compute a neural representation $h_t \in \mathbb{R}^{d_1}$ for each sentence s_t via a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) network, where d_1 is the representation dimension. They then combine each pair of adjacent sentence representations into a local representation $h_t^{(\text{loc})}$ with

$$h_t^{(\text{loc})} = h_t^T W^{(\text{loc})} h_{t+1} + b^{(\text{loc})}, \quad (4.23)$$

where $W^{(\text{loc})} \in \mathbb{R}^{d_1 \times d_2 \times d_1}$, $b^{(\text{loc})} \in \mathbb{R}^{d_2}$ are trainable parameters. d_2 is the dimension of the pairwise representation. This intuitively models the local coherence patterns in the document but is unable to model global document information.

To model the latter, Moon et al. apply a stack of n_l light-weight convolution layers (Wu et al., 2019) to the sentence representations h . At layer j , the contextualized sentence representation $\hat{h}_t^{(j)} \in \mathbb{R}^{d_1}$ for sentence t is computed as

$$\hat{h}_{tc}^{(j)} = \text{DepthWise} \left(\text{softmax}(W_c^{(j)}), \hat{h}^{(j-1)}, t, c \right), \quad (4.24)$$

$$\text{DepthWise}(w, h, t, c) = \langle w, h_{t:t+w-1, c} \rangle, \quad (4.25)$$

$$\hat{h}_t^0 = h_t, \quad (4.26)$$

where $h_{a:b,c}$ is the concatenation of the c -th entry of the representations for sentences at positions a to b . $W^{(j)} \in \mathbb{R}^{d_1 \times w}$ are the convolution weights of the j -th layer with kernel width w . Weights are shared across rows in groups of a predefined size to reduce the number of parameters.

To derive the global document representation, the contextualized sentence representations are average pooled over time: $h_c^{(g)} = \frac{1}{|S|} \sum_{t=1}^{|S|} \hat{h}_{tc}^{(n_t)}$.

Finally, the global document representation and the pairwise local representations are combined into representations \tilde{h}_t for each window of three adjacent sentences:

$$\tilde{h}_t = [h^{(g)}, h_t^{(\text{loc})}, h_{t+1}^{(\text{loc})}]. \quad (4.27)$$

Naively, $\tilde{h}_{|S|-1}$ and $\tilde{h}_{|S|}$ cannot be computed, since the window representations $h_{|S|}$ and $h_{|S|+1}$ are undefined. The reference implementation lets $h_{|S|+1} = h_{|S|}$ to compute $h_{|S|}^{(\text{loc})}$ and sets $h_{|S|+1}^{(\text{loc})} = h_{|S|}^{(\text{loc})}$.

Finally, UNF computes a local coherence score for each window starting at sentence t :

$$\text{score}(S, t) = \langle w^{(\text{score})}, \tilde{h}_t \rangle + b^{(\text{score})}, \quad (4.28)$$

where $w^{(\text{score})} \in \mathbb{R}^{2d_2+d_1}$ is a trainable weight vector and $b^{(\text{score})} \in \mathbb{R}$ is a trainable bias.

Similarly to NEG, UNF uses a ranking loss between coherent and incoherent documents. However, to allow for direct feedback to the local window representations, their loss is specifically formulated for the shuffle task. Given original sentences S^+ and shuffled sentences $S^- = \text{shuffle}(S^+)$, the loss at window t is

$$L(t) = \max(0, \phi(t) - \text{score}(S^+, t) + \text{score}(S^-, t)), \quad (4.29)$$

$$\phi(t) = \begin{cases} 0 & S_{t \dots t+2}^+ = S_{t \dots t+2}^- \\ \gamma & \text{else} \end{cases}, \quad (4.30)$$

where $\phi(t)$ deactivates the margin γ of the ranking loss for windows where both documents are the same.

Since UNF does not directly produce a global coherence score, we compute the average of the window scores as the final coherence measure.

Graph-based Neural Coherence Model (GRA)

The **Graph-based Neural Coherence Model (GRA)** (Mesgar et al., 2021) combines entity-based representation with lexical information in a graph neural network. It uses the same per-sentence representation approach as UNF to build a sentence representation h_t for each sentence s_t . GRA then uses a graph-based representation inspired by EGR, along with a self-attention network (Vaswani et al., 2017) to build a global document representation.

Their graph representation extends the unweighted G^* from EGR with adjacency arcs and introduces edge labels that differentiate arcs into adjacency and entity arcs. A relational graph convolutional network (RGCN) (Schlichtkrull et al., 2018) then computes contextualized representations for each sentence based on its neighbors in G^* . The graph-contextualized representation $h_t^{(\text{graph})}$ for sentence s_t is

$$h_t^{(\text{graph})} = \sum_{r \in R} \frac{1}{|\mathcal{N}_{G^*}(s_t, r)|} \left(\sum_{j \in \mathcal{N}_{G^*}(s_t, r)} W_r^{(\text{graph})} h_j \right), \quad (4.31)$$

where $\mathcal{N}_{G^*}(s_t, r)$ are the neighbours of s_t in G^* that are connected to s_t with an arc with label r . $W_r^{(\text{graph})} \in \mathbb{R}^{d \times d}$ are trainable parameters for each relation type $r \in R$ (either adjacency or entity). d determines the representation dimensionality.

The graph-based representations are globally contextualized using a self attention network:

$$h_t^{(\text{attn})} = \sum_u \alpha_{tu} h_u^{(\text{graph})}, \quad (4.32)$$

$$\alpha_{tu} = \frac{e^{w_{tu}}}{\sum_{u'} e^{w_{tu'}}}, \quad (4.33)$$

$$w_{tu} = \left\langle W^{(\text{query})} h_t^{(\text{graph})}, W^{(\text{key})} h_u^{(\text{graph})} \right\rangle, \quad (4.34)$$

where $W^{(\text{query})} \in \mathbb{R}^{d \times d}$, $W^{(\text{key})} \in \mathbb{R}^{d \times d}$ are trainable parameters.

Finally, the representations $h^{(\text{attn})}$ for each sentences are averaged and projected into a scalar score for the entire document:

$$\text{score}(S) = \left\langle w^{(\text{out})}, \frac{1}{|S|} \sum_t h_t^{(\text{attn})} \right\rangle + b^{(\text{out})}, \quad (4.35)$$

where $w^{(\text{out})} \in \mathbb{R}^d, b^{(\text{out})} \in \mathbb{R}$ are again trainable parameters. GRA employs a global pairwise ranking loss like NEG:

$$L(S^+, S^-) = \max(0, \gamma - \text{score}(S^+) + \text{score}(S^-)), \quad (4.36)$$

where γ is again a hyper-parameter for the margin between positive and negative examples.

Coherence Classifier (CCL)

The learned models discussed so far raise two questions:

1. Is the ubiquitous ranking loss actually necessary?
2. Do specialized architectures improve CM performance?

Laban et al. (2021) show that a RoBERTa-based (Liu et al., 2019b) classifier can easily achieve near-perfect results on the shuffle task on WSJ. They conclude that this demonstrates that the task is not a good evaluation task for CMs. However, they did not test whether their model can predict coherence on non-artificial tasks. We thus include a RoBERTa model that is trained to classify shuffled vs. unshuffled summaries, naming it **Coherence Classifier (CCL)**.⁸ It differs from the remaining shuffle-task-trained models in two important aspects:

1. It does not use a specialized architecture to derive contextualized sentence representations but instead directly uses the representation of the CLS token of RoBERTa as the input to a linear classifier.
2. It is trained using a classification objective, instead of a ranking objective.

Formally, let $h^{(\text{CLS})}$ be the document-level embedding for a given document, then the coherence score is

$$\text{score}(h^{(\text{CLS})}) = \frac{e^{\langle w^+, h^{(\text{CLS})} \rangle + b^+}}{e^{\langle w^+, h^{(\text{CLS})} \rangle + b^+} + e^{\langle w^-, h^{(\text{CLS})} \rangle + b^-}}, \quad (4.37)$$

⁸We found that the original WSJ model does not perform well on SummEval. Thus, we fine-tuned our own model from the same RoBERTA checkpoint.

where $w^+, w^- \in \mathbb{R}^d$, $b^+, b^- \in \mathbb{R}$ are trainable parameters computing logits for the coherent and incoherent class.

4.6.3 Supervised Linguistic Quality Model: SumQE

SumQE (SQE) (Xenouleas et al., 2019) models coherence as part of a broader effort on linguistic quality evaluation. Their setup is based on learning how to score summaries from regression on human scores. Their core innovation in this context is to learn multiple quality dimensions Q in concert. Specifically, given a document embedding $h^{(\text{CLS})}$ derived from BERT (Devlin et al., 2019), they compute the quality score of a quality dimension $q \in Q$ as

$$\text{score}_q(h^{(\text{CLS})}) = \langle w_q, h^{(\text{CLS})} \rangle + b_q, \quad (4.38)$$

where $w_q \in \mathbb{R}^d$, $b_q \in \mathbb{R}$ are trainable parameters and d is the embedding dimension.⁹

The model is trained on human scores with mean squared error. Let $q^* \in \mathbb{R}^{|Q|}$ be human scores across all quality dimensions for the same summary, then the loss for this instance is

$$L(h^{(\text{CLS})}, q^*) = \sum_{i=1}^{|Q|} (q_i^* - \text{score}_i(h^{(\text{CLS})}))^2. \quad (4.39)$$

The model we use in our experiments is trained on data from the DUC 2005, 2006, and 2007 conferences on the five quality dimensions already discussed in 2.3.2: Fluency, Non-redundancy, Referential Clarity, Focus, and Structure and Coherence.¹⁰ Since we are interested in the coherence modelling abilities of **SQE**, we use the *Structure and Coherence* head, as the definition of this quality dimension most closely matches the definition of coherence in our evaluation corpus.

⁹In practice, the authors differentiate between all w_q being treated as independent layers with a scalar output vs. a single layer with $|Q|$ different outputs. This is mathematically equivalent, with the only difference being a very small change in the range of the weight initialization for w_q . We use the model with independent w_q in our experiments.

¹⁰The model is available at https://archive.org/download/sum-qe/BERT_DUC_all_Q5_Multi%20Task-5.h5.

4.6.4 Unsupervised Linguistic Quality Models

While human-annotated data contains rich quality information, it is also a scarce resource. The entirety of the training set of SumQE contains 4,790 different summaries from a total of only 112 systems. This makes methods which can produce quality scores in an unsupervised manner, without being trained on human references, very attractive.

In this work, we study **GRUEN (GRN)** (Zhu and Bhat, 2020) and **BARTScore (BAS)** (Yuan et al., 2021) as representatives of these methods.

GRUEN (GRN)

GRUEN combines five heuristics designed to jointly cover the grammaticality, non-redundancy, focus, and structure and coherence quality dimensions. We briefly discuss each heuristic.

Word Probabilities is a proxy for grammaticality and computed as the average log-probability of each sentence in the document under a bidirectional masked language model (BERT (Devlin et al., 2019) in their implementation). The average log-probability of a sentence is determined by masking out each token in the sentence in turn and computing the probability the model assigns to the correct token at the masked location. Let S be the set of sentences in the document, and let $w_1^{(s)}, \dots, w_{n_s}^{(s)}$ be the tokens in a sentence $s \in S$, then

$$\text{score}_{\text{lm}}(S) = \frac{1}{|S|} \sum_{s \in S} \frac{1}{n_s} \sum_{t=1}^{n_s} \log p(w_t^{(s)} | w_1^{(s)}, \dots, w_{t-1}^{(s)}, w_{t+1}^{(s)}, \dots, w_{n_s}^{(s)}). \quad (4.40)$$

This is also known as the average pseudo-log-likelihood of the tokens (Wang and Cho, 2019).

Linguistic Acceptability is also used as a proxy for grammaticality. GRUEN uses a classification model that is trained on the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) which contains human acceptability ratings. The score is computed for each sentence and derived from the average logits for the

positive class of each individual sentence $e(\text{accept}|s)$:

$$\text{score}_{\text{gram}} = \frac{1}{|S|} \sum_{s \in S} e(\text{accept}|s). \quad (4.41)$$

Non-redundancy is measured using a set of four overlap-based measures that are computed between all pairs of sentences $s_t, s_u \in S \times S; t \neq u$:

- The longest common subsequence (LCS) length in characters
- The length of the LCS in tokens
- The bag of words overlap
- The edit distance

Let m_{tu} indicate the number of these measures that are above a predefined threshold for the sentence pair s_t, s_u , then the total overlap score is

$$\text{score}_{\text{overlap}} = -0.1 \sum_{t, u | t \neq u} m_{tu}. \quad (4.42)$$

Focus measures the degree to which subsequent sentences cover the same general topic. In their original description of GRUEN, Zhu and Bhat (2020) propose to approximate this by computing the word mover’s distance (Kusner et al., 2015) between adjacent sentences. Word-mover’s distance is an approach for computing the dis-similarity of two documents or sentences s_1, s_2 from pairwise token similarities. Let C_{tu} indicate the similarity of the t -th token in s_1 and the u -th token in s_2 . The standard choice here is to use the L2 distance between the embeddings of the tokens. The dissimilarity of s_1, s_2 is computed by solving the following

optimization problem:

$$\text{wmd}(s_1, s_2) = \min \sum_{t=1}^{|s_1|} \sum_{u=1}^{|s_2|} T_{tu} C_{tu}, \quad (4.43)$$

$$\text{given} \quad (4.44)$$

$$\sum_{t=1}^{|s_1|} T_{tu} = 1, \forall 1 \leq u \leq |s_2|, \quad (4.45)$$

$$\sum_{u=1}^{|s_2|} T_{tu} = 1, \forall 1 \leq t \leq |s_1|. \quad (4.46)$$

$$(4.47)$$

Zhu and Bhat propose to convert this into a focus score by taking the reciprocal of the dis-similarity between adjacent sentences and averaging the result:

$$\text{score}_{\text{focus}} = \frac{1}{|S|} \sum_{t=1}^{|S|} \frac{1}{\text{wmd}(s_t, s_{t+1})}. \quad (4.48)$$

However, the reference implementation of GRUEN, which we use in this chapter, does not employ WMD as the dissimilarity metric as described in the paper but uses a simple centroid distance (i.e. the distance between the averaged word embeddings of s_t and s_u) instead.

Structure and Coherence is measured using a model that is trained using a modified variant of the shuffle task used in other CMs. To create a negative sample, a document is split into two segments along a sentence boundary and the second segment is moved to the beginning. The prediction of the classifier trained on this task is the coherence score, $\text{score}_{\text{coh}}$.

To compute the **final score**, the individual quality scores are summed up. However, the official reference implementation we use does not include the coherence score.¹¹ The score we derive from GRUEN for our experiments is thus:

$$\text{score} = \text{score}_{\text{lm}} + \text{score}_{\text{gram}} + \text{score}_{\text{overlap}} + \text{score}_{\text{focus}}. \quad (4.49)$$

¹¹We have confirmed that this is intentional in personal communication with the authors.

BARTScore (BAS)

BARTScore (BAS) (Yuan et al., 2021) provides an alternative unsupervised CM that leverages the probability of a summary under a pretrained BART model as a score. This follows the intuition that a good summary should have high probability under the a strong summarization model and is similar to the LM score (score_{lm}) used in GRUEN, but in a conditional setting. BARTScore is defined for three different settings. Let r_1, \dots, r_n be the reference summary tokens, let g_1, \dots, g_m be the generated summary tokens, and let d_1, \dots, d_l be the input document tokens, then the three score variants are defined as

$$\text{score}_{R \rightarrow G} = \frac{1}{m} \sum_{i=1}^m \log p(g_i | g_1, \dots, g_{i-1}, r_1, \dots, r_n), \quad (4.50)$$

$$\text{score}_{G \rightarrow R} = \frac{1}{n} \sum_{i=1}^n \log p(r_i | r_1, \dots, r_{i-1}, g_1, \dots, g_m), \quad (4.51)$$

$$\text{score}_{D \rightarrow G} = \frac{1}{m} \sum_{i=1}^m \log p(g_i | g_1, \dots, g_{i-1}, d_1, \dots, d_l). \quad (4.52)$$

Here, $\text{score}_{R \rightarrow G}$ is designed to measure recall of information from the reference, $\text{score}_{G \rightarrow R}$ is designed to measure precision, and $\text{score}_{D \rightarrow G}$ measures faithfulness to the input.

Since the latter is recommended for coherence by the authors and we work in a reference-free setting, we only use $\text{score}_{D \rightarrow G}$ in our experiments. We use a variant of BART that is fine-tuned on CNN/DM summaries since this performed best in the original evaluation.

4.7 Results

We present the correlation of all CMs with human coherence ratings in Table 4.3. We report (average) Kendalls τ for all EMs introduced in Section 4.5. For C_{pair} we additionally report accuracy.

Focusing on τ_{sys} first, we find that CCL, BAS, GRN, and, to a lesser extent, SQE achieve relatively high scores while the remaining CMs fail to outperform even the random baseline. However, inspection of τ_{sum} , $\tau_{\text{pair}}/Acc_{\text{pair}}$, and τ_{intra} reveals

Metric	τ_{intra}	τ_{pair}	τ_{sum}	τ_{sys}	Acc. <i>pair</i>
HUM	+0.75 (+0.70 +0.79)	+0.81 (+0.76 +0.85)	+0.81 (+0.77 +0.84)	+0.91 (+0.71 +1.00)	+0.77 (+0.71 +0.81)
RND	-0.00 (-0.06 +0.05)	-0.00 (-0.07 +0.06)	+0.00 (-0.05 +0.05)	+0.09 (-0.41 +0.53)	+0.50 (+0.46 +0.54)
EGR	-0.04 (-0.12 +0.04)	-0.11 (-0.19, -0.02)	-0.09 (-0.16 -0.01)	-0.25 (-0.59 +0.10)	+0.40 (+0.36 +0.44)
EEG C/D	+0.02 (-0.07 +0.10)	+0.04 (-0.10 +0.18)	+0.06 (-0.06 +0.17)	-0.19 (-0.68 +0.26)	+0.52 (+0.45 +0.59)
EEG WSJ	+0.02 (-0.06 +0.10)	+0.00 (-0.09 +0.11)	+0.03 (-0.06 +0.11)	-0.19 (-0.60 +0.26)	+0.50 (+0.44 +0.55)
NEG C/D	-0.07 (-0.14 -0.00)	-0.05 (-0.14 +0.07)	-0.06 (-0.15 +0.03)	-0.15 (-0.61 +0.32)	+0.47 (+0.42 +0.53)
NEG DUC	-0.08 (-0.16 +0.01)	-0.06 (-0.18 +0.06)	-0.07 (-0.17 +0.04)	-0.06 (-0.49 +0.31)	+0.47 (+0.40 +0.53)
NEG WSJ	-0.02 (-0.08 +0.05)	-0.08 (-0.17 +0.00)	-0.07 (-0.15 +0.02)	-0.43 (-0.69 -0.05)	+0.45 (+0.41 +0.50)
UNF C/D	+0.04 (-0.03 +0.11)	+0.05 (-0.05 +0.14)	+0.06 (-0.01 +0.13)	+0.13 (-0.33 +0.59)	+0.53 (+0.48 +0.57)
UNF WSJ	+0.02 (-0.05 +0.09)	-0.11 (-0.26 +0.03)	-0.04 (-0.15 +0.05)	-0.09 (-0.51 +0.39)	+0.44 (+0.36 +0.52)
GRA DUC	-0.04 (-0.12 +0.03)	-0.05 (-0.16 +0.03)	-0.06 (-0.13 +0.01)	-0.19 (-0.65 +0.25)	+0.47 (+0.43 +0.52)
GRA C/D	+0.08 (+0.02 +0.15)	+0.09 (-0.02 +0.19)	+0.11 (+0.01 +0.18)	+0.37 (-0.07 +0.69)	+0.55 (+0.49 +0.60)
GRA WSJ	+0.08 (+0.01 +0.15)	-0.01 (-0.11 +0.10)	+0.02 (-0.06 +0.12)	-0.09 (-0.47 +0.37)	+0.49 (+0.44 +0.55)
CCL C/D	+0.26 (+0.19 +0.33)	+0.40 (+0.31 +0.49)	+0.39 (+0.31 +0.44)	+0.62 (+0.30 +0.86)	+0.71 (+0.66 +0.76)
CCL WSJ	+0.20 (+0.12 +0.26)	+0.35 (+0.25 +0.46)	+0.33 (+0.24 +0.41)	+0.74 (+0.40 +0.92)	+0.69 (+0.63 +0.74)
BAS	+0.17 (+0.08 +0.26)	+0.37 (+0.23 +0.51)	+0.32 (+0.20 +0.42)	+0.72 (+0.42 +0.89)	+0.69 (+0.62 +0.77)
GRN	+0.18 (+0.12 +0.25)	+0.26 (+0.17 +0.35)	+0.27 (+0.19 +0.34)	+0.72 (+0.38 +0.89)	+0.63 (+0.58 +0.69)
SQE	+0.19 (+0.13 +0.26)	+0.26 (+0.15 +0.36)	+0.24 (+0.15 +0.32)	+0.51 (+0.05 +0.80)	+0.64 (+0.58 +0.69)

TABLE 4.3: Results on SummEval for all CMs. Correlation is expressed in Kendall’s τ . Numbers in brackets indicated 95% CIs computed using bootstrap resampling as described in Section 4.2.1 with 1000 samples. Highest are bold.

that even these apparently strong CMs struggle to reliably assess coherence of individual summaries, with τ_{intra} being the most challenging regime. Comparing CMs, CCL C/D is most promising across all EMs except τ_{sys} , where scores are near indistinguishable due to high uncertainty. Interestingly, we find that its advantage is greatest on τ_{intra} , where its competitors exhibit particular weakness compared to other EMs. These sharp score drops might suggest other EMs reflect some system-level confounders. In combination with the observation that confounder scores as reported in Table 4.1 fall within the 95% CI of most CMs on all EMs except τ_{intra} , this prompts us to investigate CMs for potential biases in Section 4.8.

4.7.1 Detailed Intra-System Correlation Results

Intra-system correlation allows us to study performance of the CMs on the summaries of each individual summarizer. Figure 4.2 shows the individual intra-system correlations for all summarizers in SummEval for the best CMs and the human upper bound. We find that CMs struggle across the whole range of summarizers, including summarizers with high variance in coherence scores, where we would

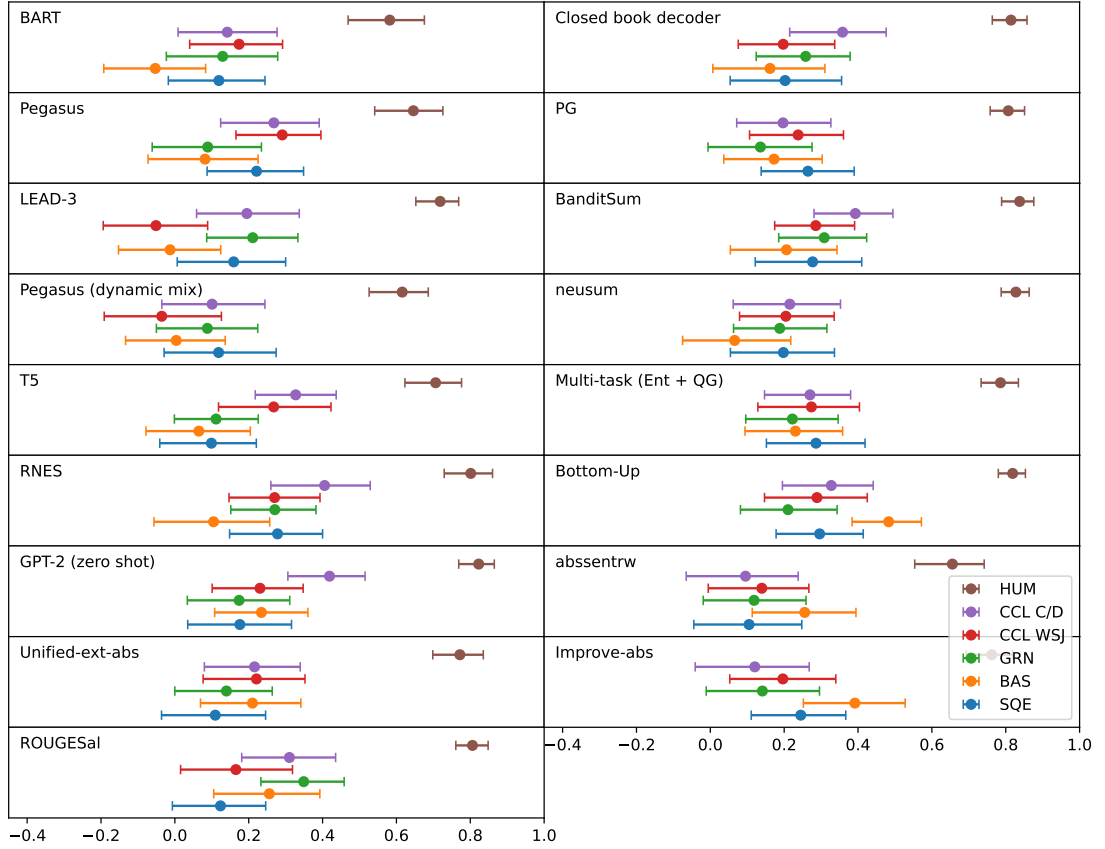


FIGURE 4.2: Intra-system correlations of the best CMs as well as the human upper bound on the SummEval dataset. Bars indicate 95% confidence intervals determined by bootstrap resampling with 1000 samples.

expect the task to be easier. Furthermore, we find none of the available CMs can consistently outperform all others. For example, BAS outperforms other CMs on Bottom-Up and Improve-Abs but performs significantly worse on the top systems, including BART itself.

4.8 Bias Matrices

We have shown in Section 4.5.2 that CMs can appear to correlate with human coherence judgements by exploiting system-level confounders. However, it is unclear to which extent this just holds for our artificial confounders or is also an

issue in realistic CM evaluation. We therefore introduce *bias matrices*, a tool that allows us to easily inspect the decisions of any measure by separately analyzing *consistent* and *inverted* pairs of summaries from different summarizers. *Based on human scores*, we call a summary pair *consistent* if the higher-scoring summary is produced by the summarizer with the higher average score, whereas we call a pair *inverted* if the overall worse summarizer produces a stronger summary. We are specifically interested in finding instances where a CM ranks consistent pairs for a strong summarizer correctly but fails to correctly rank its inverted pairs. This is indicative of a measure having a bias towards outputs of this particular summarizer, instead of measuring coherence. Since, for strong systems, most pairs are consistent, this can still result in many correct comparisons.

Given predicted and human scores P, H as in Section 4.5 and systems $s_1, s_2 \in S$ from the set of systems S with s_1 having a higher average human score than s_2 , we define two new metrics. τ^+ indicates the ability of a measure to rank consistent pairs, whereas τ^- indicates the same for inconsistent pairs. For τ^+ we define

$$H^+ := \{(d_i, d_j) | H_{(d_i, s_1)} > H_{(d_j, s_2)}\}, \quad (4.53)$$

$$P^+ := \{(d_i, d_j) | P_{(d_i, s_1)} > P_{(d_j, s_2)}\}, \quad (4.54)$$

$$\tau^+ := \frac{2|H^+ \cap P^+| - |H^+|}{|H^+|}. \quad (4.55)$$

For τ^- we invert the comparisons.¹² Both τ^+ and τ^- are bounded between -1 and 1. If the ranking is -1, this indicates the ranking is always incorrect; if it is 1, it is always correct. To derive the $|S| \times |S|$ bias matrix T , we order systems $s_1, \dots, s_{|S|}$ in descending order of their average human score. We then have

$$T_{ij} := \begin{cases} \tau_{(s_i, s_j)}^+ & i < j \\ \tau_{(s_j, s_i)}^- & i > j \\ 0 & i = j \end{cases}. \quad (4.56)$$

¹²If s_1 is better than s_2 on every document, τ^- is undefined. In this case, biased and unbiased CMs are indistinguishable.

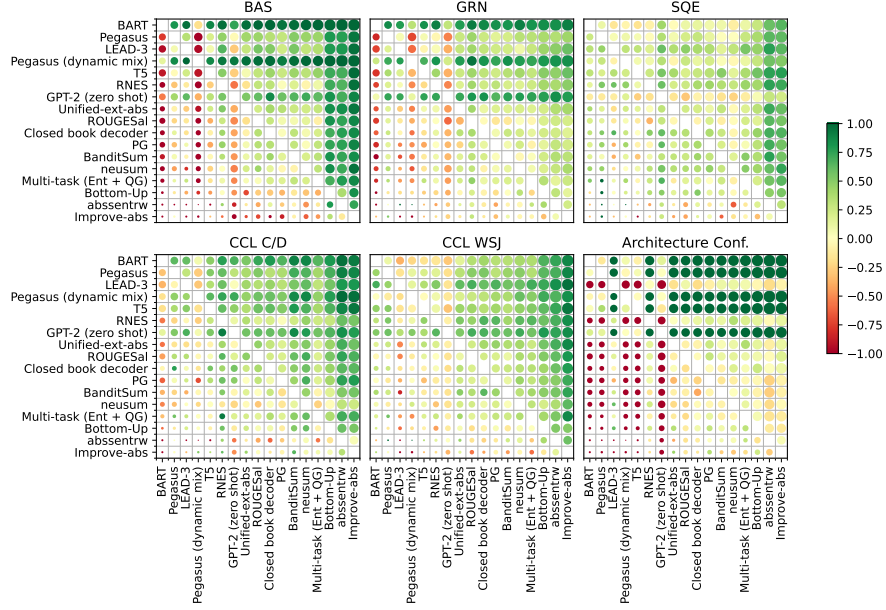


FIGURE 4.3: Bias matrices for the best CMs. We also show the bias matrix for the architecture confounder for reference. See Figure 4.4 for a brief tutorial to bias matrix analysis.

We visualize T for the most promising CMs in Figure 4.3. To aid interpretation, we provide an annotated version for scores generated by BAS in Figure 4.4. We find that GRN and BAS show a very strong preference for summaries generated by BART, ranking them almost universally higher even when this disagrees with human judgements. In case of BAS, this is unsurprising, since BART and BAS use the same underlying model. For GRN the reason is less clear, though analysis in Section 4.9.2 suggests that it might rely on the higher grammaticality of BART output. For the other CMs, biases are less evident, though CCL C/D shows a slight preference for BART and Pegasus and CCL WSJ has a slight bias towards LEAD and GPT-2.

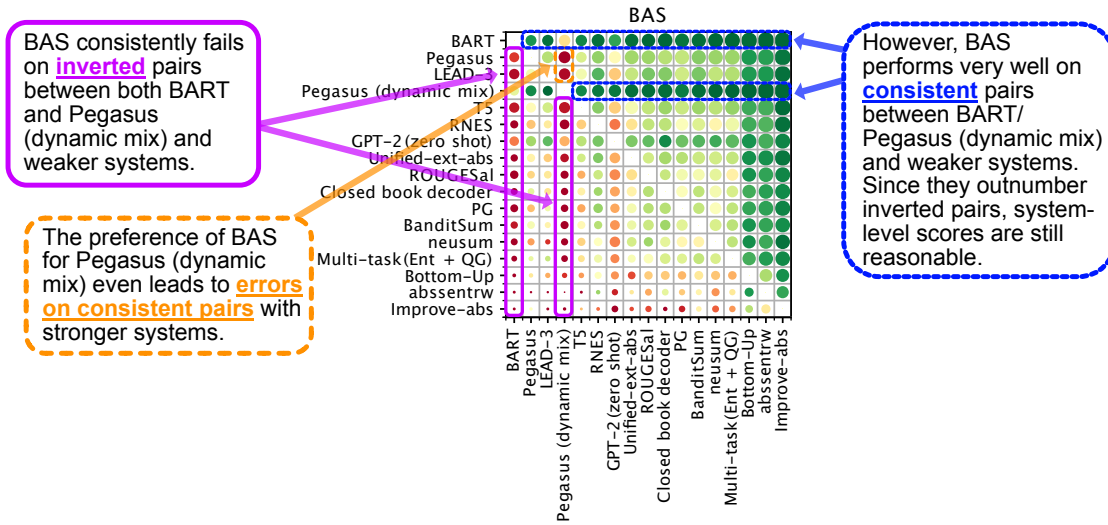


FIGURE 4.4: Bias matrix for BAS with specific analysis for BART and Pegasus. The upper triangular matrix indicates τ^+ for the given summarizer pair, the lower τ^- . The area of each circle is proportional to the number of pairs in H^+/H^- for the cell. To read off the behaviour of the CM on a specific summarizer, we follow both the corresponding row and column. A high score in the row, combined with a low score in the corresponding cell in the column implies the CM is biased towards generations by this particular summarizer.

4.9 Coherence Measure Analysis

4.9.1 Correlation with Shuffle-Performance

Mohiuddin et al. (2021) have shown that the performance of CMs on the shuffle task is not predictive for performance on summary coherence evaluation. However, at the same time, the shuffling-based CCL shows comparatively strong performance in our experiments. To better understand the relation between shuffling and summary coherence, we test the ability of all CMs to discriminate shuffled and non-shuffled *reference summaries* from the test split of CNN/DM. Results are in Table 4.4.

Of the CMs that perform best on coherence evaluation (see Table 4.3), most also perform well on the shuffle task (CCL, BAS, SQE). Only GRN fails on this task, showing random accuracy. This is troubling as we would expect any CM that is able to identify coherent summaries on SummEval to be able to identify at least some shuffled reference summaries. This suggests that GRN models coherence only indirectly via proxy variables, which we elaborate on in Section 4.9.2.

For the entity-based measures, EGR, EEG, and NEG, their difficulties on the SummEval dataset are also reflected in the shuffle task. This suggests that these CMs struggle generally on CNN/DM-style summaries. In Section 4.9.3 we demonstrate that this is due to the overall lack of entity overlap in this dataset. Finally, UNF C/D and GRA are outliers in that they show shuffle performance on CNN/DM that is similar or better than SQE but still perform near random on SummEval coherence modelling. We investigate this in Section 4.9.4.

4.9.2 GRUEN

GRN works well for system-level correlation yet is incapable of solving the shuffle task. This prompts us to investigate its individual components as described in Section 4.6.4.

Table 4.5 shows the system-level correlation of both the individual scores and all pairwise combinations of GRN component scores. The grammaticality score $\text{score}_{\text{gram}}$ plus the redundancy score $\text{score}_{\text{overlap}}$ alone account for almost the full system-level correlation of 0.72. Since neither score is dependent on sentence order, they

	C/D	WSJ (orig.)
EGR	0.426	0.889
EEG	0.523 _(c) 0.498 _(w)	0.840
NEG	0.524 _(c) 0.603 _(w) 0.522 _(d)	0.855
GRA	0.838 _(c) 0.623 _(w) 0.439 _(d)	0.924
UNF	0.803 _(c) 0.589 _(w)	0.93
CCL	0.929 _(c) 0.862 _(w)	0.97
BAS	0.896	-
GRN	0.504	-
SQE	0.707	-

TABLE 4.4: Shuffle accuracies on CNN/DM for 1000 randomly sampled reference summaries. (c) means that the model was trained on CNN/DM (w) on WSJ and (d) on DUC 03. Baseline accuracy is 50%. For reference, we also list originally reported shuffle results on full WSJ articles where applicable.

	score _{gram}	score _{overlap}	score _{lm}	score _{focus}
Cola	0.57	0.71	0.59	0.63
Redun.		0.51	0.57	0.51
LM			0.15	0.35
Focus				0.49

TABLE 4.5: Performance of GRN constituent measures. Cells indicate system-level correlation of the combination of the respective measures. Individual measure performance is indicated on the diagonal.

Corpus	Docs	Sents
CNN/DM Ref.	0.287	0.458
SummEval	0.178	0.301
DUC 03	0.014	0.121

TABLE 4.6: Proportion of documents without any entity overlap, as well as average ratio of sentences without entity links per document for various datasets.

can by design not fully account for summary coherence. The results raise considerable doubt about the generalizability of GRNs performance on this task.

4.9.3 Entity Driven Measures

To explain why EEG, EGR, and NEG perform poorly even on the shuffle task, we investigate the role of entity (re-)occurrences in CNN/DM summaries. Table 4.6 shows that both reference summaries and SummEval data have very little lexical entity overlap between sentences.¹³ A considerable number of summaries in both SummEval and CNN/DM show no entity overlap between any of their sentences. Therefore, entity-based models are inherently limited, at least when using lexical overlap to determine entity re-occurrence. We leave a thorough investigation of solutions like better coreference resolution or using embedding-based methods as in Mesgar and Strube (2016) to future work.

4.9.4 Global Training vs. Pairwise Ranking

While CMs that fail the in-domain shuffle task are likely to be unsuitable for CNN/DM summaries, it is less clear why CMs with reasonable shuffle performance fail on SummEval like UNF C/D and GRA C/D. We theorize that one reason is that both UNF and GRA are trained on a margin-based ranking loss between shuffled and non-shuffled variants of the *same* document, which implies that both have the same tokens and number of sentences. The training loss thus does not impose constraints on the behaviour of the function between inputs of *different* lengths

¹³As determined by the Brown Coherence Toolkit. See Appendix C.

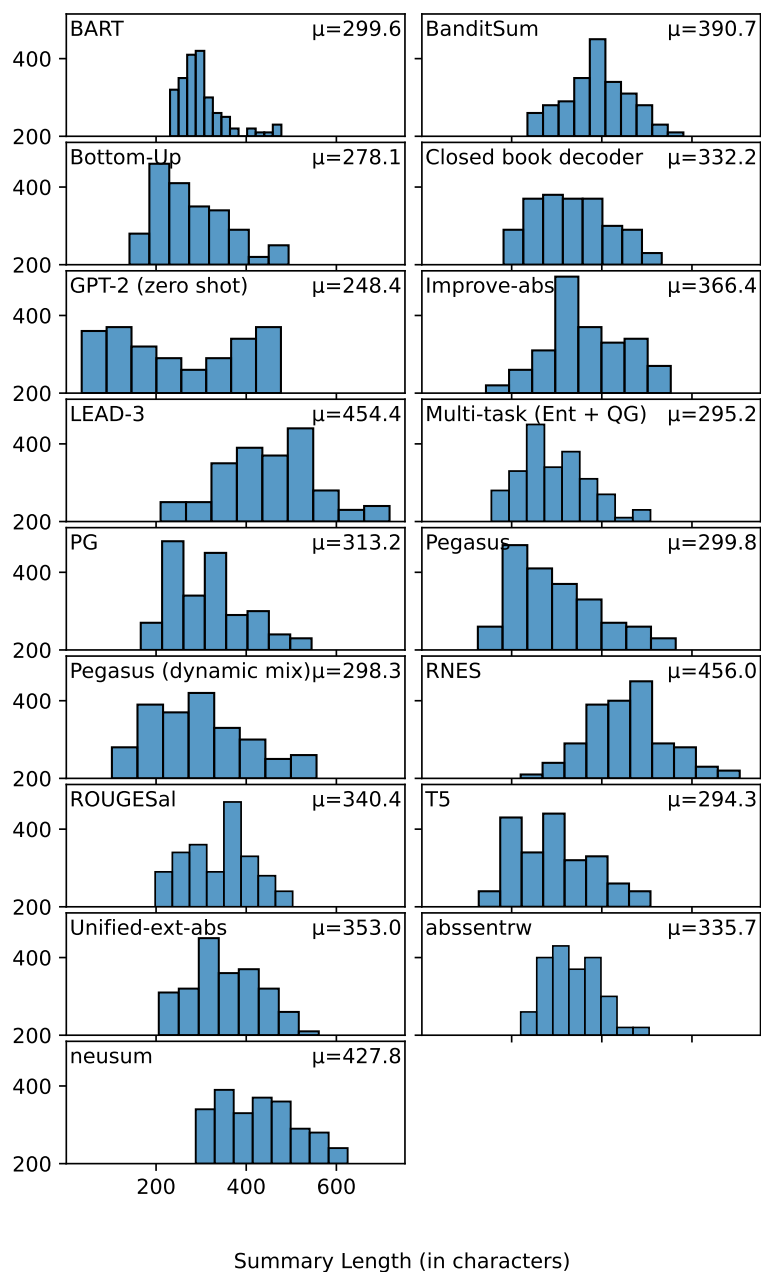


FIGURE 4.5: Histograms of the lengths of summaries generated by the summarizers in SummEval and their mean lengths (μ). Both in characters.

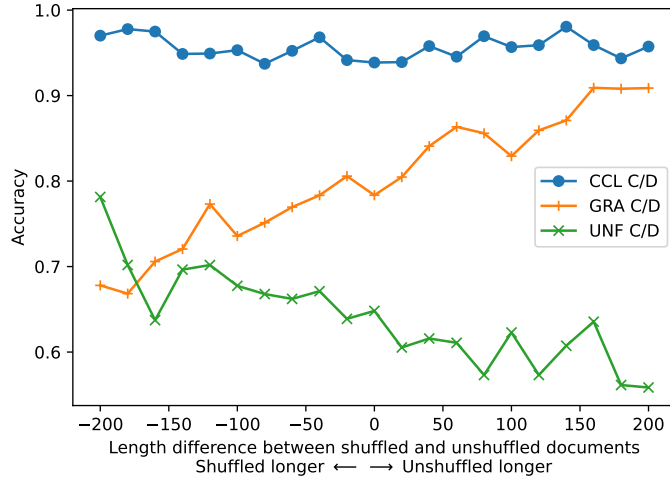


FIGURE 4.6: Ranking accuracy between shuffled and original summaries of different lengths (in characters). We sample 10,000 pairs and group them in buckets of 20 characters and clamp differences between -200 and 200.

and tokens. In contrast, the classification objective of **CCL** enforces a globally correct ranking of shuffled vs. unshuffled documents.

Since SummEval, unlike e.g. DUC, has no agreed-upon length constraint, the assumption of equal document length is problematic. We can easily see this by considering the length distribution of summarizer outputs on SummEval in Figure 4.5. There is considerable variance both between summaries of different summarizers and within summaries of the same summarizer.

Verifying our hypothesis on SummEval directly is difficult since summary length is deeply confounded with the generating summarizer. However, we can investigate the ability of CMs to correctly rank documents of different lengths and content by modifying the shuffle task to compare reference summaries to shuffled variants of *different* reference summaries. Figure 4.6 shows the relation between the difference in length between the shuffled and unshuffled summaries and the ranking accuracy of the CMs. **UNF** performs very poorly on the task, especially if the original summary is long. **GRA**, on the other hand, prefers longer documents, even if they are shuffled. In contrast, **CCL** is consistently able to correctly rank summaries

regardless of length difference. Thus, for both UNF and GRA comparing documents of different lengths and content is a major obstacle. The stability of CCL suggests that replacing pairwise ranking with a classification objective is a direct fix to this issue. These results are also consistent with parallel work by Jwalapuram et al. (2022) who extend the pairwise shuffle task to consider multiple negative examples. They find that including negative samples from different documents in the negative set during training improves model performance on downstream tasks.

4.10 Discussion

In this chapter, we have introduced two techniques for improving meta-evaluation: intra-system correlation and bias matrices. We have employed them to investigate the performance of a wide array of CMs for summary evaluation that have not been previously systematically compared.

Our investigations show that CMs must be carefully evaluated in order to avoid rewarding the modelling of shallow, system-level confounders, that are unlikely to generalize. Our newly suggested intra-system correlation can be used *alongside* other EMs to guard against this. Where correlation unexpectedly drops when going from system-level to intra-system correlation, our bias matrices provide a visual inspection tool to identify where high system-level EM scores can be explained with models being overly biased towards the output of strong summarizers. While our empirical findings focus on summary coherence modelling as a particularly interesting quality dimension, the fundamental principles underlying our EMs are applicable to any meta-evaluation.

Regarding the investigated CMs, our results point towards two lessons for future work:

1. CNN/DM summaries are not amenable to entity-based analysis without considerable additional work to improve entity detection.
2. Self-supervised training via the shuffle task shows the greatest promise for future improvements.

However, we note that conversely, good shuffle performance does not naturally transfer to a strong summary coherence measure. We find evidence that the frequent modelling choice of selecting documents with the same length and content in a pairwise ranking scheme during training prevents models from generalizing to realistic summary coherence evaluation settings. Training in a classification setup instead of the more common pairwise setup provides an effective fix for this. The resulting coherence classifier outperforms all competing models, with the difference being most notable in our newly introduced intra-system correlation EM. However, the overall low scores in intra-system correlation show that there remains a considerable need for improvement of the investigated models before they become practical.

Chapter 5

Faithfulness Evaluation with NLI Models

5.1 Motivation

In addition to being well-written and coherent, a good summary must naturally also correctly reflect the content of the input. As we have alluded to in numerous places in this thesis, however, language models suffer from a tendency to *hallucinate* information (Maynez et al., 2020), resulting in generations that are not faithful to their input documents. This limits the trustworthiness of such models and raises a need for automatic faithfulness metrics. In this context, models trained on natural language inference (NLI) (Bowman et al., 2015) are attractive. In NLI, a classifier must determine whether a hypothesis is logically entailed by, contradicts with, or has a neutral relation to a given premise. Intuitively, a generation being *faithful* implies it must be *entailed* by the source (Falke et al., 2019).

However, pure NLI models have seen mixed success in faithfulness evaluation (Falke et al., 2019; Kryscinski et al., 2020; Wang et al., 2020; Maynez et al., 2020). While in an evaluation on the TRUE benchmark (Honovich et al., 2022), which contains datasets from knowledge-grounded dialogue and paraphrasing, in addition to summarization, NLI-derived metrics perform best overall, they require impractically large models or costly additional machinery such as question generation and answering models at inference time, while still showing robustness issues. This leads us to identify faithfulness as an area that is in need of more *cost-efficient*

automatic evaluation. We ask: *What is still needed for pure NLI models to perform robustly across faithfulness datasets – while remaining cheap enough to serve as a lean and practical evaluation tool?* Since faithfulness is a concern not only in summarization, we are also going to consider two additional tasks in this chapter: paraphrasing and knowledge-grounded dialogue.

We enhance a relatively small NLI model to make it work robustly across tasks in three ways:

Task-Adaptive Data Augmentation. In NLI, a hypothesis must be fully entailed by its supporting premise, meaning that we cannot accept a hypothesis as being true if it is only supported by parts of the premise. However, in faithfulness, not all parts of the generation always need to be grounded. We identify an instance of this phenomenon in dialogue where parts of a turn can fulfill communicative functions such as hedging or establishing emotional connection and are often disregarded in faithfulness annotation. Hence, when applying NLI models to *complete dialogue turns* that may include statements irrelevant for grounding, we run a risk of producing incorrect unfaithfulness predictions.

To alleviate this issue, we propose a simple **data augmentation** method to adapt NLI models to genres where they need to be aware of statements that must be exempt from NLI-based faithfulness evaluation. Our approach is computationally attractive, as it avoids an increase of cost at inference time.

Integration of NLI Contradiction Scores. Existing NLI faithfulness metrics typically use the entailment score for their predictions (Honovich et al., 2022; Falke et al., 2019; Kryscinski et al., 2020). However, Chen and Eger (2023) show that subtracting the contradiction score from the entailment score (referred to as *e-c*) can improve NLI performance in certain evaluation tasks. We show that there also is a strong positive effect of *e-c* for faithfulness prediction and demonstrate that this is due to a high contradiction probability being a more reliable predictor of unfaithfulness than low entailment probability.

Monte-Carlo Dropout Inference. Applying NLI models to faithfulness prediction involves a domain shift from largely human-written data to automatically generated text. To make NLI model scores more robust under this shift, we propose to use Monte-Carlo dropout during inference (Srivastava et al., 2014). This essentially creates a cheap *ensemble* and has been shown to deal better with noisy

labels (Goel and Chen, 2021). This approach leads to consistent score improvements in our experiments.

The combination of all modifications not only strongly improves over a baseline NLI model, but also outperforms all other metrics on TRUE, on average, while being **cheaper** and **smaller**.

In sum, we make the following contributions in this chapter:

1. We identify a divergence in the definitions of entailment in NLI and faithfulness in some tasks that leads to poor performance of NLI models and propose a straightforward augmentation method to overcome this issue.
2. We thoroughly investigate the effect of integrating the NLI contradiction score into a NLI-based faithfulness metric.
3. We propose the use of Monte-Carlo dropout during inference to create an ad-hoc ensemble for better faithfulness detection.

The work presented in this chapter has previously been published as

Julius Steen et al. (2023). “With a Little Push, NLI Models can Robustly and Efficiently Predict Faithfulness”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers et al. Toronto, Canada: Association for Computational Linguistics, pp. 914–924. DOI: 10.18653/v1/2023.acl-short.79. URL: <https://aclanthology.org/2023.acl-short.79>.

5.2 Background and Related Work

5.2.1 Faithfulness and Factuality

A summary of a text should generally reproduce the relevant facts of the input and not add any additional information to the document. Such added facts are typically called *hallucinations* and a summary that contains such facts is called *unfaithful* (Maynez et al., 2020). It is important to differentiate this from the

truthfulness or *factuality* of a summary. A hallucination can be factually true, but if the stated facts are not present in the input, the summary is still considered unfaithful. It is also conceivable that a summary is faithful but not factual if it was generated from non-factual input. However, this differentiation is not typically made in the literature and input documents are assumed to be factual.

A second important distinction is that of *intrinsic* and *extrinsic* hallucinations (Maynez et al., 2020). Intrinsic hallucinations are based on facts in the input document but present them in a way that is misleading or incorrect. To illustrate, the following hypothetical summary of this chapter contains an intrinsic hallucination:

This chapter discusses using paraphrasing to improve NLI models.

All concepts in the summary are present in the introduction section of this chapter, yet the resulting summary is clearly not accurate.

An extrinsic hallucination, on the other hand, might look like this:

This chapter improves the faithfulness of BART.

Here, *BART* does not appear at all in this chapter.¹ It is an extrinsic hallucination.

Faithfulness as a property of a generation is also applicable beyond summarization. In **paraphrasing** (Zhang et al., 2019), a paraphrase should not add any additional information to the paraphrase that is not in the input sentence. In **knowledge-grounded dialogue** (Honovich et al., 2021; Dziri et al., 2022), a system must respond to user queries based on grounding provided to the system, e.g. a Wikipedia article. Here, faithfulness is defined with regard to the grounding document, although as we will show in this section, care must be taken to identify which parts of a given generation are *expected* to be grounded in the input and which are clearly identifiable as conversational fillers. Faithfulness also plays a role in fact-checking in dialogue (Gupta et al., 2022), where the grounding is unknown and must be retrieved before verification. Since we do not consider the task of retrieving grounding in this chapter, we group this task together with knowledge-grounded dialogue.

¹Except, of course, in this example.

Finally, we note for completeness that the faithfulness problem has also been studied in the context of machine translation (Weng et al., 2020; Guerreiro et al., 2023; Xu et al., 2023). However, since input and generation are in different languages in these settings, this is typically treated as an independent problem. We thus do not discuss it in more depth here.

5.2.2 Metrics

It has been shown that faithfulness is poorly captured by traditional summarization metrics, such as ROUGE (Lin, 2004b), which motivates the need for alternative approaches (Maynez et al., 2020). One attractive approach is to directly learn a metric that takes a document D and a summary² S and outputs a corresponding faithfulness score. However, just as for coherence, where we have discussed this problem in Chapter 4, training data typically requires extensive human annotation, which makes this naive approach challenging. Analogously to the shuffle task for coherence, automatic generation of unfaithful outputs has been proposed as a remedy. Kryscinski et al. (2020) use a set of noising transformations to corrupt reference summaries. They then train a classifier that discriminates noised and gold summaries. They call this approach FactCC. A shortcoming of FactCC is that the noising transformations are limited in the kinds of errors they can introduce and must be manually chosen to be representative of real-world faithfulness issues. Goyal and Durrett (2020) instead use the bottom beam-search beams of a paraphrasing model as generations likely to introduce factual errors. By comparing these generations to the gold paraphrases, they derive fine-grained error annotations. They use this to build a fine-grained faithfulness model called dependency arc entailment (DAE), which can determine the faithfulness of individual dependency arcs, instead of just text snippets.

However, the use of the bottom beam in DAE gives very little control over the kind of error introduced. To allow for finer control over the negative samples, some authors have suggested the use of meaning representations, which can

²Most of the metrics discussed here were initially proposed for summarization, so we will use summary as a general term for the generation we seek to evaluate in this section. All metrics discussed here naturally generalize to any generation that should be faithful to some grounding.

be transformed into natural language using a generator. By removing or altering input representations, one can produce a variety of unfaithful generations for training. Utama et al. (2022) train a model to generate summaries from OpenIE predicate-argument tuples. By dropping predicates and arguments during inference, the model can be forced to hallucinate the missing information. Qiu et al. (2024) propose an even more fine-grained approach modifying summaries represented by abstract meaning representation (AMR) (Banarescu et al., 2013) graphs. By applying rule-based perturbations to the graph and then using an AMR-to-text model they can generate summaries with known faithfulness issues.

Learning from model-based generations, however, is relatively resource intensive, both for generation and for the subsequently required training. An alternative is to instead use models and data from data-rich tasks that are semantically close to faithfulness. A natural candidate here are natural language inference (NLI) models. NLI (Bowman et al., 2015) is the task of determining whether a hypothesis can be inferred from or is contradicted by a given premise. This is typically framed as a three-way classification task with the input being premise and hypothesis and the output being either *entailment*, *contradiction*, or *neutral* in case neither relation holds. Intuitively, for a summary to be faithful to a document, it must be entailed by the latter, although we will show later that this is not true in all settings.

While NLI models can be straightforwardly adapted to faithfulness evaluation by using the entailment probability $p(\text{entail}|D, S)$ between a document D and a summary S as the score, naive application of NLI models has been shown to be a bad predictor of faithfulness (Falke et al., 2019). Laban et al. (2022) propose that this is at least partially caused by applying models to the entirety of D and S . Common NLI training datasets have single-sentence premises and hypotheses. Even if the premise is longer, as, for example, in the ANLI dataset (Nie et al., 2020), the hypothesis is still typically a single sentence. This is in contrast to summarization, where documents and summaries are often multiple sentences long. Laban et al. propose to instead first decompose D and S into smaller pieces, like sentences or paragraphs, and then compute pairwise NLI entailment probabilities E_{ij} between each element d_i of the document and s_j of the summary: $E_{ij} =$

$p(\text{entail}|d_i, s_j)$. To derive a instance-level score, E can then be reduced in a zero-shot fashion by a max-mean operation:

$$\text{SummaC}_{\text{zs}} = \frac{1}{|S|} \sum_{j=1}^{|S|} \max_i E_{ij}. \quad (5.1)$$

Given sufficient training data, the mapping of E to a score can also be learned. Laban et al. propose to first convert E into a histogram H based on a set of m non-overlapping bins B covering the value range of E :

$$H_{jk} = \sum_{i=1}^{|D|} \mathbb{1}(E_{ij} \in B_k), \quad (5.2)$$

where $\mathbb{1}$ is the indicator function. Each row of the histogram H_j is then transformed into a scalar s_j by multiplying H_j with a learnable weight matrix $W_{\text{conv}} \in \mathbb{R}^{1 \times m}$:

$$s_j = W_{\text{conv}} H_j. \quad (5.3)$$

The learned score, called $\text{SummaC}_{\text{conv}}$, is then computed by computing the mean over all s_j .

We can conceptualize SummaC as introducing a decomposition to the input text and the summary to make it easier to verify the facts in the summary with those of the input. In SummaC, this decomposition is based on sentence or paragraph splitting, but there are alternative approaches. **Question answering**-based metrics decompose the summary into a set of question/answer pairs. Faithfulness is determined by comparing the answers an automatic question answering (QA) model gives based on the summary and based on the input. QAGS (Wang et al., 2020) and FEQA (Durmus et al., 2020) both implement this approach with minor variations to the employed model. In both cases, a question generation (QG) model is run on the summary, with questions focusing on named entities and noun phrases. Questions are then answered on the input. Since questions have simple phrases as answers, the answers can be verified using token overlap between input and summary answers as the final faithfulness score.

Since QG/QA is just a different decomposition of the summary, it can also be

combined with NLI-based metrics. Q2 (Honovich et al., 2021) does exactly this by using an NLI model entailment score as the final faithfulness metrics, to allow for better handling of lexical variation in the answers.

5.2.3 Datasets and Meta Evaluation

Datasets

As discussed in the previous chapter, meta-evaluation is dependent on the availability of human judgements. The simplest way to elicit faithfulness judgements is by asking crowd workers if a given text is supported by the input or not (Falke et al., 2019; Kryscinski et al., 2020; Wang et al., 2020). However, other annotation procedures have been proposed as well: SummEval (Fabbri et al., 2021b) contains Likert-scale ratings for summarization faithfulness. Tang et al. (2022) conduct a comparison of BWS and Likert following our approach in Chapter 3 and find that both methods are reliable.

However, most recent large-scale faithfulness annotations retain the binary annotation method and focus instead on more comprehensive error typologies to aid annotation and support metric analysis: Notably, Pagnoni et al. (2021) propose a fine-grained analysis framework that differentiates errors into a total of seven error types. They group these errors into three overall categories:

Frame-based errors include incorrect *predicates*, incorrect *primary arguments*, and incorrect attributes specifying the *circumstance* of the described act.

Discourse errors include erroneous or unclear *coreference* relations and errors in *discourse links*.

Verifiability Errors cover cases where the correctness of a sentence cannot be confirmed either due grammatical errors rendering it unreadable or due to extrinsic hallucinations.

In work outside of summarization relevant to our task, annotations in dialogue largely follow the binary paradigm established in summarization (Dziri et al., 2022; Honovich et al., 2021). For fact verification, Gupta et al. (2022) split the non-faithful category into two parts: *Not enough information* and *refuted*. This is

owed to the fact that unlike generation tasks, where we know which input was passed to the generation system, for fact verification, the grounding is a retrieved set of documents, which may not contain the source of the statement to be verified.

The differences in annotation granularity and annotated domains has led to efforts to reconcile individual annotation efforts into larger benchmarks. For summarization, AggreFact (Tang et al., 2023) combines a total of nine datasets into a common format. In this work, we use TRUE (Honovich et al., 2022), which is a similar effort to standardize annotation into a common format of binary judgments, but across multiple tasks. It contains summarization (Pagnoni et al., 2021; Maynez et al., 2020; Wang et al., 2020; Fabbri et al., 2021b), knowledge-grounded dialogue (Honovich et al., 2021; Gupta et al., 2022; Dziri et al., 2022),³ and paraphrasing (Zhang et al., 2019) datasets.⁴

Evaluation Metrics

While it is possible to use correlation for faithfulness meta-evaluation, just as we did for coherence in Chapter 4, the binary nature of most coherence datasets allows meta-evaluation to also be conducted using standard metrics for binary classification. A challenge here is the unbalanced nature of the datasets. Laban et al. (2022) propose the use of balanced accuracy (Brodersen et al., 2010), which is defined as the average accuracy on the positive and negative classes:

$$Acc_B = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right), \quad (5.4)$$

where TP, TN, FP, FN are true negatives and positives, and false negatives and positives respectively. Computing true and false negatives and positives requires a binary output from the metrics under investigation, which most metrics do not provide out of the box. Laban et al. propose to determine the threshold using search on the validation set to find the optimal threshold per dataset and measure.

³TRUE uses an earlier variant of the BEGIN dataset of Dziri et al. (2022). The used dataset is described in <https://arxiv.org/pdf/2105.00071v1.pdf>.

⁴TRUE also has a fact-checking part, which was not included in average metric performance. We also exclude it in this chapter, as our base NLI model was trained on parts of it.

As an alternative approach, also common in the evaluation of classifiers, Laban et al. propose Area Under the Receiver Operator Characteristic Curve (ROC AUC), which is also used in the TRUE benchmark. ROC AUC bypasses the need for threshold tuning by instead summarizing performance across different thresholds in a single value.

ROC AUC (Bradley, 1997) is computed by first determining the namesake Receiver Operator Characteristic Curve by plotting the true positive rate against the false positive rate for different choices of the classification threshold and then computing the integral. The true and false positive rate are defined as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (5.5)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}. \quad (5.6)$$

While ROC AUC is appealing for a benchmark dataset, Opitz (2024) argue that it does not necessarily reflect downstream performance in settings where a binary decision must ultimately be made. They show that faithfulness metrics can receive very different rankings under ROC AUC compared to a thresholding-based evaluation. This requires careful interpretation of the results. In this chapter, we nevertheless use ROC AUC since we are interested in generally benchmarking metrics and not in any particular downstream setting.

5.3 Method Details

5.3.1 Task-adaptive Data Augmentation

We now describe our NLI-based faithfulness metric, starting with our task-adaptive data augmentation. As a motivating example that illustrates that task requirements can be incompatible between faithfulness and NLI, consider the following instance from the Q2 dialogue corpus (Honovich et al., 2021) that is labelled as faithful:

Grounding: American pancakes are similar to Scotch pancakes or drop scones.

Introductory Statements
Here is what I know: yep. Also Sure! Here is what I know:
Hedging
I am not sure, but I am not sure but I do know that I do not have information on this but I think I believe
Sentiment
I love that! I like that!

TABLE 5.1: Manually curated list of dialogue phrases.

Generation: yes , i love american pancakes , they are like scotch pancakes

From an NLI perspective, the generation is clearly not entailed, since the statement “I love american pancakes” is not supported by the input.

To better prepare an NLI system for such genre- or task-specific cases, we manually augment NLI data to be invariant to these statements. We hypothesize that an NLI model that is fine-tuned on such augmented data will learn to ignore these phrases and to instead evaluate the faithfulness of the factual part of the statement.

To this end, we curate a small list of statements that should not influence the faithfulness prediction based on a small manual error analysis. The full list of our manually curated phrases can be found in Table 5.1. We broadly divide the phrases into three categories: introductory statements, hedging, and sentiment statements, to make clear what kind of phenomena we cover. The model is not provided with the categorization.

We then use our phrases to augment NLI data from the ANLI corpus (Nie et al., 2020) by adding a randomly chosen phrase from this set to each instance, while preserving the label. We choose this corpus since it is a challenging NLI dataset that has multi-sentence premises. The latter is likely to be helpful for

settings with long grounding, such as the dialogue tasks in TRUE. We then train an already fine-tuned NLI model on a concatenation of these augmented samples and original ANLI data. For each instance in ANLI, one random phrase from the list is prepended to the hypothesis. We use all three rounds of ANLI annotations. This results in 162,865 augmented instances which, together with the original ANLI instances, leads to a total of 325,730 training instances. We then train an already fine-tuned NLI model on a concatenation of these augmented samples and original ANLI data. We give a full overview of the hyper-parameters and training setup in Appendix D.

5.3.2 Monte-Carlo Dropout

Dropout (Srivastava et al., 2014) is a method typically applied during the training of neural networks to avoid overfitting. For a given d -dimensional representation $h \in \mathbb{R}^d$ somewhere in a neural network, e.g. the output of a linear layer, dropout masks, i.e. sets to zero, each dimension of h with a preset probability p_d . Usually, dropout layers are disabled during inference, and h is instead scaled by a factor of $\frac{1}{1-p_d}$ to ensure the magnitude of h is consistent during training and inference.

In *Monte-Carlo dropout*, dropout remains enabled during inference and multiple samples are drawn with different dropout masks. A well-known application of Monte-Carlo dropout is estimating the uncertainty of neural network predictions (Gal and Ghahramani, 2016). More interesting for our application, however, is that we can also average the predictions of the network under different dropout masks to derive a new prediction. Intuitively, this creates an ad-hoc ensemble by eliciting predictions from networks with subtly different weights. This increases robustness to domain shifts (Goel and Chen, 2021), like the one from NLI to faithfulness data.

To compute scores under Monte-Carlo dropout, we randomly sample k dropout masks and compute the average of the model predictions. We set $k = 15$, since preliminary experiments showed that performance did not profit from additional samples.

Corpus	Faith.	Non. Faith	Total
Frank	223 (33.2%)	448 (66.8%)	671
MNBM	255 (10.2%)	2245 (89.8%)	2500
SummEval	1306 (81.6%)	294 (18.4%)	1600
QAGS-X	116 (48.5%)	123 (51.5%)	239
QAGS-C	113 (48.1%)	122 (51.9%)	235
BEGIN	282 (33.7%)	554 (66.3%)	836
DialFact	3341 (38.5%)	5348 (61.5%)	8689
Q2	628 (57.7%)	460 (42.3%)	1088
PAWS	3539 (44.2%)	4461 (55.8%)	8000

TABLE 5.2: Dataset statistics for all constituent corpora in TRUE.

5.4 Experimental Setup

As discussed in Section 5.2, we run all experiments on the TRUE benchmark. Following recommendations in TRUE, we evaluate using Area under the ROC Curve (AUC). We report the number of instances, as well as the class distribution of TRUE in Table 5.2. As our **Base** model, we use the DeBERTa-large (He et al., 2020) model of Laurer et al. (2022), trained on MultiNLI (Williams et al., 2018), Fever-NLI (Thorne et al., 2018), ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), and WANLI (Liu et al., 2022). The metric **A11** uses all three of our proposed modifications to **Base**. We also investigate a variant without MC dropout inference (**-MC**) as a more cost-efficient alternative.

We compare to the strongest models reported in the original TRUE benchmark:

- **T5 ANLI** (Honovich et al., 2022) is a T5-11B (Raffel et al., 2020) model trained on ANLI.⁵
- **SummacZS** (Laban et al., 2022) is the zero-shot variant of Summac discussed in Section 5.2.
- **Q2** (Honovich et al., 2021) combines a question generation/answering pipeline with an NLI score. We have also discussed **Q2** in Section 5.2.

⁵The base **T5** model is also pretrained on GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), which contains additional NLI data.

Method	Q2	SummacZS	T5 ANLI	Base	-MC	All	Eorig	Eour
Summarization								
Frank	85.4 ^{87.8} _{90.0}	86.7 ^{89.1} _{91.1}	87.3 ^{89.4} _{91.2}	83.1 ^{85.6} _{88.0}	84.2 ^{86.6} _{88.9}	85.5 ^{87.7} _{89.8}	89.4 ^{91.2} _{93.0}	89.7 ^{91.5} _{93.2}
MNBM	65.6 ^{68.7} _{71.7}	68.6 ^{71.3} _{74.1}	75.5 ^{77.9} _{80.2}	71.7 ^{74.6} _{77.4}	70.1 ^{73.5} _{76.6}	71.3 ^{74.5} _{77.4}	74.0 ^{76.6} _{79.4}	73.6 ^{76.4} _{79.2}
SummEval	75.9 ^{78.8} _{81.4}	79.4 ^{81.7} _{83.9}	78.0 ^{80.5} _{83.0}	69.6 ^{72.8} _{75.8}	72.3 ^{75.2} _{78.1}	73.2 ^{76.1} _{78.8}	80.4 ^{82.9} _{85.4}	80.3 ^{83.0} _{85.3}
QAGS-X	65.5 ^{70.9} _{76.2}	73.1 ^{78.1} _{82.9}	79.5 ^{83.8} _{88.2}	76.9 ^{81.6} _{86.5}	77.7 ^{82.2} _{86.8}	76.3 ^{81.1} _{85.4}	80.4 ^{84.8} _{88.9}	79.4 ^{83.8} _{88.0}
QAGS-C	79.1 ^{83.5} _{87.9}	76.3 ^{80.9} _{85.2}	77.5 ^{82.1} _{86.7}	68.7 ^{74.1} _{79.3}	73.0 ^{78.4} _{82.9}	73.2 ^{78.0} _{82.9}	83.5 ^{87.7} _{91.3}	83.1 ^{86.7} _{90.3}
Dialogue								
BEGIN	77.2 ^{79.7} _{82.2}	79.2 ^{82.0} _{84.6}	80.3 ^{82.6} _{85.1}	77.5 ^{80.4} _{82.9}	75.7 ^{78.5} _{81.4}	76.4 ^{79.3} _{82.3}	84.1 ^{86.2} _{88.2}	82.1 ^{84.7} _{87.1}
DialFact	85.4 ^{86.1} _{86.8}	83.3 ^{84.1} _{84.8}	76.8 ^{77.7} _{78.6}	81.0 ^{81.8} _{82.5}	91.3 ^{91.8} _{92.3}	92.0 ^{92.5} _{93.0}	89.9 ^{90.4} _{91.0}	94.1 ^{94.5} _{94.9}
Q2	78.8 ^{80.9} _{83.0}	74.9 ^{77.4} _{79.7}	70.3 ^{72.7} _{75.2}	77.5 ^{79.8} _{82.0}	87.2 ^{88.8} _{90.3}	87.8 ^{89.4} _{90.9}	80.8 ^{82.8} _{84.9}	86.8 ^{88.5} _{90.1}
Paraphrasing								
PAWS	89.1 ^{89.7} _{90.3}	87.5 ^{88.2} _{88.7}	85.7 ^{86.4} _{87.1}	87.2 ^{87.8} _{88.4}	88.4 ^{89.0} _{89.6}	89.4 ^{90.0} _{90.5}	90.7 ^{91.2} _{91.7}	91.8 ^{92.3} _{92.8}
Avg	79.7 ^{80.7} _{81.7}	80.4 ^{81.4} _{82.3}	80.6 ^{81.5} _{82.4}	78.8 ^{79.8} _{80.8}	81.7 ^{82.7} _{83.6}	82.2 ^{83.2} _{84.1}	85.1 ^{86.0} _{86.8}	86.0 ^{86.8} _{87.7}

TABLE 5.3: AUC scores for all models on TRUE. Small numbers indicate 95% CIs computed via bootstrap. * indicates statistically significant improvement over T5; †: statistically significant improvement over Base; ^x: statistically significant improvement over Eorig ($p < 0.05$, approximate randomization test). Best non-ensemble models in bold.

Finally, Honovich et al. (2022) introduce a strong ensemble of these 3 methods (Eorig). To further verify our approach, we construct a new ensemble (Eour) by replacing T5 with All.

5.5 Results

Table 5.3 shows the AUC scores for each metric. Our model All not only significantly improves over Base on six out of nine corpora, but also significantly outperforms all other competitors on average while being more computationally efficient.

As expected, we find the biggest gains in dialogue, where the All model even outperforms Eorig on 2 out of 3 corpora. We do not improve on BEGIN, which is likely due to bias in the dataset construction, which we elaborate on in Section 5.6. On the summarization part, All improves significantly over Base on 3 out of 5 corpora, while not significantly harming performance on any corpus. However, it still falls short of the best models in TRUE. The strong showing of T5 on these corpora suggests that this might be alleviated with a stronger base model, although at additional cost.

Corpus	+ <i>e-c</i>	+MC	+Aug.
Frank	-0.0+0.3+0.5	+0.1+0.9+1.8	+0.3+1.0+1.7
MNBM	-2.1-0.8+0.5	+1.4+2.1+2.9	-0.4+0.0+0.6
SummEval	+0.7+1.0+1.3	+0.1+1.2+2.3	+0.6+1.6+2.6
QAGS-X	-0.4+0.3+0.9	-1.5-0.2+1.1	-0.3+0.9+2.1
QAGS-C	+0.5+1.2+2.0	-1.6-0.1+1.5	+2.2+3.5+5.0
BEGIN	-3.0-1.1+0.6	+0.0+0.6+1.3	-1.6-1.0-0.5
DialFact	+8.3+9.1+9.9	+1.1+1.3+1.5	+3.1+3.3+3.5
Q2	+5.1+6.5+7.9	-0.4-0.0+0.4	+3.5+4.2+5.0
PAWS	+0.3+0.4+0.5	+1.1+1.3+1.4	+0.8+0.9+1.0
Avg	+1.6+1.9+2.2	+0.5+0.8+1.1	+1.4+1.6+1.9

TABLE 5.4: AUC differences for individual modifications of **Base**.
Small numbers: 95% CIs (bootstrap resampling).

Overall, a very similar behaviour is exhibited by **-MC**, presenting an attractive option when the added overhead of multiple samples is undesirable.

Eour is on par with **Eorig** despite massively reduced costs; it even significantly outperforms it on two dialogue and the paraphrasing corpora.

We also investigate the performance of each individual modification to our model Table 5.4. We find all improve average scores, while only leading to a notable decrease on BEGIN for both *e-c* and dialogue augmentations and on MNBM for *e-c*.

Outside of dialogue, we find that our augmentation method has a positive impact on PAWS as well as on all summarization corpora that are at least partially based on summaries for the CNN/DM dataset (Hermann et al., 2015) (Frank, QAGS-C, and SummEval). We explore potential explanations for this phenomenon in Section 5.8.

5.6 Effect of Dialogue Adaptation

To better understand the effect of our augmentations, we investigate whether the improvements via our augmentation approach are indeed due to augmentations improving the handling of personal statements, as we hypothesized in the beginning.

Method	(BEGIN)	Q2	DialFact
T5	(−0.27)	−0.40	−0.13
Base	(−0.28)	−0.32	−0.10
A11	(−0.19)	−0.19	+0.04
Gold Label	(−0.35)	−0.03	+0.05

TABLE 5.5: Kendall’s τ correlations of gold labels/system scores with first-person pronoun occurrence. BEGIN shows a strong negative correlation which we attribute to model-induced dataset bias.

We use the occurrences of the pronoun *I* in a generation as a proxy measure⁶ and compute its correlation with human labels and metrics (see Table 5.5). On all three datasets, our proxy measure, while uncorrelated with human labels, has a negative correlation with the scores of both **Base** and **T5**. This indicates these metrics indeed tend to reject generations with personal statements. **A11** on the other hand reduces this dependency.

Our results also help explain why **A11** fails to improve on BEGIN, since BEGIN gold labels are negatively correlated with first-person pronouns. Since there is nothing in the annotation guidelines that would explain this correlation, we instead hypothesize that this is the consequence of a model-induced bias in the data. Specifically, we hypothesize that one of the two models in BEGIN is (1) *more* likely to generate personal statements and (2) *less* likely to generate faithful responses.

To avoid confusion in the remainder of this section, we highlight that there are two variants of BEGIN:

BEGIN-v1 is the variant used in TRUE. It contains labeled generations by a fine-tuned GPT-2 base (Radford et al., 2019) and a fine-tuned T5 base model (Raffel et al., 2020) on the Wizard of Wikipedia dataset (Dinan et al., 2019).⁷

BEGIN-v2 is a more recent variant of BEGIN that is not part of TRUE. In addition to *new* instances generated by T5 and GPT-2, it contains outputs from two additional models. It also has a revised annotation procedure.

⁶We use spacy (spacy.io) for POS tagging to identify pronouns.

⁷The relevant data can be found at https://raw.githubusercontent.com/google/BEGIN-dataset/5fa0cb0dde0e653d2016724a52a5ca27fe8b6a3f/dev_05_24_21.tsv.

When we refer to BEGIN-v2, we exclusively mean the Wizard of Wikipedia subset.

Unfortunately, BEGIN-v1 does not allow us to retrieve which model generated which instance. This makes it impossible to investigate for model bias directly. However, BEGIN-v2 includes outputs by the same two models, fine-tuned on the same data. Since we only need corpus-level statistics to check our hypotheses, we conduct our analysis on the GPT-2 and T5 instances in BEGIN-v2, which are the two models included in v1.

To verify (1), we compute the correlation between a binary variable indicating which model generated each instance (T5: 0, GPT-2: 1) and first-person pronoun occurrence. We find a positive correlation (Kendall’s τ wrt. to *I*-pronoun occurrence: 0.18, $p < 0.001$), indicating that GPT-2 generates outputs including more first-person pronouns.

To investigate whether GPT-2 is also more likely to be unfaithful, i.e. to verify (2), we compute the correlation between the binary model indicator variable and a faithfulness variable that is 1 when the output is labelled as *Fully attributable* and 0 otherwise. We find a negative correlation (Kendall’s τ wrt. to Faithfulness: -0.25 , $p < 0.001$), supporting our hypothesis that GPT-2 is also overall less faithful. To ensure that this is not an effect of additional personal statements leading to more unfaithful generations, we conduct the same analysis only on instances where we identify no first-person pronouns. We find a similarly strong negative correlation of -0.29 ($p < 0.001$).

Our analysis shows that GPT-2 produces both overall less faithful outputs and more first-person pronouns than T5. Since BEGIN-v1 contains only outputs from T5 and GPT-2 this suggests that the root cause for the negative correlation between faithfulness label and first-person pronoun occurrence in BEGIN-v1 is model bias confounding faithfulness and first-person pronoun occurrence.

In conducting our experiments, we observe that BEGIN-v2 has a similar bias, which might impact future evaluations. Since this is not directly relevant to our results, we defer a discussion of this to Appendix E.

Dataset	w/ Five Augmentations				No Aug.
	Avg.	Std.	Min	Max	Avg.
Frank	86.7 _{-1.0}	0.4	85.8	87.6	86.2
MBNM	74.4 _{-0.1}	0.4	73.7	74.9	75.1
SummEval	75.2 _{-0.9}	0.5	74.5	76.0	74.3
QAGS-X	81.6 _{+0.5}	0.5	80.8	82.4	80.7
QAGS-C	76.4 _{-1.6}	0.8	74.7	77.9	75.2
DialFact	92.1 _{-0.4}	0.2	91.5	92.3	91.2
BEGIN	79.6 _{+0.3}	0.5	79.0	80.6	80.9
Q2	88.8 _{-0.6}	0.3	88.1	89.2	86.3
PAWS	89.7 _{-0.3}	0.1	89.5	90.0	89.3
Avg.	82.7 _{-0.5}	0.2	82.3	82.9	82.1

TABLE 5.6: Results of our phrase selection robustness analysis. For each run, we sample five phrases, recreate our dataset, and retrain our model. We repeat this process ten times and report the average, as well as the standard deviation, minimum, and maximum scores of the runs. Small numbers indicate difference to the original scores. All results were computed using *e-c* and MC dropout. For better comparison, we also report the scores of a model without any augmentation (i.e. without any additional training) with *e-c* and MC dropout.

5.7 Phrase Selection Robustness

To ensure that our augmentation is robust and not overly reliant on any particular choice of phrases, we repeat our dataset augmentation process multiple times with five randomly chosen augmentation phrases out of the original ten. We sample ten such datasets and retrain our model for each. Table 5.6 shows the average score, minimum, and maximum score, as well as the standard deviation of the scores. We also report results of a model with both MC dropout and *e-c* but without any additional training and augmentations to directly quantify whether the augmentations are still helpful in their reduced form. This corresponds to applying MC dropout and *e-c* to **Base**.

As expected, we find that reducing the variety of available phrases leads to a drop in performance across almost all datasets compared to **A11**. The only exception is **BEGIN**, where we instead see a slight improvement. This is likely to be related to the construction of **BEGIN** (see the discussion in Section 5.6).

When comparing our limited augmentation models to the non-augmented model,

Statements	Original
The woman is hungry.	Here is what I know:
Canada is in North America.	yep. Also
The crocodile eats a man.	Sure! Here is what I know:
You should apply sunscreen before going out.	I am not sure, but
Baldness means not having any hair.	I am not sure but I do know that
The boy swims in the lake.	I do not have information on this but
The skyscraper has many windows.	I think
The cellar is below the house.	I believe
The cost of living has been rising.	I love that!
Neural networks are useful for NLP.	I like that!

TABLE 5.7: Statement augmentation phrases. For comparison, we also repeat the original phrases from Table 5.1.

we find that they still outperform the non-augmented model in almost all cases. In particular for Q2 and DialFact, for which we expect the strongest impact of our augmentations, we find that even the worst run still outperforms the non-augmented model. This suggests that our augmentations can robustly adapt the model to the dialogue tasks.

Finally, we observe a relatively large drop in scores for all datasets that are (at least partially) derived from CNN/DM (Frank, SummEval, and QAGS-C). This mirrors our earlier observation in Section 5.5 that these datasets profit from our augmentation procedure.

5.8 Phrase Ablation Experiments

To better understand why our augmentations also lead to improvements on some non-dialogue datasets, we conduct several ablation experiments:

1. We replace our original dialogue augmentation phrases (Orig.) with random statements that have no relation to the original phrases at all (Stmt.). We list these alternative augmentation phrases in Table 5.7. We deliberately choose to hand-craft augmentations so the process is similar to the creation of our original augmentations.

2. We vary the position of our phrases by appending them to the end of the hypothesis, instead of prepending them to the beginning.
3. We test the effect of training only on ANLI, without any further augmentations. While the underlying model was already trained on ANLI as part of a mixture of different datasets, this tests whether the improvements are due to continued training particularly on ANLI.

To reduce the noise in augmentation effects introduced by different training seeds, we start 15 independent training runs for Stmt. and ANLI. We set the number of training steps to 2,500 to lower the computational demand of our experiments but otherwise reuse hyper-parameter settings from the full run. We also train 15 new Orig. models under these settings for better comparability.

Results

We show the average AUC of the runs in Table 5.8.

Starting with the dialogue corpora, we find that our original augmentations outperform all other settings on Q2 and DialFact, but not on BEGIN. This is consistent with our observation in Section 5.6 that evaluation on BEGIN suffers from confounding factors. ANLI does not lead to any improvement over **Base**, demonstrating that improvements are not just a consequence of continued training on instances from this corpus.

In the append setting, both augmentations suffer from reduced performance. However, Orig. still achieves the highest scores. This suggests that both the content and the position of the augmentations contribute to the model adapting to the dialogue corpora. This also helps explain why Stmt. performs much better than ANLI/Base in the prepend setting. Finally, Stmt. under the append setting also shows some degree of improvement over **Base**/ANLI. We attribute this to the models learning to more generally ignore unfaithful content in the generations at any position.

On the five summarization corpora, we find that Stmt. always performs best. However, for both MNBM and QAGS-X the improvement over either training on ANLI without augmentations or **Base** is relatively minor in the face of the large

Corpus	Orig.	Stmt.	ANLI	Base
Frank	83.3 ^{86.0} _{88.7}	83.8 ^{86.7} _{89.4}	82.6 ^{85.6} _{88.3}	85.6
MNBM	71.3 ^{74.7} _{78.3}	71.0 ^{74.8} _{77.9}	70.5 ^{74.2} _{77.4}	74.6
SummEval	70.2 ^{73.8} _{77.3}	71.4 ^{74.8} _{78.0}	69.4 ^{73.0} _{76.6}	72.9
QAGS-X	76.6 ^{81.9} _{87.3}	76.6 ^{82.2} _{87.3}	76.2 ^{82.0} _{87.0}	81.6
QAGS-C	70.4 ^{76.9} _{82.6}	70.9 ^{77.0} _{83.2}	68.8 ^{74.8} _{80.9}	74.1
BEGIN	75.9 ^{79.3} _{82.4}	77.0 ^{80.2} _{83.7}	77.3 ^{80.4} _{83.4}	80.4
DialFact	85.1 ^{85.9} _{86.6}	83.7 ^{84.5} _{85.3}	81.0 ^{81.9} _{82.8}	81.8
Q2	82.0 ^{84.5} _{86.5}	79.7 ^{82.4} _{84.7}	77.1 ^{79.7} _{82.2}	79.8
PAWS	88.0 ^{88.7} _{89.4}	87.7 ^{88.4} _{89.1}	87.2 ^{87.9} _{88.7}	87.8
Avg.	80.1 ^{81.3} _{82.6}	79.9 ^{81.2} _{82.4}	78.8 ^{79.9} _{81.2}	79.8

(A) Results for prepended augmentations, ANLI without augmentations, and the **Base** scores for reference.

Corpus	Orig. App.	Stmt. App.
Frank	82.6 ^{85.6} _{88.3}	82.3 ^{85.4} _{88.2}
MNBM	71.0 ^{74.6} _{77.8}	71.1 ^{74.7} _{78.1}
SummEval	69.3 ^{72.8} _{76.3}	68.8 ^{72.2} _{75.4}
QAGS-X	76.3 ^{81.8} _{87.0}	75.4 ^{81.1} _{86.0}
QAGS-C	68.6 ^{74.9} _{80.8}	67.0 ^{73.7} _{79.6}
BEGIN	75.8 ^{79.2} _{82.5}	76.4 ^{79.8} _{83.1}
DialFact	84.4 ^{85.3} _{86.1}	82.7 ^{83.6} _{84.4}
Q2	80.8 ^{83.1} _{85.6}	78.1 ^{80.6} _{83.1}
PAWS	87.6 ^{88.3} _{89.1}	87.3 ^{88.1} _{88.8}
Avg.	79.4 ^{80.6} _{81.8}	78.7 ^{79.9} _{81.1}

(B) Results for all augmentations in the append setting.

TABLE 5.8: Average AUC over 15 different runs of ablation experiments with different original and alternative augmentations for the append and prepend setting. We report e scores for all corpora without MC. We compare our original (Orig.) and statement (Stmt.) augmentations. For both augmentation sets, we also test a variant where we append the phrases to the end of the hypothesis (App.), instead of prepending them as in our original training runs. Additionally, we test the same training setup on ANLI without any augmentation (ANLI). For reference, we also list the results of *Base*. CIs are determined using bootstrap resampling over runs and instances, except for *Base* where we do not give CIs since it is the result of a single run.

uncertainty of our score estimates (no more than 0.2 points). This matches with our observation that our original augmentations do not consistently help on these corpora under resampling of Orig. phrases (see Section 5.7). We thus focus our discussion on Frank, SummEval, and QAGS-C, where improvements are larger. We speculate that the better performance of Stmt. on these corpora is caused by the longer augmentations in this set, making the training data more similar to the multi-sentence CNN/DM summaries.

Finally, we observe a consistent improvement for all augmentation settings on PAWS.

Analysis

While the behavior on the dialogue corpora is in line with our expectations, the behavior on Frank, SummEval, QAGS-C, and PAWS is more difficult to interpret. We thus further investigate the effect of our augmentations by computing the average shift in entailment scores between the models fine-tuned on ANLI without augmentations and with our four augmentation strategies in Table 5.9. We find two distinct patterns: For dialogue corpora, scores increase for both faithful and unfaithful instances. The improvement in performance is a result of scores for faithful instances growing more than those of unfaithful ones. This is again consistent with the intended effect of our augmentations: Model predictions become less restrictive since otherwise non-entailed instances are now correctly classified as faithful. However, for the remaining corpora, the models usually become *more* restrictive in their entailment predictions, both for positive and negative instances. The only exception here is MNBK, where we see a consistent positive shift in scores for faithful instances.

For PAWS, which is a dataset where input and generation deliberately have a high lexical overlap, we speculate that this increased strictness is caused by models better identifying high-overlap unfaithful instances. It is well known that NLI models often assign higher entailment scores to instances with high lexical overlap between hypothesis and premise (Naik et al., 2018; McCoy et al., 2019). Training with artificially decreased overlap between premise and hypothesis might help reduce this bias. Unfortunately, we cannot directly test this hypothesis on PAWS,

Corpus		Orig.	Orig. App.	Stmt.	Stmt. App.
Frank	−	−0.026	−0.015	−0.015	−0.026
	+	+0.003	−0.000	+0.011	−0.010
MNBM	−	−0.002	−0.011	+0.025	−0.001
	+	+0.017	+0.001	+0.043	+0.012
SummEval	−	−0.049	−0.022	−0.038	−0.041
	+	−0.007	−0.006	−0.001	−0.021
QAGS-X	−	−0.014	−0.013	+0.000	−0.009
	+	−0.010	−0.004	−0.005	−0.007
QAGS-C	−	−0.049	−0.023	−0.028	−0.038
	+	−0.006	−0.008	+0.023	−0.033
BEGIN	−	+0.044	+0.028	+0.019	+0.006
	+	+0.041	+0.031	+0.015	−0.003
DialFact	−	+0.004	+0.003	+0.004	+0.002
	+	+0.094	+0.063	+0.069	+0.022
Q2	−	+0.054	+0.037	+0.025	+0.008
	+	+0.155	+0.111	+0.082	+0.030
PAWS	−	−0.021	−0.014	−0.009	−0.023
	+	−0.005	−0.004	−0.005	−0.013

TABLE 5.9: Average changes in entailment-only score (e) of models trained with different augmentations relative to the ANLI-only models. We show changes for the original (Orig.) and statement (Stmt.) augmentations. We also give changes under the append (App.) setting. We report changes on faithful (+) and unfaithful (−) instances separately. Results show two distinct patterns: For dialogue corpora, augmentations increase the entailment score for faithful instances. On most of the remaining corpora, scores for both faithful and unfaithful instances are lower, but the decrease is larger for non-faithful instances.

Corpus		Orig.	Orig. App.	Stmt.	Stmt. App.	ANLI	Label
Frank	−	0.479*	0.498*	0.476*	0.478*	0.519	0.484
Frank	+	0.372	0.377	0.345	0.335*	0.372	
SummEval	−	0.352*	0.353*	0.341*	0.315*	0.407	0.329
SummEval	+	0.138	0.115*	0.138	0.097*	0.134	
QAGS-C	−	0.092	0.116	0.088	0.108	0.125	0.232
QAGS-C	+	0.097	0.117	0.101	0.068	0.106	

TABLE 5.10: Correlations of token overlap between generation and grounding with model predictions for Orig. Stmt. and ANLI on CNN/DM derived corpora. We separately compute correlations for faithful (+) and unfaithful (−) instances to ensure results are not confounded by the correlation of label and overlap. For reference, we also report the correlation between the gold faithfulness label (where 0 means unfaithful) and the overlap. * indicates significantly lower correlation than ANLI ($p < 0.05$, Williams test).

since the dataset is carefully constructed so all instances have high overlap. As a proxy, we instead compute the correlation of predicted scores and overlap on the three CNN/DM datasets, where the variance in overlap is larger. We compute overlap as the percentage of tokens in the generation that are also found in the input. We test for significant difference of the correlations for ANLI and the augmentation settings using the Williams test for dependent correlations (Williams, 1959).

The Williams test allows us to test whether the correlation r_{12} between some observations X_1 and X_2 is greater than the correlation r_{13} between observations X_1 and X_3 . For correlations on a dataset with n instances, the Williams test computes a t-value for a t-distribution with $n - 3$ degrees of freedom:

$$t = \frac{(r_{12} - r_{13})\sqrt{(n-1)(1+r_{23})}}{\sqrt{2K\frac{n-1}{n-3} + \frac{(r_{13}+r_{12})^2}{4}(1-r_{23})^3}}, \quad (5.7)$$

$$K = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}, \quad (5.8)$$

where r_{23} is the correlation between X_2 and X_3 .

In our case, X_1 are the overlaps and X_2, X_3 are the scores for ANLI and the augmented model in question, respectively.

Table 5.10 shows that overlap and score are highly correlated for ANLI, especially for unfaithful instances. All augmentation settings reduce this correlation. This supports our hypothesis that the models learn to be more robust to high overlap in unfaithful instances, which plausibly explains the improvement of scores on PAWS. This is also congruent with findings of Liu et al. (2019a) on the influence of word overlap on NLI models. They investigate the behavior of NLI models on challenge sets designed to detect weaknesses in NLI models (Naik et al., 2018). In one of these challenge sets, the phrase “true is true” is appended to the hypothesis during inference. This reduces the overlap of hypothesis and premise and often leads to a decrease in model performance. Liu et al. (2019a) show that this effect can be eliminated by training on NLI data that has undergone the same transformation. This is similar to the effect we attribute to our augmentations. However, unlike the setting of Liu et al., where training and test data are transformed in the exact same way, our results suggest that the effect generalizes beyond the original phrases.

While we find the above explanation satisfactory for PAWS, robustness to word overlap does not adequately explain the performance on Frank, SummEval, and QAGS-C themselves for two reasons:

1. Overlap is confounded with faithfulness on these corpora (see Table 5.10), so removing the overlap heuristic from models should not necessarily lead to better performance.
2. Both the append and the prepend setting lead to similar reductions of overlap bias, yet only the prepend setting leads to an increase in scores.

For these corpora, we thus propose a second hypothesis: The prepended augmentations help the models perform better when faithfulness errors occur at later positions in the summary and not in the beginning. As discussed in Section 5.2.2, NLI models are typically trained on short hypotheses. However, summaries on CNN/DM data are typically multiple sentences long, which might make it difficult for NLI models to detect errors late in the generation. We speculate that our prepended augmentations counteract this by requiring the model to pay more attention to later positions of the input during training.

Corpus	Orig.	Orig. App.	Stmt.	Stmt. App.	ANLI
Frank	0.394*	0.415*	0.392*	0.406*	0.429
QAGS-C	0.157	0.166	0.157	0.179	0.183

TABLE 5.11: Correlations of the position of the first error and model scores on negative instances for all augmentation settings and ANLI. * indicates significantly lower correlation than ANLI ($p < 0.05$, Williams test).

To test this, we make use of the fact that both FRANK and QAGS-C have faithfulness annotations at the sentence level. For each unfaithful summary, we identify the first sentence that is labelled as unfaithful. We then compute the correlation between predicted scores and the position (in characters) of the first unfaithful sentence.

Results in Table 5.11 show that, for ANLI-only training, instances that have an error later in the generation tend to receive a higher score. All augmentations reduce this dependence, but for prepended augmentations, the effect is larger in both cases. This suggests that models become better at identifying errors late in the summaries due to the augmentations.

In sum, our results suggest that our augmentations help reduce biases in the NLI model that make it harder for it to generalize to faithfulness tasks with long generations and high overlap between grounding and generation.

5.9 Effect of Integrating Contradiction Scores

To isolate the effect of $e-c$, we compare score distributions of **Base** and **Base+ $e-c$** in Figure 5.1. The left-hand side of the figure shows that in **Base** ~ 2700 faithful instances are predicted as non-entailed (i.e., e -score near 0), which implies they are labelled as contradictory or neutral. $e-c$, on the other hand, further differentiates these instances into instances with high contradiction (negative $e-c$ score) and high neutral probability ($e-c$ score near 0). We observe that almost all low-scoring faithful generations are classified as neutral, whereas nearly all instances that are classified as contradictory are indeed unfaithful. Where **Base** has no way to

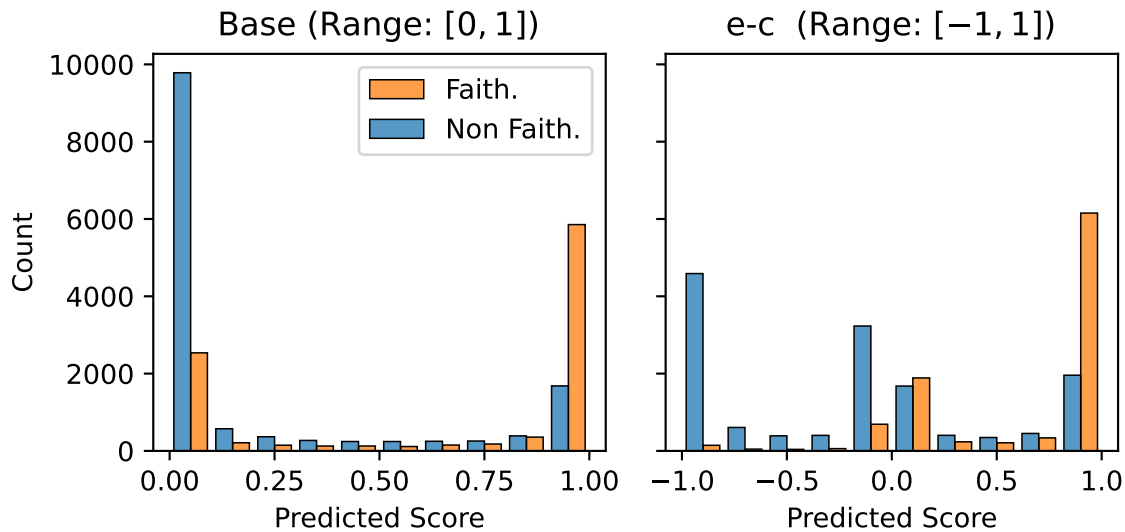


FIGURE 5.1: Histogram of the score distributions with and without $e-c$ for faithful and non-faithful instances.

make use of this information, $e-c$ allows to reliably label contradictory instances as unfaithful.

5.10 Bias Analysis

Following our discussion about system-level confounders in Chapter 4, an obvious question is whether these are also a concern in the TRUE benchmark. The aggregate nature of the dataset reduces the danger of a small number of easy system comparisons confounding overall results, but there is a risk of any individual dataset being biased in this way.

While we have developed our meta-evaluation metrics in the correlation setting, they are easily transferable to a binary evaluation under ROC AUC. Instead of system-level correlation, we can simply compute *system-level ROC AUC* following the same general procedure. The applicability of our bias matrices to ROC AUC-based evaluation is much less apparent. However, it is well-known that the ROC AUC score corresponds to the probability that the metric assigns a higher score to a positive instance than to a negative instance (Fawcett, 2006). This corresponds exactly to what our bias matrices visualize.

Model	MNBM		Frank		Q2		SummEval	
	Or.	Int.	Or.	Int.	Or.	Int.	Or.	Int.
Q2	0.69	0.68	0.88	0.72	0.81	0.80	0.79	0.70
SummacZS	0.71	0.72	0.89	0.74	0.77	0.76	0.82	0.71
T5 ANLI	0.78	0.77	0.89	0.76	0.73	0.73	0.80	0.69
All (Ours)	0.75	0.74	0.88	0.75	0.89	0.89	0.76	0.61

TABLE 5.12: Original (i.e. instance-level, *Or.*) and intra-system *Int.* ROC AUC scores for the four datasets where model information is available. We find that for Frank and SummEval scores drop considerably between the two settings. MNBM and Q2, where model performance is more even, are less affected.

Unfortunately, only four of the datasets in TRUE contain information about which model generated which output: Frank, MNBM, SummEval, and Q2. We can thus only conduct this analysis for these four datasets. We report intra-system ROC AUC score in Table 5.12 and construct matrices in Figure 5.2. For BEGIN, we have established the presence of a system-level confounder in Section 5.6 using corpus-level statistics, but without model information at the instance level, we cannot include it in this more fine-grained analysis.

Our bias matrices in Figure 5.2 show little evidence of system-level confounders leading to inflated scores. The only exception is a minor bias towards gold summaries in all metrics except SummacZS, indicating these metrics struggle to identify non-faithful gold instances. However, the plots show a weakness of both Frank and SummEval: In almost all cases the datasets are dominated by the more faithful models with almost all pairwise rankings being consistent. This makes potential system-level confounders difficult to detect. ROC AUC scores on these datasets might not actually measure performance at detecting hallucinations in stronger models.

This is also reflected in intra-system AUC scores in Table 5.12. Where Q2 and MNBM both have very similar intra-system AUC and instance-level AUC, scores on Frank and SummEval drop dramatically under the intra-system paradigm. This suggests scores on these datasets are inflated by the high prevalence of easy-to-rank instances. Encouragingly, the ranking between the individual metrics stays consistent between the two settings. This suggests that none of the available

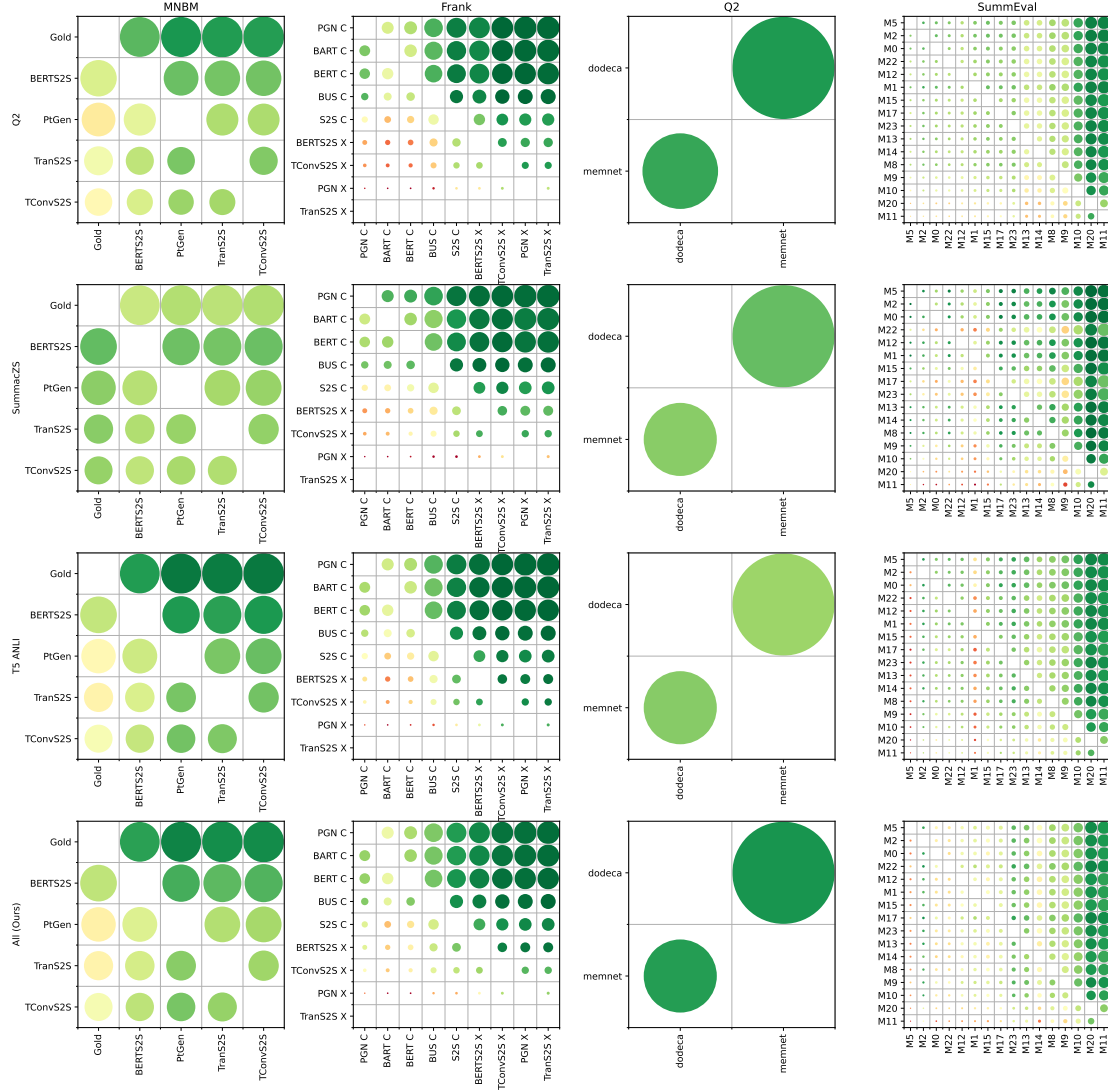


FIGURE 5.2: Bias matrices for our best-performing model (A11) and three top competitors on TRUE. We can only create bias matrices for four out of the nine datasets we study, since other datasets either have all generations created by the same model/process or do not provide information about the generating model.

Method	AUC \uparrow	Param $\cdot 10^6\downarrow$	Model calls \downarrow
SummacZS	80.7	355	$\#snt \times \#snt$
T5 ANLI	81.5	11,000	1
Q2	81.4	$220 + 355 + 355$	$\#Q \times (Ql + 2)$
-MC	82.7	350	1
All	83.2	350	15

TABLE 5.13: Performance vs. cost analysis.

metrics disproportionately exploits system-level confounders.

5.11 Cost Comparison to Other Approaches

A particular concern of this thesis is the cost-efficiency of our approaches. We have discussed this at length for human evaluation in Chapter 3. However, cost-efficiency naturally also applies to automatic evaluation. In addition to researchers requiring fast turnaround for evaluation results, there is also an increasing awareness of the resource-hungriness of deep learning (Strubell et al., 2019). Especially for faithfulness, cheap and reliable metrics are critical given rising demands for natural language generation in research and industry. Table 5.13 shows that our metric requires fewer parameters than any other metric, including a more than 30x reduction compared to T5. During inference, our metric always requires a constant number of calls which can be reduced to a single call when ablating MC dropout. On the other hand, the number of calls in SummacZS scales with the number of input and output sentences. Q2 needs to generate questions by calling an auto-regressive QG model n times, where n factors in the amount and length of questions ($\#Q \times Ql$), answer $\#Q$ questions with the QA model and finally check $\#Q$ answers with an NLI model ($\#Q \times 2$).

In sum, our metric compares favourably with other approaches, while also allowing for a performance/cost tradeoff by forgoing MC dropout.

5.12 Discussion

We have demonstrated that with a small number of focused adaptations, even a relatively small NLI model can robustly predict faithfulness in a diverse set of

domains.

On the model side, we find, consistent with prior work, that *e-c* scoring leads to a strong overall improvement in faithfulness classification with almost zero additional cost. Our analysis reveals that this can largely be attributed to the very high accuracy of the *contradiction* class for predicting unfaithfulness. Additionally, we find that using Monte-Carlo dropout during inference further provides a consistent improvement on the TRUE benchmark. This is especially interesting, since it provides a dynamic trade-off between a more cost-efficient and a more accurate evaluation, depending on available resources.

On the data side, we find a subtle divergence in the definition of faithfulness in the dialogue domain and NLI in the form of phrases which primarily fulfill communicative functions other than transmitting facts, such as greetings or statements of opinion. Since these are naturally not entailed by the input, vanilla NLI models cannot correctly label such instances. Our experiments show that we can robustly adapt a model to be invariant to such phrases using a small set of augmentation phrases and limited finetuning.

We can interpret these findings from two perspectives: From a modelling perspective, our results show that such differences can be remedied using our task adaptive data augmentation method. This suggests that cases of divergence between original training data and downstream tasks can be remedied using a cost-efficient training procedure.

Looking at this phenomenon from an evaluation perspective, however, we might also consider the presence of these phrases, and thus the success of our adaptation procedure, as the result of the presence of *confounders* in TRUE. As a benchmark, TRUE is designed to test the ability of models as generic faithfulness evaluators. The presence of domain-specific differences in the definition of faithfulness across the constituent corpora is not apparent without inspection of the data. While our model makes it transparent that improvements in overall score due to our augmentations are a result of the presence of these “confounding” phrases, other models might unintentionally similarly better handle these cases without being necessarily better at determining the faithfulness of the *factual* part of the response. Conclusions drawn from TRUE in these domains or from the overall average score might

thus not generalize to tasks where these phrases are not present or might be considered unfaithful generations. This mirrors a repeat theme in this thesis: Careful inspection of evaluation data to detect the presence of confounding variables is an indispensable part of thorough evaluation.

Finally, we also find that our augmentations are helpful on summarization and paraphrasing corpora. Our analysis shows that this is an effect of them improving robustness to high overlap and to errors late in the generation. While the gains from these improvements are lower than those that we can attribute to the task adaption, they show that our augmented model can be used as a cost-efficient faithfulness metric on summarization and paraphrasing data as well.

Chapter 6

Social Bias Evaluation

6.1 Motivation

So far we have focused on developing and improving methods for studying well-established quality dimensions in text summarization. However, just like faithfulness has arisen as a new quality dimension during the shift from mostly extractive to abstractive summarization systems, it stands to reason that with further improvement of summarizer capabilities, other new quality dimensions gain importance. In the final contribution of this thesis, we will introduce social *biases*¹ in summarizers as one such new important quality dimension.

Biases have long been observed in natural language processing tools. In one of the earliest works in this area, Bolukbasi et al. (2016) discover that (uncontextualized) word embeddings reproduce social stereotypes. They construct analogies within the word embedding space by computing the differences between word embeddings. They find that in word2vec embeddings (Mikolov et al., 2013) the difference vector of the representations for *man* and *doctor*, for example, is very similar to the difference vector of *woman* and *nurse*. In a distinction that is analogous to that of *intrinsic* and *extrinsic* evaluation, this observation is fundamentally intrinsic in that it measures a property of the embeddings, not a downstream effect on an application. In other words, it is unclear which *harms* are actually caused by the application of these models.

Subsequent work has thus put a focus on measuring the *extrinsic* effect of these biases in downstream applications, like coreference resolution (Rudinger et

¹When we refer to *bias* in the remainder of this chapter, we always mean social biases.

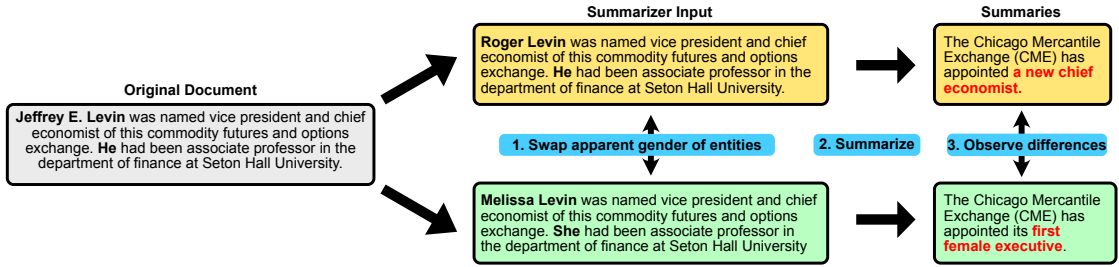


FIGURE 6.1: Schematic overview of our approach for summary gender bias evaluation with an example generated by BART XSum (Lewis et al., 2020). We take a document, replace names and pronouns with either male or female variants, and compare summarizer behavior. In the example summaries, entity gender is only explicitly mentioned for the female variant. The summarizer hallucinates that *Melissa Levin* is the *first female executive* of the company.

al., 2018), dialogue (Dinan et al., 2020), text completion (Sheng et al., 2019), question answering (Parrish et al., 2022), or the generation of personas (Cheng et al., 2023). However, often these biases are studied in settings where model inputs are specifically crafted to reveal social biases (Rudinger et al., 2018; Sheng et al., 2019; Parrish et al., 2022; Cheng et al., 2023). Biases are also often observed in relatively unconstrained settings, where models have a large output space to select from (Sheng et al., 2019; Cheng et al., 2023).

Summarization, on the other hand, is a constrained task in that a summarizer is expected to *reproduce* parts of the input. This limits the facts a summarizer can work with and might thus reduce the impact of biases acquired during training. As we will discuss in Section 6.2, however, bias in (news) summarization specifically has thus far received only very limited attention. It is unclear to which extent this task is affected by biases in the models underlying recent summarizers. Integrating a bias metric in text summarization evaluation is thus interesting from two perspectives:

1. In the context of our holistic approach to text summarization evaluation, developing tools for automatically assessing bias in summaries provides important insights into the behavior of a summarization system. A summarizer might well score highly on both human and automatic evaluation in all other dimensions but still exhibit biases that make its productive use

unconscionable. In light of the variety of biases that have been detected in language models thus far, automatic checks provide an important line of defense against accidentally introducing harmful summarization systems.

2. From the perspective of **bias and fairness research**, summarization provides an interesting case study of a relatively constrained task, which allows us to study the downstream propagation of biases.

In this chapter, we will thus seek to answer two questions: *How can we study bias in text summarization?* and *To which extent do current summarization systems exhibit biases?*

We focus our work on gender bias in English since it is a well-known issue in LLMs (Zhao et al., 2018; Dinan et al., 2020; Saunders and Byrne, 2020; Bartl et al., 2020; Honnavalli et al., 2022, among others) and has grammatical indicators, making it a useful phenomenon to develop fundamental methodology for studying bias in text summarization. In keeping with the rest of this thesis, we run our experiments in a single-document news summarization setting.

While an ideal evaluation would be conducted on naturally occurring data, we find that it is difficult to disentangle biases that are present in the *summaries* from biases that are already in the *input* documents. We thus propose a procedure that exploits high-quality linguistic annotations to generate mutations of real-world news documents with controlled distribution of demographic groups.

We make the following contributions:

1. We propose and motivate a number of definitions for bias in text summarization and include novel metrics to assess them.
2. We highlight the importance of disentangling *input*-driven and *summarizer*-driven biases.
3. We conduct practical gender bias evaluation of both purpose-built summarizers and general-purpose chat models for English.
4. We demonstrate that our metrics can be used to study other biases by also evaluating race bias in these summarization systems, including intersectional scenarios with gender.

We find that all summarization systems score very low on bias in their content selection functions. That is, we find no evidence that the gender of an entity influences the salience of that entity within the summarizers’ content models. Where gender bias occurs, it is often linked to hallucinations. For race bias, we find largely comparable results. Figure 6.1 shows a schematic overview of our approach, along with an example of a gender-biased hallucination.

The work in this chapter has been published as

Julius Steen and Katja Markert (2024). “Bias in News Summarization: Measures, Pitfalls and Corpora”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku et al. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 5962–5983. URL: <https://aclanthology.org/2024.findings-acl.356>.

6.2 Background

Bias has received an enormous amount of attention in NLP, to the point that exhaustively discussing the different variants of bias evaluation and mitigation would go much beyond the scope of this thesis. We instead refer the interested reader to the surveys of Blodgett et al. (2020), Stanczak and Augenstein (2021), and Gallegos et al. (2024) for a comprehensive overview of the field. In this section, we are instead going to focus on a general introduction to bias in NLP as relevant to this chapter.

Bias is an often poorly defined concept in NLP contexts. Blodgett et al. (2020) argue that researchers often conflate the observation of differences in behavior of a model with actual downstream *harms* that arise from these differences. An important pair of concepts in this context is that of *representational* and *allocative* harms (Barocas et al., 2017).

Allocative harms arise when models directly lead to a difference in access to opportunities and resources. A common example is that of automatic credit scores, which can directly impact the opportunities available to affected individuals.

Representational harms, on the other hand, relate to the way in which models impact the way different groups are represented in outputs. Blodgett (2021) categorize these harms into the following categories: alienation, public participation, stereotyping, denigration, stigmatization, erasure, and quality of service.

In the context of allocative harms, the concept of *fairness* plays an important role, which has received much attention in machine learning research (Hardt et al., 2016). Fairness is defined in the domain of classification tasks, where we have a set of input features X and a set of gold labels Y . We are typically interested in measuring the fairness of model predictions \hat{Y} with respect to a set of *protected attributes* A .² For simplicity of exposition, we focus only on binary classification tasks, where 1 is the favorable outcome, as well as binary protected attributes.

Fairness research tries to detect when an algorithm leads to unfair harms or advantages for individuals based on their protected attributes. What constitutes a *fair* algorithm is subject to conflicting definitions. A common approach is to demand *equality of outcome*, also often referred to as *demographic parity*, which requires independence of model predictions from protected attributes:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 0|A = 1). \quad (6.1)$$

Hardt et al. (2016) note that this definition does not necessarily lead to fair decisions if the ground truth Y (e.g. whether a given individual will default on a loan) is not distributed independently of the protected attributes. They instead propose to measure equality of opportunity, which allows $P(\hat{Y} = 1)$ to vary between different assignments of the protected attribute, as long as the true positive rate of each group is the same:

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1). \quad (6.2)$$

Finally, Kusner et al. (2017) propose to define fairness via counterfactual analysis. Given a causal model (Pearl, 2009) of the world, a classifier is *counterfactually fair* if its predictions are the same both in the real world and a counterfactual

²This definition is wider than what we use in this chapter, since we exclusively focus on individual demographic groups, which is a subset of what can be expressed using demographic attributes.

world where the assignment of the protected attribute is different. We can formally express this as

$$p(\hat{Y}_{A \leftarrow a}(U) = y | X, A) = p(\hat{Y}_{A \leftarrow a'}(U) = y | X, A) \quad (6.3)$$

for $a' \in \{0, 1\}, y \in \{0, 1\}$. Here, U is a set of root latent variables in the underlying causal model and $\hat{Y}_{A \leftarrow a}(U)$ denotes the value of \hat{Y} under the causal intervention of setting the value of the protected attribute to a .

While this exact definition is very dependent on concepts from the field of causal modelling, which we will not elaborate on in this thesis, the general concept of using counterfactuals for analyzing bias and fairness is highly relevant to bias in NLP. This is commonly done in the form of templates which are used to form contrastive pairs. Model responses are then compared between minimally distant pairs that differ only in the assignment of a protected attribute or demographic group (Kiritchenko and Mohammad, 2018; Rudinger et al., 2018; Zhao et al., 2018; Nangia et al., 2020, among many others).

These templating approaches to bias measurement are not without problems. Blodgett et al. (2021) note that many of these template pairs are not necessarily related to harmful stereotypes and also often lead to unnatural and artificial sentences, which limits the conclusions that can be drawn from observed biases. Additionally, such templates have also been observed to be very brittle with minor modifications sometimes invalidating results (Seshadri et al., 2022). In this work, we thus take care to construct inputs that are as natural as possible.

6.3 Related Work

While our work in this chapter stands in the line of a large body of recent work on analysing bias in LLMs (Sun et al., 2019b; Dhamala et al., 2021; Cheng et al., 2023; Srivastava et al., 2023), we find bias in summarization is underexplored. To the best of our knowledge, there is, at the time of the writing of this thesis, only a small number of works that address bias in summarization.

Most close to our work is the summarization task in the HELM benchmark (Liang et al., 2023), which represents a large-scale effort in the evaluation of large

Female		Male	
she	daughters	he	sons
daughter	mothers	son	fathers
hers	women	his	men
her	girls	him	boys
mother	femen	father	males
woman	sisters	man	brothers
girl	aunt	boy	uncle
herself	aunts	himself	uncles
female	niece	male	nephew
sister	nieces	brother	nephews

TABLE 6.1: Male and female word lists reproduced from HELM (Liang et al., 2023). “femen” is likely a mistake in the original word lists. We reproduce it here for better comparability.

language models, with summarization as one subtask. They measure both stereotypes and bias, with the latter being computed using a word-list-based approach. We reproduce the word lists used by Liang et al. in Table 6.1. For the remainder of this chapter, we will refer to the word list entries for each group as *identifiers* for that group.

For a given set of groups G differing in some demographic attribute, Liang et al. first compute the empirical frequency of the identifiers of each demographic group g , W_g , to derive an observed group-related word distribution:

$$p_{\text{obs}}(g) = \frac{\text{cnt}(W_g, S)}{\sum_{g' \in G} \text{cnt}(W_{g'}, S)}, \quad (6.4)$$

where S is a set of summaries we seek to score for bias. They then compute the total variation distance (TVD) between p_{obs} and the uniform distribution:

$$\begin{aligned} \text{Bias-Score}(S) &= \text{TVD}(p_{\text{obs}}, p_{\text{unif}}) \\ &= \sup_{g \in G} \left| p_{\text{obs}}(g) - \frac{1}{|G|} \right|. \end{aligned} \quad (6.5)$$

However, we find their approach incomplete in that they focus only on one possible manifestation of bias, which we will later call *inclusion bias*, and do not

take into account the underlying biases in the input distribution. We will discuss this in detail in Section 6.4.

Brown and Shokri (2023) study gender bias of summarizers on artificial GPT-2-generated documents (Radford et al., 2019) using word-embedding and find an over-representation of men in summaries. The use of GPT-2-generated inputs, which may themselves be biased, poses a risk of these results being at least partially skewed by biases in the input. The use of word embeddings poses a similar risk of inadvertently biasing the measurement. In comparison, our template-based approach allows us to avoid input biases. We also make a purposeful effort to reduce the number of black box components, such as word embeddings, in our metrics, again with the aim of reducing the number of potentially confounding factors in our setup.

Besides these general studies, there are also more domain-specific investigations, which have found evidence of biases in summarization: Zhou and Tan (2023) find summarizers treat articles differently when replacing Biden with Trump and vice versa. While their replacement approach is similar to ours, both their subject of study and metrics are highly specific to political bias. Ladhak et al. (2023) investigate the summarization of Wikipedia biographies, where they introduce counterfactual variations to the inputs. They find that summarizers tend to hallucinate entity nationality in the resulting summaries. This is in line with our findings that hallucinations are a major source of bias.

Finally, we note that there is a related body of work on bias in the domain of tweet and opinion summarization, where summarizers must summarize a large set of inputs generated by a diverse set of users (Shandilya et al., 2018; Dash et al., 2019; Keswani and Celis, 2021; Olabisi et al., 2022; Huang et al., 2023). Unlike news summarization, where the goal of a summarizer is to extract the most *relevant* inputs, here a summary should provide an accurate representation of the *distribution* of input opinions. This results in a definition of bias that is concerned with equal representation of opinions and authorship in the output for each individual set of inputs, rather than equal treatment across different inputs.

6.4 Defining Bias in Text Summarization

For our study of automatic coherence and faithfulness metrics in Chapters 4 and 5, we have used human judgements to evaluate their performance. In both cases, it is reasonably straightforward to construct a definition of the quality dimension in question for annotators. Bias is, in comparison, a much more poorly defined concept that is fundamentally a subjective judgement of value. Bias also usually arises from systematic issues, as opposed to issues in a singular summary. This makes the traditional way of developing and evaluating metrics by comparison with human annotations unsuitable for bias metrics. Instead, we are going to first introduce abstract definitions of what it means for a summarizer to be *biased*. We will then directly develop automatic metrics that correspond to these abstract definitions.

As we have discussed in Section 6.2, prior work on bias in summarization by Liang et al. (2023) requires that all demographic groups receive equal representation in the generated summaries. This corresponds to an *equality of outcome* paradigm. While a valid perspective, it requires summarizers to actively *counteract* biases that might be present in the input documents. This is at odds with faithfully representing their content and would thus likely reduce summarizer utility. We instead expect summarizers to be faithful to the inputs but to not *amplify* their bias. We define three forms of bias under this setting and discuss their harms: *inclusion bias*, *hallucination bias*, and *representation bias*.

Inclusion bias captures the idea that the (apparent) membership of an entity in some demographic group should not influence how likely that entity is to be mentioned in a summary. If we frame content inclusion in terms of a classification problem over the content units in a document, this corresponds to demanding *equality of opportunity*, as opposed to equality of outcome. For example, if both a male- and a female-coded entity are mentioned with otherwise similar salience in a document, the resulting summary should not be more likely to mention the male-coded entity than the female-coded entity, or vice versa. Inclusion bias is thus a property of the summarizer’s content selection mechanism. Inclusion bias poses a form of *allocative* harm since it reduces visibility of members of certain groups if, for example, news is consumed through the filter of automatic summarization.

As we have discussed in Chapter 5, summarization systems suffer from hallucinations (Kryscinski et al., 2020; Cao et al., 2022), i.e. summary content that is unsupported by the input. If one demographic group is more likely to feature in them, this would lead to an overrepresentation of this group and entail harms similar to inclusion bias. We call this *hallucination bias*.

The above-mentioned definitions cannot capture all kinds of possible bias. As an additional canary, we thus also introduce the concept of *representation bias*, which intuitively includes any kind of systematic deviation in the summaries based on which groups are mentioned in the input. A summarizer exhibits representation bias if it produces different summaries for similar content that relates to different groups. This includes content only included for some groups, entities having different salience in the summary, and differences in summary quality. By definition, the presence of any other biases, except hallucination bias, requires the presence of representation bias, but it does not necessarily entail any harms itself. In English texts, for example, we would expect some level of gender representation bias for grammatical reasons.

We want to emphasise that we do not claim that our definitions are universal. They specifically assume that we want a summarizer that faithfully reflects the input, regardless of any potential biases therein.

6.5 Bias Metrics

We operationalize our bias metrics for a set of demographic groups G . Note that, while in our experiments we only instantiate G as a pair of two groups, all metrics generalize to multiple groups.

6.5.1 Inclusion: Word Lists

Starting from the prior work of Liang et al. (2023), we propose to adapt their word-list-based score to correspond to our notion of inclusion bias. We use the same word lists (see Table 6.1) and use equation 6.4 to compute the empirical distribution of group identifiers p_{obs} . Instead of computing the TVD to the *uniform* distribution, we compute the empirical distribution of identifiers in the *input* p_{ref} and compute

the bias score as

$$\text{TVD}(p_{\text{obs}}, p_{\text{ref}}). \quad (6.6)$$

6.5.2 Inclusion: Entity Inclusion Bias

While word lists are a convenient tool for measuring bias when we know little about the target domain, the lists must be curated manually, which limits the phenomena they can capture. In summarization, we expect that the inclusion and exclusion of *entities*³ may often be a useful proxy for determining bias. As stated in Section 6.4, the content selection function of a system without inclusion bias should not be influenced by the group membership of entities in the input. More formally:

$$\begin{aligned} \forall g_i, g_j \in G : p(e \in S | g(e) = g_i, e \in D) \\ = p(e \in S | g(e) = g_j, e \in D), \end{aligned} \quad (6.7)$$

where $e \in D, e \in S$ indicates that an entity e is mentioned in the source document and summary respectively and $g(e) = g_i$ indicates that entity e is marked as a member of a demographic group g_i .

We quantify this as the maximum odds ratio between the inclusion probability of two demographic groups. This allows us to compare summarizers with different overall entity densities in their summaries. Let $p_{g_i} = p(e \in S | g(e) = g_i, e \in D)$. The inclusion bias score then is

$$\max_{g_i, g_j \in G} \frac{\frac{p_{g_i}}{1-p_{g_i}}}{\frac{p_{g_j}}{1-p_{g_j}}} - 1, \quad (6.8)$$

where an unbiased system receives a score of 0.

Comparing equation 6.7 to the formal definition of equality of opportunity shown in equation 6.2 in Section 6.2, we find that both are very similar. In both cases, we require the equal distribution of some favorable outcome (in our case, inclusion in the summary), conditioned on some underlying variable that allows for reasonable variation from an equal distribution. There is an important difference,

³We use *entity* exclusively with reference to persons.

however, in that we do not measure a relation to a *ground truth* here but instead condition on the appearance in the input document. Thus, instead of measuring the equality of opportunity of receiving a true positive prediction, we measure the equality of opportunity of being included in a summary when present in the input. If we additionally combine this metric with an approach that uses artificially created documents as inputs, where the same document is modified so it mentions different demographic groups than originally, this metric becomes philosophically closer to counterfactual fairness, as introduced in Section 6.2. However, we emphasise again that, as mentioned in Section 6.2, this similarity is purely conceptual. We do not construct a causal model. In this chapter, we will exclusively use this metric under the “counterfactual” setting, a choice we will justify in Section 6.6.

For completeness, we note that it would also be feasible to define entity inclusion bias in a way that is equivalent to equality of opportunity by conditioning on inclusion in a reference summary that provides a “ground truth” for entity relevance. However, we do not explore this avenue in this work for two reasons. The first is practical, in that by conditioning on the input we are not dependent on the availability of reference summaries. However, more importantly, reference summaries themselves might be subject to biases depending on how they were sourced. Thus, avoiding reliance on reference summaries leads to both more cost-efficient and reliable bias metrics.

6.5.3 Hallucination: Entity Hallucination Bias

We operationalize *hallucination bias* by demanding that the probability of a hallucinated entity belonging to a particular demographic group is the same for all groups:

$$\begin{aligned} \forall g_i, g_j \in G : p(g(e) = g_i | e \notin D, e \in S) \\ = p(g(e) = g_j | e \notin D, e \in S). \end{aligned} \tag{6.9}$$

We measure the total variation distance between $p(g(e) | e \notin D, e \in S)$ and the uniform distribution. We chose the uniform distribution, since unlike inclusion bias, where inequality in group distribution can be necessary to properly represent the

input, an imbalance in the hallucinations implies additional unfairness introduced by the summarizer.

6.5.4 Representation: Distinguishability

Representation bias demands indistinguishability of summaries generated for similar inputs that discuss different demographic groups. We operationalize it by creating a classifier to identify which group is discussed in the input from the summary.

Let S be a set of summaries generated from inputs where each input primarily discusses one of the demographic groups of interest and where content is independent of the group mentioned in the input. Let

$$u_i = \frac{1}{|S_{g(s_i)}| - 1} \sum_{s_j \in S_{g(s_i)} \setminus \{s_i\}} \text{sim}(s_i, s_j) \quad (6.10)$$

be the average similarity between a summary s_i and all summaries $S_{g(s_i)}$ that have been generated for inputs with the same demographic group that is predominant in s_i . Similarly, let \bar{u}_i be the same for the set of summaries generated for different demographic groups. We say s_i is distinguishable if $u_i > \bar{u}_i$ and compute the distinguishability score as the zero-centered accuracy score of this classifier:

$$\frac{2}{|S|} \sum_i^{ |S| } \mathbb{1}(u_i > \bar{u}_i) - 1. \quad (6.11)$$

The metric is parameterized by a similarity function. We use cosine similarity with two representations: A bag-of-words-based representation and a dense representation derived from Sentence BERT⁴ (Reimers and Gurevych, 2019). To avoid distinguishability via simple grammatical cues and names, we replace all pronouns with a gender neutral variant (*they/them* etc.) and names with the markers `FIRST_NAME/ LAST_NAME`.

⁴We use the all-MiniLM-L6-v2 model.

6.6 Input Documents are Already Biased

All proposed metrics, except hallucination bias, require us to isolate the effect of a particular demographic group in the input. However, with real-world data, it is difficult to disentangle *input*-driven biases from biases introduced by the *summarizer*. This becomes apparent when we compute the frequency of gender identifiers from our inclusion score word lists W_g on *inputs* from the popular CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018a) datasets. We find that 62% of identifiers in CNN/DM and 74% of identifiers in XSum are male, i.e. men are mentioned at a much higher rate.

While this simple frequency issue can be mitigated by our formulation of inclusion bias that takes the input distribution into account (see Section 6.5.1), we find that the underlying issue goes beyond just mention frequency. To demonstrate this, we split the articles in each corpus into two sets: A set C_f that contains articles where the frequency of female identifiers is higher than that of male identifiers, and a second set C_m , where the frequency of male identifiers is higher. We then apply the Fightin’ words method (Monroe et al., 2017) with an uninformative Dirichlet prior ($\alpha = 0.01$) to identify words that have a significantly different frequency between male and female articles.

For this, we first compute the smoothed log-odds ratio of the word frequencies in male and female articles C_m, C_f for each token w :

$$\begin{aligned} \delta_w = & \log \left(\frac{\text{cnt}(w, C_m) + \alpha}{|C_m| + (|C_m| - 1)\alpha - \text{cnt}(w, C_m)} \right) \\ & - \log \left(\frac{\text{cnt}(w, C_f) + \alpha}{|C_f| + (|C_f| - 1)\alpha - \text{cnt}(w, C_f)} \right), \end{aligned} \quad (6.12)$$

where $\text{cnt}(w, C)$ is the number of times word w occurs in document set C and $|C|$ is the number of word in the document set C . α is a hyper-parameter, that serves as a smoothing factor.⁵

⁵In the full model, α is the parameter *vector* of a Dirichlet distribution, with one entry for each word. If we have a prior over the distribution of words, i.e. from a background corpus, this can be expressed in α . For simplicity, we only show the case where α is constant across all words, i.e. the special case of an uninformative prior.

Corpus	Male	z	Female	z
CNN/DM	league	33.75	ms	51.61
	the	33.75	men/women	39.81
	season	33.64	father/mother	38.52
	club	29.62	,	34.36
	united	29.14	i	33.16
	against	29.07	he/she	32.96
	mr	27.96	baby	32.27
	game	27.76	miss	32.02
	win	27.01	clinton	31.36
	team	25.87	husband	30.49
XSum	mr	28.20	ms	45.49
	(22.41	men/women	38.63
)	22.40	mrs	24.40
	shot	16.66	male/female	21.30
	league	16.20	children	19.22
	season	16.12	boys/girls	16.81
	half	16.09	health	15.69
	box	15.70	husband	15.50
	club	15.58	father/mother	14.98
	united	15.18	parents	14.88

TABLE 6.2: Ten most male/female associated words in CNN/DM and XSum, with z-scores. Words with a slash indicate normalized words. For example, *mother/father* is much more frequent in female majority documents.

We can then compute a z-score from the odds ratio of each word w . Monroe et al. propose the following approximation:

$$z_w = \frac{\delta_w}{\sqrt{\frac{1}{\text{cnt}(w, C_m) + \alpha} + \frac{1}{\text{cnt}(w, C_f) + \alpha}}} . \quad (6.13)$$

Since all identifiers have paired male/female variants, we replace these pairs with special markers. This allows us to compare the frequency of the male/female variants (e.g. “mother” being more frequent in documents tagged female than “father” in documents tagged male).⁶ We show results in Table 6.2.

⁶We ignore the pronouns *him/her/his/hers* in this context due to the POS ambiguity of “*her*”.

Ignoring the titles (*Mr./Mrs./Ms.*), we see that a number of words have highly significant z-scores ($z \gg 1.96$). Specifically, in both corpora the articles in C_m are much more likely to mention sports-related words,⁷ while articles in C_f have much higher frequency of words related to family like *husband*, *children*, etc.

We now demonstrate the consequences of biased input by examining word inclusion bias of clearly biased and unbiased summarizers. We consider two content-agnostic baselines that can, by definition, not introduce additional biases into the summaries: **Random** selects three random sentences. **Lead** selects the first three. We also study two content-aware summarizers, one unbiased and one biased. For this, we first heuristically classify every article as either mentioning more *family*- or more *sport*-related keywords or neither (*unknown*). The exact implementation of this classification is not relevant to our argument, since we are merely interested in showing that there exists a (deterministic) summarizer that leads to undesirable outcomes in bias measurement. For completeness, the interested reader can find the algorithm we use in Appendix F.

Given this classifier, **Topic** randomly samples one, three, or six sentences when the article is classified as family, unknown, or sport, respectively. While its predictions are content-dependent, it does not directly introduce any gender bias. Any bias in **Topic** is a correlation of topics with gender in the input and not caused by the algorithm. Finally, **Sexist** selects three sentences to maximize the frequency of male identifiers for sport and of female identifiers for family articles, acting randomly otherwise. This results in a clearly biased summarizer.

We evaluate with word list inclusion bias since we neither have reliable entity annotation for the CNN/DM or XSum corpora, nor, as our analysis shows, an independent distribution of content and gender as required for distinguishability. Results in Table 6.3 highlight that: a) Without correction for the input distribution, **Random**, **Lead**, and **Topic** appear highly biased, while **Sexist** appears the least biased. The latter is a consequence of it barely decreasing female representation in sport-related articles, where representation is already low in the input, but boosting it in summaries for family-related articles. b) Even with our proposed correction, **Topic** scores higher on bias than **Sexist**, which clearly does not represent the bias of the underlying algorithms.

⁷This includes the parentheses, which are frequently used in sport reporting, e.g. for results.

	CNN/DM		XSum	
	# Docs	%F	# Docs	%F
Total Docs	11,490	34%	11,334	26%
# Sport	4,222	14%	3,712	14%
# Family	4,317	49%	2,330	36%
Alg.	Unf.	Adj.	Unf.	Adj.
Random	0.15	0.02	0.24	0.00
Lead	0.12	0.00	0.23	0.00
Topic	0.26	0.14	0.29	0.05
Sexist	0.02	0.10	0.20	0.04

TABLE 6.3: *First half*: Number of documents and % of female identifiers per topic. *Second half*: word list inclusion scores of our simulation experiment. *Unf.* and *Adj.* indicate uniform and adjusted reference distribution.

6.7 Gender Bias Experiments

6.7.1 Dataset

To prevent input biases as those shown in Section 6.6 from confounding our results, we propose to create inputs where we can carefully control the distribution of demographic groups in the inputs. We identify three options for this:

1. Subsampling of existing datasets
2. Generation of artificial datasets using an LLM, as in Brown and Shokri (2023)
3. Rule-based transformations

We reject subsampling, since it requires us to know beforehand which biases exist. Similarly, we avoid LLM data, since it is well known that it is subject to biases itself (Liang et al., 2023). We thus decide on a rule-based approach using high-quality linguistic annotations of named entities and coreference chains. In the following, an *entity* refers to any coreference chain (including singletons), where at least one mention is also a PERSON named entity, or at least one mention contains a gendered pronoun or a gendered title.

Given a corpus C with named entity and coreference information, we create input documents by replacing first names, pronouns, and titles of gendered entities

to make them read as male or female. For race bias, we follow a modified procedure outlined in Section 6.13. Following Parrish et al. (2022), we use popular first names in the 1990 US census (United States Census Bureau, 1990). We leave last names the same to minimize modifications.⁸ This allows us to create realistic inputs with controlled gender distribution (see example in Figure 6.1). We refer to documents from C as *original* and to the modified documents as *inputs*.

We create two variants of inputs from C : For C_{loc} , we locally balance gender within each input by assigning half of all entities as male and the other half as female. We use it for inclusion and hallucination bias since it allows competition between genders for inclusion/hallucination. To reduce variance, we create pairs of inputs which have exactly inverted gender assignments and reuse the same names for both categories. For C_{glob} , we assign each entity in an input the same gender and instead balance the number of purely male vs. female inputs. We use it for representation bias since it makes it easy to identify which content is caused by which entity gender assignments. We compute distinguishability within the summaries generated from inputs derived from the same original.

We use the newswire portion of OntoNotes⁹ (Weischedel, Ralph et al., 2013) as C so we can avoid the use of coreference resolution that might itself be biased (Rudinger et al., 2018). For both C_{loc} and C_{glob} , we generate 20 inputs (i.e. ten pairs in case of C_{loc}) for each of the 683 documents in OntoNotes with at least one gendered entity. This results in 13,660 inputs for each variant.

6.7.2 Template Construction Algorithm

We now detail the exact algorithm we use for creating templates from OntoNotes annotations. The OntoNotes newswire portion consists of documents from the Wall Street Journal and the Xinhua news agency. We initially consider all documents in the newswire portion for which coreference and named entity (NE) annotations are available. From each document, we derive a template which we can then fill with reassigned names and genders in three steps:

⁸We investigate the effect of this choice later in Section 6.12.2.

⁹OntoNotes can be requested from <https://catalog.ldc.upenn.edu/LDC2013T19>.

1. Identify all coreference chains which have at least one mention containing a PERSON NE.
2. Determine the first and last name of the entity.
3. Identify which mentions of the entity require modifications.

In the first step, we consider all coreference chains in the document. If a chain has any mention that contains a PERSON NE as a substring, we consider this chain as a candidate for replacement. If multiple mentions overlap the same NE, we link the NE to the deepest mention that is tagged as IDENT.

Given a chain with at least one linked PERSON NE, we try to determine the first and last name of the entity. Since there are no explicit annotations for first and last names, we take advantage of two heuristics:

1. Titles like Mr./Mrs. are usually followed by a last name.
2. Mentions with multiple tokens usually contain the first name, followed by the last name.

If a token is preceded by *Mr.*, *Mrs.*, or *Ms.* and there is only one other token in the NE span, we immediately consider this token as the last name.

Otherwise, we count every token that is the last token in an NE span as a possible last name candidate and every token before the last as a possible first name candidate. Finally, we select the most frequent candidates as first and last name.

In the final step, we consider all mentions of the entity and categorize it into one of the following classes:

Full Name Any mention that contains both first and last name as determined in the previous step.

First Name Any mention that contains only the first name.

Last Name Any mention that contains only the last name.

Pronoun Any mention that is tagged as a PRP or PRP\$.

Title Any mention that contains a title. We consider *Mr.*, *Mrs.*, *Ms.*, *Sir*, and *Lady*.

One shortcoming of OntoNotes for our application is that it does not contain singleton annotations. However, singletons are important, since they still require gender adaptation to avoid biasing the input. We solve this by treating every PERSON NE that is not assigned to a chain in the first step as a singleton.

We only consider documents for generation where we find at least one entity with either a first name, gendered personal pronoun, or title mention. During the generation of input documents, each entity is assigned a gender and a name. We then fill each mention according to its category by inserting the corresponding pronoun, title, or name.

6.8 Metric Implementation Details

6.8.1 Entity Alignment

Entity inclusion and hallucination bias require rudimentary cross-document coreference resolution between each summary s and input d . OntoNotes gives us access to gold entities and coreference chains E_d in the input d , but we lack the same in the summary s . As a first step, we thus identify all named entities E_s in the summary with an NER tool.¹⁰ This leaves us with the problem of aligning input entities E_d and summary entities E_s . While cross-document coreference is difficult in the general case (Singh et al., 2011), we rely on two assumptions to build a transparent heuristic instead:

1. Most entities are going to be referred to similarly in the input and the summary.
2. If an entity is hallucinated, there is a high chance that it does not have any overlap with the input.

Following these assumptions, we align a summary entity e_s to an input entity e_d if e_s contains the last name of e_d , as assigned during dataset construction. We

¹⁰We use `spacy.io` (Montani et al., 2023).

Male	Female
he	she
him	her
his	hers
himself	herself

TABLE 6.4: Pronouns used for gender classification in Wikipedia articles.

additionally require that any other token in e_s is the first name assigned to e_d during dataset construction or a title. Note that to avoid incorrectly identifying hallucinations, we additionally require that at least one of the tokens in the entity does not appear in the source to count as hallucinated. Manual verification in Section 6.11.1 finds this procedure performs well.

6.8.2 Identifying Gender of Hallucinated Entities

We also require a way of determining the gender of hallucinated summary entities so we can compute hallucination bias. While we can identify the gender of entities that appear in the input from dataset construction, this is not true for hallucinated entities. We thus need to design a classification scheme. We rely on two separate lookup-based approaches:

1. Matching against English Wikipedia pages
2. Matching against the 1990 census first names

For the Wikipedia-based approach, we try to find an English Wikipedia page with a title that exactly matches the named entity detected in the summary (including redirects). To limit false hits, we only consider pages that are in a category that contains the words “births”, “deaths”, or “people”. The latter allows matching categories such as “people from X”, while the first two allow matching categories like “Y deaths”, where Y is a date. We ignore pages with only a single word in the title due to the high likelihood of misidentification.

To determine entity gender from the Wikipedia article text, we use the number of occurrences of the pronouns shown in Table 6.4 and select the gender whose

pronouns appear more frequently. If we have a tie in the number of pronouns or if we get conflicting gender predictions due to multiple people with different genders (according to pronoun count) sharing the same name, we classify the gender as unknown. There is a risk that the better coverage of male entities in Wikipedia (Wagner et al., 2015) might influence our bias metric. We thus manually inspect the failure cases of this step and find no evidence that the failure rate is higher for female names.

If we do not find a matching entity in Wikipedia, we turn to the list of first names from the 1990 US census we also used in the construction of our dataset. The census contains gender frequencies for each included name. We resolve duplicates to the most frequent gender if it is at least twice as frequent as the less frequent one, and eliminate them as ambiguous otherwise. We classify an entity as male if any token is present in the list of male first names, and as female if any token is present in the female list. We do not classify an entity as either gender if it contains names from both lists.

6.9 Summarizers

6.9.1 Models

We study both **purpose-built** summarizers and **chat** models. For **purpose-built summarizers** we use BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020a), both transformer models fine-tuned for summarization. We use the XSum and CNN/DM¹¹ models. For **chat models** we choose Llama-2 chat (Touvron et al., 2023) 7b, 13b, and 70b models with the standard system prompt.

For the chat models, we randomly select one prompt per summary from a list of ten prompts designed to elicit summarizing behavior. We list them in Table 6.5.

Please summarize the following old text
Please summarize the following old article
Summarize the following old text
Summarize the following old article
Give a summary of the following old text
Give a summary of the following old article
Give me a summary of the following old article
Give me a summary of the following old text
I need a summary of the following old article
I need a summary of the following old text

TABLE 6.5: Prompts used for the Llama-2 models. We specify that the articles/documents are “old” since we found in preliminary experiments that this reduces instances where Llama-2 chat 7b would refuse to summarize documents that contained dates or can be implicitly dated.

6.9.2 Gender Bias Summary Statistics

Table 6.6 gives the average number of tokens and entities per summary for the gender bias experiments, as well as the percentage of entities tagged as hallucinated for the summarizers. For comparison, we also report the number of tokens and entities in summaries generated on original documents. All summaries are generated using default model settings in the transformers¹² library. We find that different summarizers produce summaries of varying lengths, with summaries from summarizers trained on XSum being by far the shortest and Llama-2 summaries being the longest. Hallucinations are most frequent for XSum-based summarizers, which, as discussed in Section 2.1, is an expected consequence of the dataset construction.

6.10 Gender Bias Results

Table 6.7 shows that all summarizers score low on both inclusion bias metrics, indicating that the content selection of all studied summarizers does not carry any significant gender bias *in this particular setting*. Remarkably, we find that all

¹¹Taken from <https://huggingface.co>

¹²<https://huggingface.co/docs/transformers/en/index>

Corpus	C_{loc}			C_{glob}		
	Avg. Tok.	Avg. Ent.	% Hal.	Avg. Tok.	Avg. Ent.	% Hal.
BART CNN/DM	60.76 σ : 8.83	0.97 σ : 1.34	4.65	60.88 σ : 8.78	0.99 σ : 1.39	4.01
BART XSum	23.55 σ : 6.71	0.27 σ : 0.59	51.28	23.59 σ : 6.72	0.28 σ : 0.58	47.67
Pegasus CNN/DM	56.23 σ : 16.74	0.87 σ : 1.25	3.29	56.19 σ : 16.90	0.86 σ : 1.22	3.32
Pegasus XSum	24.69 σ : 15.99	0.22 σ : 0.54	33.69	24.74 σ : 16.24	0.22 σ : 0.57	32.09
LLama2 7b	164.40 σ : 42.73	0.97 σ : 1.66	2.97	165.38 σ : 42.55	0.99 σ : 1.69	3.18
LLama2 13b	163.80 σ : 39.04	1.55 σ : 2.10	2.95	163.87 σ : 39.14	1.56 σ : 2.10	2.89
LLama2 70b	147.87 σ : 41.88	1.79 σ : 2.25	1.24	148.11 σ : 41.82	1.79 σ : 2.26	1.39

(A) Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on gender bias data. σ indicates standard deviation.

Corpus	Avg. Ent.	Avg. Tok
BART CNN/DM	60.60 σ : 9.55	1.00 σ : 1.31
BART XSum	22.81 σ : 6.48	0.25 σ : 0.52
Pegasus CNN/DM	55.29 σ : 17.51	0.79 σ : 1.18
Pegasus XSum	22.90 σ : 10.44	0.19 σ : 0.47
LLama2 7b	175.52 σ : 34.63	1.22 σ : 1.96
LLama2 13b	166.86 σ : 39.43	1.63 σ : 2.17
LLama2 70b	151.15 σ : 41.63	1.89 σ : 2.34

(B) Average number of tokens and entities of summaries generated on original documents. σ indicates standard deviation.

TABLE 6.6: Summary statistics for summaries generated on C_{loc} and C_{glob} for gender bias, and on original documents.

	BART		Pegasus		Llama-2 chat		
	CNN	XSum	CNN	XSum	7b	13b	70b
Word List Inclusion	0.00 s: 0.00,0.01 d: 0.00,0.03	0.03 s: 0.02,0.05 d: 0.00,0.11	0.02 s: 0.02,0.03 d: 0.00,0.06	0.04 s: 0.01,0.06 d: 0.00,0.11	0.04 s: 0.02,0.05 d: 0.01,0.07	0.07 s: 0.06,0.07 d: 0.04,0.09	0.04 s: 0.04,0.05 d: 0.02,0.06
Entity Inclusion	0.02 s: 0.01,0.04 d: 0.00,0.04	0.02 s: 0.00,0.06 d: 0.00,0.11	0.03 s: 0.01,0.04 d: 0.01,0.05	0.01 s: 0.00,0.05 d: 0.00,0.08	0.00 s: 0.00,0.03 d: 0.00,0.03	0.04 s: 0.03,0.06 d: 0.02,0.06	0.02 s: 0.00,0.03 d: 0.00,0.03
Entity Hallucination	0.39 s: 0.36,0.42 d: 0.28,0.47	0.37 s: 0.37,0.38 d: 0.31,0.43	0.38 s: 0.35,0.40 d: 0.14,0.50	0.31 s: 0.30,0.33 d: 0.22,0.39	0.38 s: 0.34,0.41 d: 0.30,0.45	0.44 s: 0.42,0.46 d: 0.40,0.48	0.41 s: 0.39,0.43 d: 0.29,0.48
Distinguishability (Cnt.)	0.21 d: 0.19,0.24	0.24 d: 0.20,0.26	0.15 d: 0.13,0.18	0.13 d: 0.11,0.16	0.05 d: 0.03,0.07	0.09 d: 0.06,0.11	0.07 d: 0.04,0.09
Distinguishability (Den.)	0.22 d: 0.19,0.24	0.24 d: 0.21,0.27	0.15 d: 0.13,0.18	0.14 d: 0.12,0.17	0.04 d: 0.02,0.06	0.06 d: 0.04,0.09	0.05 d: 0.03,0.07

TABLE 6.7: Results of our bias metrics. In all cases, a zero score indicates no evidence of bias. We indicate the 95% bootstrap confidence intervals when resampling original documents (d) and when resampling among the different entity assignments sampled during dataset construction (s). We do not compute (s) for distinguishability, since we cannot independently resample scores for input documents generated from the same original document here.

summarizers carry a bias towards male entities in their hallucinations. We study this in more detail in Section 6.12.1.

All summarizers show some degree of distinguishability, with BART summaries showing the most pronounced differences between summaries for male- and female-coded documents. As noted in Section 6.4, this is not in itself sufficient to establish whether this leads to harm to any particular group. We thus analyse this further in Section 6.12.3.

6.11 Validating our Metrics

While we have carefully derived and implemented our bias measurements, we have no “gold” standard to test whether our measurements are valid. This is especially true in cases where we have not found any bias in Section 6.10, i.e. for inclusion bias. While we can further investigate results that indicate bias to find the underlying causes, validating negative results is more challenging. This motivates us to conduct a number of tests to rule out potential problem sources and increase confidence in the reliability of our metrics:

1. We manually verify our alignment algorithm works as intended.

2. We check whether our modified input documents lead to degraded summary quality, which might indicate that our modified input documents are insufficiently natural.
3. We test whether making content words conform to changed entity gender impacts results.
4. We test whether our method is capable of detecting inclusion bias in clearly biased summarizers.

We note that even with this additional validation, it is impossible to *guarantee* the validity of our metrics. However, we believe that in conjunction with our careful avoidance of potential confounding factors when constructing our metrics, these validations make our results sufficiently trustworthy.

6.11.1 Validation of our Alignment Algorithm

To validate that the algorithm aligning input and summary entities outlined in Section 6.8.1 works as intended, we conduct a manual annotation study on the gender bias data. We annotate ten samples each for all systems on both C_{loc} and C_{glob} . This results in a total of 140 input-summary pairs. Since we are interested in validating the alignment, as opposed to the named entity recognizer, we only sample from among all instances where the summary has at least one named entity.

We then manually check the automatic alignment. For each instance, we annotate the following:

1. The number of entities in the source that are incorrectly aligned with an entity in the summary.
2. The number of entities in the summary that are erroneously tagged as hallucinated when they are supported by the input. Since hallucinated entities only affect the hallucination bias score when our gender name classification algorithm assigns an apparent gender to the entity, we report how many of these incorrectly tagged entities receive a gender classification and thus

# Input entities	688
# Summary entities	315
# Input entities with alignment in summary	208
# Incorrect entity alignments	2
# Summary entities tagged as hallucinated	40
...of these with gender classification	18
# Erroneously tagged hallucinations	13
...of these with gender classification	2

TABLE 6.8: Results of our manual annotation of entity alignments. Note that, since we do not have coreference information in the summary, a single input entity can be aligned with multiple summary entities. This may happen in case the name is repeated more than once.

might affect the hallucination score. We conduct this annotation on hallucinations before our additional safeguard requiring at least one token in the entity to not be present in the source.

Results in Table 6.8 show that our alignment procedure generally works very well. The low number of incorrect alignments can be attributed to the strict matching criteria between summary and source entities as described in Section 6.8.1. While a third of hallucinations are incorrect, we find that this has little impact on bias scores, since all except two of these hallucinations do not receive a gender classification and thus do not affect the hallucination bias score.

A qualitative analysis reveals that these incorrectly tagged hallucinations are often caused by more complicated coreference settings. For example, five of the incorrectly identified hallucinations are a result of a document discussing a family “*The Beebes*”, which does not get correctly identified as an entity in the input by our approach, since we focus on mentions of individuals. We also find a failure case where the replacement in the input is incomplete since names are part of nested entities that are not of PERSON type. For example, “*Bush*” in “*The Bush administration*” does not receive a PERSON tag and thus the entity “Bush” cannot be aligned to the input. Since, in our case, these entities are a) not gendered and b) appear in the source document and are thus not taken into account for hallucination bias, this shortcoming of the alignment heuristic does also not affect bias scores.

6.11.2 Summary Quality

Degradation in summary quality between original documents and inputs might be indicative of our inputs being insufficiently natural. This would cast doubt on the generalizability of our observations. We thus conduct an automatic quality evaluation to ensure this is not the case. We evaluate relevance since we are primarily interested in content selection effects that might bias our measurements.

Since we do not have access to gold summaries, we use an unsupervised evaluation method. Following the recent success of using large language models in reference-free evaluation for text generation (Liu et al., 2023a; Chiang and Lee, 2023; Shen et al., 2023), we use GPT 3.5 to elicit ratings for the generated summaries. We prompt the model using the reason-then-score prompt of Shen et al. (2023):¹³

“Score the following Summary given the corresponding Article with respect to relevance from one to five, where one indicates “irrelevance”, and five indicates “perfect relevance”. Note that relevance metrics the Summary’s selection of important content from the Article, whether the Summary grasps the main message of the Article without being overwhelmed by unnecessary or less significant details.

Article: {article}

Summary: {summary}

Provide your reason in one sentence, then give a final score:”

For each system, we evaluate all 683 summaries generated from the original documents which are used as templates for C_{loc} and C_{glob} . For C_{loc} and C_{glob} themselves, we only evaluate summaries generated for two randomly selected inputs per original document to conserve resources. For C_{loc} , we ensure these two inputs form a pair with inverted gender assignment. For C_{glob} , we select one male- and one female-only input. This results in 1366 ratings per system.

¹³We use the `gpt-3.5-turbo-1106` model. This model is more recent than the one used in the evaluation of Shen et al. but allows us to fit the entirety of the documents and summaries into the available tokens.

System	C_{loc}	σ	C_{glob}	σ	Original	σ
Pegasus XSum	4.23	1.45	4.24	1.45	4.28	1.42
Pegasus CNN/DM	4.57	1.03	4.59	1.01	4.70	0.89
BART XSum	4.32	1.38	4.34	1.37	4.30	1.40
BART CNN/DM	4.81	0.66	4.84	0.60	4.86	0.60
Llama-2 7B	3.50	1.83	3.50	1.82	3.85	1.68
Llama-2 13B	4.99	0.18	4.98	0.22	4.99	0.15
Llama-2 70B	5.00	0.05	4.99	0.08	4.99	0.14

TABLE 6.9: GPT-3.5 RTS scores for summaries generated on C_{loc} , C_{glob} and on original documents. For C_{loc} , C_{glob} we evaluate summaries for two inputs each from each original document ($n = 1366$). For the original documents, we evaluate all summaries ($n = 683$). We find only minor differences in quality between summaries on $C_{\text{loc}}/C_{\text{glob}}$ and original documents, indicating that our procedure does not result in systematic degradation of summary quality. σ indicates standard deviation.

Table 6.9 shows that, while there is a small reduction in score for 4 out of 7 systems, performance is very similar between original and modified documents, with the latter score falling within less than one standard deviation of the original score. This indicates that our modification of the input documents does not lead to meaningful degradation in summary quality.

6.11.3 Content Words

Our automatic template generation procedure only changes names and pronominal mentions, leaving content words unchanged. This can lead to unnatural occurrences, such as *Chairman Diane Sasser*, when *Chairwoman Diane Sasser* would be more appropriate. To check whether this is an issue in our experiments, we manually extend the automatically derived templates to also modify content words. We manually annotate 100 documents with how content words should be altered depending on entity gender and rerun our experiments on this subset.

Female		Male	
daughter	niece	son	nephew
mother	nieces	father	nephews
woman	wife	man	husband
girl	wives	boy	husbands
female	actress	male	actor
sister	actresses	brother	actors
daughters	chairwoman	sons	chairman
mothers	chairwomen	fathers	chairmen
women	mum	men	dad
girls	mums	boys	dads
females	waitress	males	waiter
sisters	waitresses	brothers	waiters
aunt	mistress	uncle	lover
aunts		uncles	

TABLE 6.10: Extended word list used to identify candidate documents for annotation.

Annotation Procedure

Since we found in preliminary experiments that many documents do not require any manual intervention, we first run an automatic filter over our dataset to identify candidate documents for annotation. We use an extended variant of the word lists of Liang et al. (2023) reproduced in Table 6.10. We then randomly sample from these documents until we find 100 instances where at least one text span requires manual intervention to adapt to entity gender.

During annotation, we first identify text spans which should change in accordance with the gender of an entity in the document. We then annotate which words should be used depending on the gender of the entities in the document (e.g. generating *chairman* or *chairwoman* depending on the gender of the entity occupying that position). We also consider the case where multiple entities might influence the realization of a particular word, like *brothers*. In these cases, we also specify a neutral variant (e.g. *siblings*) to be used in case the referenced entities have different genders. All annotations were conducted by the author of this thesis.

	BART		Pegasus		Llama-2 chat		
	CNN	XSum	CNN	XSum	7b	13b	70b
Word List Inclusion	0.00 s: 0.00,0.02 d: 0.00,0.05	0.01 s: 0.00,0.05 d: 0.00,0.15	0.02 s: 0.01,0.04 d: 0.00,0.08	0.07 s: 0.03,0.11 d: 0.00,0.21	0.08 s: 0.05,0.12 d: 0.04,0.13	0.05 s: 0.03,0.06 d: 0.01,0.09	0.06 s: 0.04,0.07 d: 0.02,0.09
Word List Inclusion (Orig.)	0.05 s: 0.03,0.06 d: 0.00,0.11	0.05 s: 0.01,0.09 d: 0.00,0.18	0.06 s: 0.04,0.07 d: 0.00,0.15	0.10 s: 0.07,0.13 d: 0.01,0.25	0.06 s: 0.02,0.09 d: 0.01,0.11	0.06 s: 0.05,0.08 d: 0.02,0.11	0.04 s: 0.03,0.05 d: 0.00,0.09
Entity Inclusion	0.03 s: 0.00,0.07 d: 0.00,0.07	0.05 s: 0.00,0.16 d: 0.00,0.23	0.04 s: 0.01,0.08 d: 0.01,0.09	0.01 s: 0.00,0.10 d: 0.00,0.25	0.01 s: 0.00,0.08 d: 0.00,0.09	0.06 s: 0.02,0.09 d: 0.01,0.10	0.03 s: 0.00,0.06 d: 0.00,0.07
Entity Inclusion (Orig.)	0.03 s: 0.00,0.07 d: 0.00,0.07	0.05 s: 0.00,0.15 d: 0.00,0.22	0.04 s: 0.01,0.08 d: 0.00,0.09	0.01 s: 0.00,0.10 d: 0.00,0.24	0.01 s: 0.00,0.08 d: 0.00,0.08	0.06 s: 0.02,0.09 d: 0.01,0.10	0.03 s: 0.00,0.06 d: 0.00,0.07
Entity Hallucination	0.30 s: 0.19,0.40 d: 0.02,0.50	0.32 s: 0.29,0.34 d: 0.10,0.46	0.45 s: 0.36,0.50 d: 0.00,0.50	0.36 s: 0.31,0.39 d: 0.08,0.50	0.35 s: 0.22,0.46 d: 0.22,0.46	0.46 s: 0.40,0.50 d: 0.31,0.50	0.39 s: 0.30,0.47 d: 0.17,0.50
Entity Hallucination (Orig.)	0.26 s: 0.15,0.36 d: 0.03,0.48	0.33 s: 0.31,0.35 d: 0.12,0.47	0.50 s: 0.50,0.50 d: 0.00,0.50	0.33 s: 0.30,0.36 d: 0.05,0.50	0.17 s: 0.02,0.34 d: 0.01,0.42	0.41 s: 0.34,0.47 d: 0.22,0.50	0.40 s: 0.34,0.47 d: 0.06,0.50
Distinguishability (Count)	0.42 s: 0.36,0.49 d: 0.20,0.34	0.42 s: 0.34,0.50 d: 0.25,0.40	0.27 s: 0.19,0.35 d: 0.13,0.24	0.23 s: 0.16,0.32 d: 0.09,0.27	0.07 s: 0.01,0.12 d: 0.08,0.17	0.20 s: 0.11,0.28 d: 0.08,0.22	0.26 s: 0.19,0.33 d: 0.05,0.15
Distinguishability (Count) (Orig.)	0.27 s: 0.20,0.34 d: 0.34,0.49	0.32 s: 0.25,0.40 d: 0.31,0.48	0.19 s: 0.13,0.24 d: 0.20,0.36	0.17 s: 0.18,0.32 d: 0.09,0.26	0.12 s: 0.00,0.10 d: 0.07,0.16	0.15 s: 0.12,0.25 d: 0.09,0.26	0.10 s: 0.05,0.15 d: 0.02,0.14
Distinguishability (Dense)	0.41 s: 0.34,0.49 d: 0.20,0.35	0.40 s: 0.31,0.48 d: 0.24,0.40	0.28 s: 0.10,0.24 d: 0.09,0.26	0.25 s: 0.18,0.32 d: 0.09,0.26	0.05 s: 0.00,0.10 d: 0.07,0.16	0.19 s: 0.12,0.25 d: 0.09,0.26	0.28 s: 0.21,0.34 d: 0.02,0.14
Distinguishability (Dense) (Orig.)	0.28 s: 0.20,0.35 d: 0.34,0.49	0.32 s: 0.24,0.40 d: 0.31,0.48	0.17 s: 0.10,0.24 d: 0.09,0.26	0.18 s: 0.09,0.26 d: 0.02,0.16	0.07 s: 0.02,0.11 d: 0.02,0.16	0.09 s: 0.02,0.16 d: 0.02,0.16	0.08 s: 0.02,0.16 d: 0.02,0.16

TABLE 6.11: Results on our manually extended variants of C_{loc} and C_{glob} for gender bias with content words altered to conform to entity gender. Since our annotations cover only a relatively small subset of the whole corpus, we also report the scores of summaries generated for the same inputs without content word modification for comparison (Or.). We find that almost all scores fall within their respective confidence intervals.

Metric	Llama-2 chat 13b
Word List	0.42 s: 0.41,0.42 d: 0.40,0.44
Entity Inclusion	0.71 s: 0.68,0.74 d: 0.63,0.80

TABLE 6.12: Inclusion bias scores on Llama-2 13b prompted to induce an inclusion bias towards female entities.

Results

We report results on gender bias with our modified inputs in Table 6.11. We find scores for modified inputs are very close to original scores when taking into account confidence intervals and exhibit the same trends. This suggests that our choice not to alter content words in the input does not meaningfully influence results. However, we note that the small number of inputs makes confidence intervals relatively wide.

6.11.4 Induced Bias Detection

Since we find no inclusion bias in the studied summarization systems, we test whether our method is capable of detecting inclusion bias in clearly biased summarizers. To simulate a biased summarizer, we rerun Llama-2 13b on C_{loc} but append “*Please put a particular focus on the women mentioned in the text*” to the prompt. We manually verify that Llama-2 does not refuse this instruction.

Inclusion bias scores shown in Table 6.12 show that we can clearly detect the induced inclusion bias.

6.12 Gender Bias Analysis

6.12.1 Investigating Hallucination Bias

We now turn to a closer investigation of the biases we detect, starting with hallucination bias. Our results in Section 6.10 show that there is a consistent bias towards male entities in the hallucinations of all summarizers we investigate. To

	CNN/DM	#	XSum	#		CNN/DM	#	XSum	#
BART	greene _u	91	farai sevenzo _m	352	Pegasus	frum _u	76	boris yeltsin _m	60
	bob greene _m	69	george w. bush _m	315		david frum _m	75	obama _u	48
	david frum _m	53	mikhail gorbachev _m	104		zelizer _u	40	farai sevenzo _m	44
	frum _u	47	james baker _m	66		greene _u	28	francois mitterrand _m	40
	peter bergen _m	41	boris yeltsin _m	60		bob greene _m	25	richard cohen _m	32
	bergen _u	41	daniel ortega _m	56		julian zelizer _m	20	sharmila tagore _f	31
	saatchesi _u	25	obama _u	49		frida ghitis _f	19	helmut kohl _m	30
	bynoes _u	20	helmut kohl _m	40		ghitis _u	19	alain juppe _m	30
	frida ghitis _f	15	francois mitterrand _m	40		david weinberger _m	8	george w. bush _m	25
	hainis _u	12	george h. w. bush _m	25		bergen _u	8	k. _u	20
	# male	238	# male	1465		# male	170	# male	662
	# female	29	# female	212		# female	24	# female	153
Llama-2 chat	7b	#	13b	#	70b	7b	#		
	mikhail gorbachev _m	36	erich honecker _m	74		mikhail gorbachev _m	27		
	richard nixon _m	29	mikhail gorbachev _m	53		erich honecker _m	22		
	boris yeltsin _m	23	richard nixon _m	32		richard nixon _m	21		
	erich honecker _m	20	manuel noriega _m	32		walter sisulu _m	20		
	mclaren _u	20	george h.w. bush _m	29		alan greenspan _m	20		
	daniel ortega _m	17	daniel ortega _m	29		naguib mahfouz _m	16		
	james baker _m	14	walter sisulu _m	20		yasser arafat _m	12		
	helmut kohl _m	14	mahatma gandhi _m	18		edberg _u	12		
	eduard shevardnadze _m	12	nelson mandela _m	17		nelson mandela _m	11		
	pat nixon _f	12	james baker _m	17		george h.w. bush _m	9		
	# male	290	# male	545		# male	259		
	# female	32	# female	35		# female	26		

TABLE 6.13: Ten most frequent PERSON named entities without alignment in the generated summaries. *m/f/u* indicate entities tagged as *male/female/unknown* by our name gender classifier (see Section 6.8.2).

better understand the nature and causes of this bias, we investigate the ten most frequent hallucinations of each model in Table 6.13.

We identify two types of frequent hallucinations: For the first type, summarizers often insert entities that are related to the time of the original documents, sometimes by “hallucinating” the original name for an entity in spite of the input, or by inserting the first name for entities that are mentioned without first name in the input. The male bias here can thus be attributed to the male-dominant nature of news at article publication times. A possible cause for this is that we do not alter last names in the inputs. We will investigate – and reject – this hypothesis in Section 6.12.2. Our observations also link with recent research on *knowledge conflicts* (Wang et al., 2023; Xie et al., 2024), where language models may fail to properly reflect answer uncertainty introduced by conflicting evidence in prompt and parametric knowledge. For the Llama-2 models, we manually verify that most hallucinations can be explained in this way.

However, for the purpose-built summarizers, we find a second type of hallucinations that refer to contributors from CNN (for CNN/DM trained summarizers) or the BBC (for XSum). These usually appear when the summary attributes the text to an author. This is more problematic than the hallucination of historic entities since the hallucinated entities always incorrectly attribute authorship to already potentially well-known, mostly male figures. We find many of these follow repeated patterns. For example, in many instances, BART and Pegasus XSum would generate “In our series of letters from African - American journalists, writer and columnist [name] ...”, followed by a short summary.

6.12.2 Investigating the Effect of Replacing Last Names

To test whether our choice of minimizing modifications to the original documents by leaving last names intact during corpus construction influences our results for hallucination bias, we repeat our original experiments but additionally replace last names. We use the last names from the 2010 US census,¹⁴ again following Parrish et al. (2022). Since any effect of this experiment is likely to be limited to hallucination bias, we only conduct this experiment on C_{loc} to preserve computational resources. We compute entity hallucination scores, along with the two inclusion scores, since they are computed on C_{loc} as well. Results shown in Table 6.14 are comparable with the setting that leaves last name intact, with the exception of Llama-2 chat 13b, which shows a notable decrease in hallucination score. However, even in the latter case, it remains significantly non-zero. This suggests hallucination scores are not a side effect of our corpus construction.

6.12.3 Investigating Distinguishability

Finally, we investigate the causes of the observed distinguishability scores. Table 6.7 indicate some systematic difference between summaries generated for male- and female-coded inputs, even when accounting for expected grammatical differences (see Section 6.5.4). A possible explanation for this is a difference in summary

¹⁴https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

	BART		Pegasus		Llama-2 chat		
	CNN	XSum	CNN	XSum	7b	13b	70b
Word List Inclusion	0.01 s: 0.00,0.02 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.08	0.03 s: 0.03,0.04 d: 0.00,0.07	0.03 s: 0.01,0.05 d: 0.00,0.09	0.06 s: 0.04,0.08 d: 0.03,0.09	0.07 s: 0.06,0.08 d: 0.05,0.09	0.06 s: 0.05,0.07 d: 0.04,0.08
Entity Inclusion	0.01 s: 0.00,0.02 d: 0.00,0.03	0.05 s: 0.01,0.09 d: 0.00,0.12	0.01 s: 0.00,0.03 d: 0.00,0.03	0.04 s: 0.00,0.08 d: 0.00,0.10	0.03 s: 0.00,0.05 d: 0.00,0.06	0.02 s: 0.00,0.03 d: 0.00,0.04	0.02 s: 0.01,0.04 d: 0.00,0.04
Entity Hallucination	0.44 s: 0.41,0.47 d: 0.38,0.49	0.29 s: 0.27,0.31 d: 0.24,0.34	0.41 s: 0.39,0.44 d: 0.24,0.50	0.27 s: 0.25,0.29 d: 0.21,0.33	0.37 s: 0.32,0.43 d: 0.28,0.43	0.32 s: 0.26,0.38 d: 0.18,0.42	0.43 s: 0.38,0.47 d: 0.33,0.48

TABLE 6.14: Results for entity metrics computed on C_{loc} for gender bias with last names altered. We do not report distinguishability, since it requires a corpus in C_{glob} format. We find results are comparable with results without last name alteration. Only Llama-2 13b shows a notable decrease in hallucination score, although it still exhibits strong hallucination bias.

System	M.	F.	Diff
BART XSum	4.30	4.37	0.07 d: 0.01,0.14
BART CNN/DM	4.84	4.84	0.01 d: 0.00,0.05
Pegasus XSum	4.24	4.24	0.00 d: 0.00,0.09
Pegasus CNN/DM	4.59	4.59	0.00 d: 0.00,0.07
Llama-2 7B	3.50	3.50	0.00 d: 0.00,0.20
Llama-2 13B	4.98	4.99	0.01 d: 0.00,0.03
Llama-2 70B	5.00	4.99	0.01 d: 0.00,0.02

TABLE 6.15: GPT 3.5 RTS relevance on C_{glob} for summaries on male- and female-only inputs, along with score difference. We compute confidence intervals for the score difference as in Table 6.7.

quality between genders. We test this using the same reference-free automatic evaluation metric as in Section 6.11.2. We report average scores comparing male and female summaries in C_{glob} in Table 6.15, finding no quality differences.

Automatic evaluation can itself be biased and differences in summary quality are only one aspect of representation bias. We thus conduct a manual *qualitative* analysis. We first group all inputs that were generated for the same original document in C_{glob} . We then sort these groups by their distinguishability score and investigate the instances with the highest distinguishability.

For BART XSum, which has the highest overall distinguishability, we find there is a pattern where summaries highlight the gender of women in the context of receiving an appointment to a position of power, but do not do the same for men.

Corpus	C_{loc}			C_{glob}		
	Avg. Tok.	Avg. Ent.	% Hal.	Avg. Tok.	Avg. Ent.	% Hal.
BART CNN/DM	60.99 σ : 8.59	0.92 σ : 1.32	4.71	60.97 σ : 8.63	0.88 σ : 1.30	4.15
BART XSum	23.50 σ : 6.50	0.25 σ : 0.52	46.05	23.40 σ : 6.51	0.24 σ : 0.52	46.09
Pegasus CNN/DM	56.62 σ : 16.76	0.83 σ : 1.21	4.41	56.62 σ : 16.94	0.80 σ : 1.20	4.07
Pegasus XSum	24.66 σ : 13.78	0.21 σ : 0.50	38.31	24.66 σ : 12.67	0.19 σ : 0.49	41.31
LLama2 7b	172.59 σ : 37.64	0.88 σ : 1.55	2.12	172.48 σ : 37.70	0.89 σ : 1.58	1.86
LLama2 13b	162.54 σ : 39.04	1.45 σ : 1.98	1.28	162.25 σ : 39.37	1.34 σ : 1.86	1.50
LLama2 70b	148.14 σ : 41.63	1.71 σ : 2.18	0.46	147.70 σ : 41.98	1.61 σ : 2.05	0.44

TABLE 6.16: Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on **race** bias data with randomly assigned genders. σ indicates standard deviation. Note that while we could theoretically identify hallucinated instances for race bias using the same algorithm we use for the gender bias experiments, we cannot use these to compute hallucination bias since we do not attempt to identify entity race from names.

See Figure 6.1 for an example. We find a total of 12 instances of “first woman” and an additional 11 instances of “first female” in the summaries generated by BART XSum, but no instances of “first male” and only a single instance of “first man”. This not only hallucinates information but also forms an instance of *markedness* (Waugh, 1982; Cheng et al., 2023) by highlighting the appointment of women to positions of power as abnormal. We find no similarly problematic patterns for the remaining systems. Instead, most changes are minor variations in content selection between male and female inputs.

6.13 Extension to Race Bias

6.13.1 A Dataset for Race Bias in Summarization

While we have thus far focused only on gender bias as a well-known bias category, our methods are applicable to any group-based bias where group membership can be indicated using names. We demonstrate this by investigating race bias for

stereotypically black and white names. We use the name dictionary of Parrish et al. (2022). We change first and last names since both are relevant in communicating race. We investigate a total of five different settings with regard to the intersection of race and gender: one *random* setting, where gender and race are assigned independently, and four intersectional ones, where gender and race are assigned in tandem. Since the name inventory is smaller than for gender, we cannot generate instances for all documents. We thus only consider originals where we can generate a full set of 20 inputs under all settings, leaving us with 12,240 instances per dataset.

We give statistics for the summaries generated for the race bias experiments with random gender assignment in Table 6.16. The remaining, highly similar, tables can be found in Appendix G for completeness. We find that behavior is similar to that on the gender bias dataset for all summarizers (compare Table 6.6).

6.13.2 Results

Since word lists for race bias typically rely on last names, we only compute entity inclusion bias. We also opt not to compute hallucination bias, since we want to avoid constructing a classifier that attempts to identify entity race. Table 6.17 shows that most summarizers exhibit no entity inclusion bias, with the exception of BART XSum, which prefers to include black-associated names in the summary. We find that behavior is very similar between the random and intersectional settings. Interestingly, we find that for all summarizers that have significantly non-zero distinguishability, it is highest when black- and white-coded entities are assigned opposite genders. Similarly, for BART XSum, inclusion bias is highest in these settings, although we note that none of the differences are significant. Overall we find no strong evidence of intersectional effects in our bias metrics.

6.13.3 Investigating Distinguishability

Analogously to our analysis for gender bias in Section 6.12.3, we check quality differences as a source of distinguishability in Table 6.18. We find that scores are

Gender Assignment	BART		Pegasus		Llama-2 chat		
	CNN	XSum	CNN	XSum	7b	13b	70b
Entity Inclusion Bias							
Random	0.01 s: 0.00,0.03 d: 0.00,0.04	0.17 s: 0.11,0.24 d: 0.08,0.29	0.04 s: 0.00,0.09 d: 0.00,0.10	0.02 s: 0.00,0.04 d: 0.00,0.05	0.01 s: 0.00,0.04 d: 0.00,0.05	0.01 s: 0.00,0.02 d: 0.00,0.03	0.03 s: 0.01,0.05 d: 0.01,0.05
Black Male/White Female	0.03 s: 0.01,0.04 d: 0.00,0.05	0.19 s: 0.13,0.26 d: 0.10,0.31	0.04 s: 0.00,0.09 d: 0.00,0.12	0.02 s: 0.00,0.03 d: 0.00,0.04	0.05 s: 0.02,0.08 d: 0.01,0.08	0.02 s: 0.00,0.04 d: 0.00,0.04	0.04 s: 0.02,0.06 d: 0.02,0.06
Black Male/White Male	0.05 s: 0.03,0.07 d: 0.02,0.08	0.11 s: 0.05,0.18 d: 0.03,0.21	0.08 s: 0.03,0.13 d: 0.01,0.16	0.05 s: 0.03,0.07 d: 0.03,0.08	0.02 s: 0.00,0.05 d: 0.00,0.05	0.01 s: 0.00,0.03 d: 0.00,0.03	0.02 s: 0.00,0.04 d: 0.01,0.04
Black Female/White Male	0.03 s: 0.01,0.05 d: 0.00,0.06	0.24 s: 0.18,0.30 d: 0.13,0.38	0.05 s: 0.00,0.10 d: 0.00,0.14	0.01 s: 0.00,0.03 d: 0.00,0.04	0.01 s: 0.00,0.04 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.03
Black Female/White Female	0.01 s: 0.00,0.03 d: 0.00,0.04	0.12 s: 0.06,0.17 d: 0.02,0.23	0.02 s: 0.00,0.07 d: 0.00,0.10	0.04 s: 0.02,0.06 d: 0.01,0.07	0.01 s: 0.00,0.04 d: 0.00,0.04	0.01 s: 0.00,0.03 d: 0.00,0.03	0.02 s: 0.00,0.04 d: 0.00,0.04
Distinguishability (Cnt.)							
Random	0.19 d: 0.16,0.21	0.23 d: 0.20,0.25	0.16 d: 0.14,0.19	0.10 d: 0.08,0.12	0.01 d: -0.01,0.03	0.04 d: 0.02,0.06	0.01 d: -0.01,0.03
Black Male/White Female	0.24 d: 0.22,0.26	0.28 d: 0.25,0.31	0.18 d: 0.16,0.21	0.13 d: 0.11,0.16	0.03 d: 0.01,0.05	0.04 d: 0.02,0.06	0.07 d: 0.05,0.09
Black Male/White Male	0.20 d: 0.17,0.22	0.25 d: 0.22,0.27	0.15 d: 0.13,0.17	0.09 d: 0.06,0.11	0.02 d: -0.00,0.04	0.02 d: 0.00,0.05	0.04 d: 0.02,0.06
Black Female/White Male	0.23 d: 0.21,0.26	0.30 d: 0.27,0.33	0.26 d: 0.24,0.29	0.19 d: 0.16,0.21	0.03 d: 0.01,0.05	0.05 d: 0.03,0.07	0.10 d: 0.08,0.12
Black Female/White Female	0.21 d: 0.18,0.23	0.24 d: 0.22,0.27	0.22 d: 0.19,0.24	0.12 d: 0.09,0.14	0.01 d: -0.01,0.03	0.01 d: -0.01,0.04	0.05 d: 0.03,0.07
Distinguishability (Den.)							
Random	0.16 d: 0.13,0.19	0.21 d: 0.19,0.24	0.17 d: 0.14,0.19	0.10 d: 0.08,0.13	0.02 d: -0.00,0.04	0.03 d: 0.01,0.05	0.02 d: -0.00,0.04
Black Male/White Female	0.24 d: 0.22,0.27	0.28 d: 0.26,0.31	0.19 d: 0.17,0.22	0.13 d: 0.10,0.15	0.01 d: -0.01,0.03	0.03 d: 0.01,0.05	0.05 d: 0.03,0.07
Black Male/White Male	0.19 d: 0.17,0.22	0.23 d: 0.21,0.26	0.16 d: 0.13,0.18	0.09 d: 0.07,0.12	0.02 d: 0.00,0.04	0.02 d: 0.00,0.04	0.06 d: 0.04,0.08
Black Female/White Male	0.24 d: 0.22,0.26	0.29 d: 0.27,0.32	0.26 d: 0.24,0.29	0.18 d: 0.16,0.20	0.04 d: 0.02,0.06	0.05 d: 0.03,0.08	0.09 d: 0.07,0.11
Black Female/White Female	0.21 d: 0.19,0.24	0.23 d: 0.21,0.26	0.22 d: 0.19,0.25	0.12 d: 0.10,0.15	0.02 d: -0.00,0.04	0.02 d: 0.00,0.04	0.04 d: 0.02,0.06

TABLE 6.17: Bias scores for race bias with black/white associated names with different gender assignments. *Random* assigns gender uniformly at random, independently of race.

System	Black	White	Diff
BART XSum	4.22	4.33	0.11 d: 0.02, 0.20
BART CNN/DM	4.85	4.83	0.02 d: 0.00, 0.07
Pegasus XSum	4.26	4.27	0.00 d: 0.00, 0.10
Pegasus CNN/DM	4.65	4.64	0.01 d: 0.00, 0.09
Llama 7B	3.47	3.56	0.09 d: 0.01, 0.27
Llama 13B	4.98	4.98	0.00 d: 0.00, 0.02
Llama 70B	4.98	4.99	0.01 d: 0.00, 0.02

TABLE 6.18: Quality difference scores for race bias with random gender assignment. Confidence intervals are computed using bootstrap resampling of documents.

largely similar, with no summarization system showing significant quality differences. This mirrors our observations on gender bias. We also repeat the same qualitative analysis as outlined in Section 6.12.3 but find no problematic patterns.

6.14 Discussion

In this chapter, we have introduced bias as a crucial quality dimension for summarization evaluation. We have introduced definitions that allow us to clearly formulate expectations for what constitutes bias in summarization, along with metrics that allow us to detect these biases. We have shown that any metric of *summarizer* bias must account for confounding biases in the input and proposed a rule-based method that allows us to create realistic data with controlled entity distribution for studying summarizer bias. This provides both tools for researchers interested in detecting biases in their summarization systems as well as guidelines that help create new metrics for bias in summarization.

Our study of seven summarizers indicates that content selection is not strongly affected by either gender or race bias for black/white-coded names. We find significant gender bias in hallucinations revealing a connection between unfaithfulness and bias. This suggests increasing faithfulness as a bias mitigation strategy.

The results in this chapter might be taken to indicate that evaluating bias in content selection is not necessary. However, we caution that content selection in news summarization is known to be subject to easy heuristics like the lead “bias” (Jung et al., 2019). Summaries might be more susceptible to biases in more

complicated settings. We thus argue that bias evaluation should be an element of any holistic evaluation setup.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have set out to provide a toolbox for holistic text summarization evaluation. Taken together, the elements of this thesis provide:

- A set of best practices for cost-efficient and reliable **human evaluation** that lets researchers avoid common pitfalls in their analysis
- A robust approach to **meta-evaluation** that allows researchers to quickly identify instances of system-level confounders that threaten the generalizability of meta-evaluation results
- A comprehensive study of **summary coherence measures**, where we have identified shortcomings and promising directions for coherence modelling
- A cost-efficient **faithfulness metric** based on NLI
- Abstract definitions and concrete metrics to detect the presence of **bias** in summaries

When combined with traditional evaluation for relevance, our tools allow a user or researcher to get a holistic picture of the performance of summarization systems. We have made a particular effort to ensure this holistic evaluation remains *cost-efficient*. For human evaluation, we have discussed in Chapter 3 how we can reduce the cost of annotation by choosing appropriate annotation methods and using nested, as opposed to the more common partially crossed designs. For automatic

evaluation, cost is dependent on the computational effort required to run metrics, which is why in Chapter 5 we have made an effort to find a faithfulness measure that is based on a relatively small model and does not require costly inference procedures. This ensures evaluations can be run both frequently and thoroughly.

Beyond the pure practical utility of this thesis, a particularly important insight in this thesis is how easily the reliability of summarization evaluation can come under threat. We have, at numerous points, identified reliability shortcomings in current practices and made proposals to alleviate them:

- In our human evaluation in Chapter 3, we have identified the effect of variance in annotator preferences and input document difficulty as potential sources of Type I errors in statistical evaluation when not accounted for.
- Our meta-evaluation of coherence measures in Chapter 4 has shown that rankings of coherence measures are heavily influenced by the presence of system-level confounders, i.e. strong differences in summarizer coherence scores that happen to coincide with summarizer ranking within a coherence measure. This may lead to poor generalization of meta-evaluation results to future summarization systems.
- Our improvements to NLI models for faithfulness in TRUE in Chapter 5 are, in part, due to a better modelling of differences in task formulation in different domains. Not taking these into account may have confounded past evaluations.
- In our investigation of gender and race bias in summarization in Chapter 6, we have highlighted the importance of accounting for the input distribution, so it does not become a confounding factor in evaluation scores.

While the approaches for dealing with confounding factors must be individually tailored to the specific problems at hand, this highlights the importance of carefully designing evaluation studies to account for potential threats to reliability.

In addition to providing a usable, reliable, and cost-efficient toolbox for summarization evaluation, this thesis thus also provides a number of case studies in how to identify and address threats to reliability.

7.2 Outlook

While we believe that this thesis has tackled contemporary issues in evaluation, we expect that the shifting landscape of model capabilities will require future adjustment to evaluation practices. This thesis itself, in a way, already reflects this process. The degeneracies that have led to increased repetition, one of the quality dimensions we investigated in Chapter 3, have been effectively solved by larger models (Lewis et al., 2020; Zhang et al., 2020a) and improved sampling procedures (Holtzman et al., 2020). Similarly, coherence of large models is typically high, even in longer-form summaries (Subbiah et al., 2024). On the other hand, faithfulness and bias have arisen as (potential) new concerns. It thus becomes clear that as long as model development does not slow down, evaluation practices must continue to evolve and adapt as well. In this final part of the thesis, we will reflect on what this evolution might look like.

The most obvious likely change is a shift in evaluation dimensions, exemplified by the additions of faithfulness and bias as dimensions of interest. It is reasonable to assume future evaluations might further shift which quality dimensions are important. However, we believe our core insights regarding human and meta-evaluation will remain relevant for any quality dimension.

The increase in quality of summaries raises the question of whether crowd-sourced evaluation studies, such as the ones we conducted in Chapter 3, will continue to provide useful signals for model evaluation. For example, Clark et al. (2021) find that untrained human annotators are unable to differentiate between human-written and GPT-3-generated stories. Further enlightening the gap of expert and crowd-sourced evaluation in the context of improving text summarization systems would provide useful guidance for future evaluation studies. Our guidelines regarding cost-efficient study design in Chapter 3 might help enable researchers to conduct expert studies, which have a much higher base cost.

More fundamentally, all evaluations we have conducted in this paper have been *intrinsic*. However, as summarization systems become stronger and differences become more subtle, intrinsically evaluating the quality of a summary, especially without reference to a specific downstream task, is bound to become increasingly difficult even for experts. While intrinsic evaluation will remain important to

pinpoint specific shortcomings, such as a lack of faithfulness or coherence, we argue future efforts should focus more on *extrinsic* evaluation, in spite of the associated challenges. A similar sentiment is also mirrored by Goyal et al. (2022). While there is some recent work in this area (Pu et al., 2024), there are a number of underexplored questions:

- *Which downstream tasks should extrinsic evaluation focus on?* As we have discussed in Section 2.1, many recent summarization datasets have been created by crawling naturally occurring summaries from the web. However, with these summaries, it is often unclear which *purpose* they serve. Designing effective extrinsic evaluations will require a more careful consideration of what a summary is expected to accomplish. Ideally, summaries can be integrated into an already existing process. The extrinsic evaluation in Mani et al. (1999) was, for example, inspired by the needs of U.S. information analysts.
- *What are realistic settings for this kind of evaluation?* In the classic summarization setting, the summarizer is provided with the input and (optionally) a query and generates a single summary. However, in a realistic extrinsic evaluation, a summarizer is likely to be part of a larger system. Annotators might, for example, have access to additional resources in addition to the summary, or even work in an interactive summarization setting (Shapira et al., 2022). Designing an evaluation where summarizers are part of a realistic working environment will be crucial to identify their actual utility to the downstream task.
- *What are appropriate measures to quantify the extrinsic utility of a summary?* Both Pu et al. (2024) and Mani et al. (1999) measure downstream task performance and time required for task completion in their extrinsic evaluations. In both cases, evaluation is focused mostly on classification-based tasks, where human performance can be easily graded using classification metrics. However, this is not appropriate for all downstream tasks. As an example, consider an opinion summarization setting, where utility of a summary is defined as how well it informs a shopper’s choice of product. In this case, an appropriate measure might be how the shopper rates

the product chosen based on the summaries. For more critical use cases, assigning severity to mistakes in downstream tasks will also likely play an increasing role. Consider a scholarly multi-document summarization system that summarizes the contributions of a large set of papers. Such a system could, for example, be extrinsically evaluated by having researchers identify relevant related work from the summary. We might imagine a summarizer that generates very brief summaries that cover only the most well-known papers in the input. Such a system would save a lot of time and likely even result in decent downstream performance as measured by the number of papers correctly identified, but would encourage shallow related work sections that systematically ignore less well-known work. Such subtleties are unlikely to be well captured by one-dimensional task performance metrics and will require careful design of performance measurements.

Finally, we have treated the problems of human evaluation and automatic evaluation entirely separately in this thesis. However, with human evaluation being reliable but costly (especially when conducted with experts) and the less reliable automatic evaluation having a cost advantage, combining both might yield an evaluation framework that provides the advantages of both. This combination has naturally been explored with earlier automatic metrics. Chaganty et al. (2018) provide a statistical framework for debiasing an arbitrary metric using human judgements. However, they find little practical advantage with this combination due to the weaknesses of contemporary summarization evaluation metrics. With the advent of stronger automatic evaluation, it is reasonable to expect future improvements from more direct collaboration between human and automatic evaluation. A promising direction in this regard is to use metrics to automatically select instances where human annotation would lead to a large gain in knowledge of overall system ranking. While some forays have been made in this direction (Mohankumar and Khapra, 2022; Ruan et al., 2024), the design space of such active learning approaches for evaluation is far from completely mapped.

In sum, while we believe our work to be holistic with regard to the current state of text summarization, the nature of evaluation means it must constantly co-evolve along constantly improving summarization systems. However, the fundamental

goals of reliability and cost-efficiency will remain relevant for any future evaluation endeavors. Our work provides practical guidance for achieving this, both for human and automatic evaluation.

Bibliography

- Adams, Griffin, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad (2023). “From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting”. In: *Proceedings of the 4th New Frontiers in Summarization Workshop*. Ed. by Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini. Singapore: Association for Computational Linguistics, pp. 68–74. DOI: 10.18653/v1/2023.newsum-1.7. URL: <https://aclanthology.org/2023.newsum-1.7> (cit. on p. 17).
- Afli, Haithem, Pintu Lohar, and Andy Way (2017). “MultiNews: A Web collection of an Aligned Multimodal and Multilingual Corpus”. In: *Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 11–15. URL: <https://www.aclweb.org/anthology/W17-5602> (cit. on p. 255).
- Amidei, Jacopo, Paul Piwek, and Alistair Willis (2018). “Rethinking the Agreement in Human Evaluation Tasks”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 3318–3329. URL: <https://www.aclweb.org/anthology/C18-1281> (cit. on pp. 35, 36, 46).
- Artstein, Ron and Massimo Poesio (2008). “Survey Article: Inter-Coder Agreement for Computational Linguistics”. In: *Computational Linguistics* 34.4, pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: <https://aclanthology.org/J08-4004> (cit. on p. 33).
- Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: San Diego, California (cit. on pp. 16, 48).

- Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (2013). “Abstract Meaning Representation for Sembanking”. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Ed. by Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper. Sofia, Bulgaria: Association for Computational Linguistics, pp. 178–186. URL: <https://aclanthology.org/W13-2322> (cit. on pp. 24, 124).
- Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace (1996). “Magnitude Estimation of Linguistic Acceptability”. In: *Language* 72.1, p. 32. ISSN: 00978507. DOI: 10.2307/416793. URL: <https://www.jstor.org/stable/416793?origin=crossref> (visited on 06/15/2024) (cit. on p. 40).
- Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach (2017). “The Problem with Bias: From Allocative to Representational Harms in Machine Learning”. In: *SIGCIS conference paper* (cit. on pp. 3, 154).
- Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily (2013). “Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal”. In: *Journal of Memory and Language* 68.3, pp. 255–278. ISSN: 0749596X. DOI: 10.1016/j.jml.2012.11.001. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0749596X12001180> (visited on 08/20/2020) (cit. on pp. 31, 46, 55, 58, 60).
- Bartl, Marion, Malvina Nissim, and Albert Gatt (2020). “Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias”. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Ed. by Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 1–16. URL: <https://aclanthology.org/2020.gebnlp-1.1> (cit. on p. 153).
- Barzilay, Regina and Mirella Lapata (2008). “Modeling Local Coherence: An Entity-Based Approach”. In: *Computational Linguistics* 34.1, pp. 1–34. DOI: 10.1162/coli.2008.34.1.1. URL: <https://aclanthology.org/J08-1001> (cit. on pp. 73, 74, 81–84, 86, 92, 94).
- Belz, Anja and Eric Kow (2011). “Discrete vs. Continuous Rating Scales for Language Evaluation in NLP”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

- Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, pp. 230–235. URL: <https://aclanthology.org/P11-2040> (cit. on p. 40).
- Bhandari, Manik, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig (2020). “Re-evaluating Evaluation in Text Summarization”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 9347–9359. DOI: 10.18653/v1/2020.emnlp-main.751. URL: <https://aclanthology.org/2020.emnlp-main.751> (cit. on p. 87).
- Binh Tran, Giang, Mohammad Alrifai, and Dat Quoc Nguyen (2013). “Predicting Relevant News Events for Timeline Summaries”. In: *Proceedings of the 22nd International Conference on World Wide Web. WWW ’13 Companion*. New York, NY, USA: Association for Computing Machinery, pp. 91–92. ISBN: 978-1-4503-2038-2. DOI: 10.1145/2487788.2487829. URL: <https://doi.org/10.1145/2487788.2487829> (cit. on p. 13).
- Blodgett, Su Lin (2021). “Sociolinguistically Driven Approaches for Just Natural Language Processing”. PhD thesis. University of Massachusetts Amherst. DOI: 10.7275/20410631. URL: https://scholarworks.umass.edu/dissertations_2/2092 (visited on 06/13/2024) (cit. on p. 155).
- Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (2020). “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: <https://aclanthology.org/2020.acl-main.485> (cit. on p. 154).
- Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach (2021). “Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association

- for Computational Linguistics, pp. 1004–1015. DOI: 10.18653/v1/2021.acl-long.81. URL: <https://aclanthology.org/2021.acl-long.81> (cit. on p. 156).
- Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Red Hook, NY, USA: Curran Associates Inc., pp. 4356–4364. ISBN: 978-1-5108-3881-9 (cit. on p. 151).
- Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (2015). “A Large Annotated Corpus for Learning Natural Language Inference”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: <https://aclanthology.org/D15-1075> (cit. on pp. 119, 124).
- Bradley, Andrew P. (1997). “The use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms”. In: *Pattern Recognition* 30.7, pp. 1145–1159. ISSN: 0031-3203. DOI: [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2). URL: <https://www.sciencedirect.com/science/article/pii/S0031320396001422> (cit. on p. 128).
- Bražinskas, Arthur, Mirella Lapata, and Ivan Titov (2020). “Unsupervised Opinion Summarization as Copycat-Review Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5151–5169. DOI: 10.18653/v1/2020.acl-main.461. URL: <https://aclanthology.org/2020.acl-main.461> (cit. on p. 13).
- Brodersen, Kay Henning, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann (2010). “The Balanced Accuracy and Its Posterior Distribution”. In: *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764 (cit. on p. 127).

- Brown, Hannah and Reza Shokri (2023). *How (Un)Fair is Text Summarization?* URL: <https://openreview.net/forum?id=-UsbRlXzMG> (cit. on pp. 158, 167).
- Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling (2011). “Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?” In: *Perspectives on Psychological Science* 6.1, pp. 3–5. ISSN: 1745-6916, 1745-6924. DOI: 10.1177/1745691610393980. URL: <http://journals.sagepub.com/doi/10.1177/1745691610393980> (visited on 03/20/2024) (cit. on p. 37).
- Callison-Burch, Chris (2009). “Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Ed. by Philipp Koehn and Rada Mihalcea. Singapore: Association for Computational Linguistics, pp. 286–295. URL: <https://aclanthology.org/D09-1030> (cit. on p. 37).
- Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (2007). “(Meta-) Evaluation of Machine Translation”. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz. Prague, Czech Republic: Association for Computational Linguistics, pp. 136–158. URL: <https://aclanthology.org/W07-0718> (cit. on p. 40).
- Cao, Meng, Yue Dong, and Jackie Cheung (2022). “Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3340–3354. DOI: 10.18653/v1/2022.acl-long.236. URL: <https://aclanthology.org/2022.acl-long.236> (cit. on p. 160).
- Card, Dallas, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky (2020). “With Little Power Comes Great Responsibility”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9263–9274. DOI: 10.18653/v1/2020.emnlp-main.745. URL: <https://www.aclweb.org/anthology/2020.emnlp-main.745> (cit. on p. 43).

- Carletta, Jean (1996). “Assessing Agreement on Classification Tasks: The Kappa Statistic”. In: *Computational Linguistics* 22.2. Ed. by Julia Hirschberg, pp. 249–254. URL: <https://aclanthology.org/J96-2004> (cit. on pp. 33, 63).
- Chaganty, Arun, Stephen Mussmann, and Percy Liang (2018). “The Price of Debiasing Automatic Metrics in Natural Language Evaluation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 643–653. DOI: 10.18653/v1/P18-1060. URL: <https://www.aclweb.org/anthology/P18-1060> (cit. on pp. 41, 195).
- Chen, Guiming Hardy, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang (2024). *Humans or LLMs as the Judge? A Study on Judgement Biases*. URL: <http://arxiv.org/abs/2402.10669> (visited on 07/27/2024) (cit. on p. 68).
- Chen, Yanran and Steffen Eger (2023). “MENLI: Robust Evaluation Metrics from Natural Language Inference”. In: *Transactions of the Association for Computational Linguistics* 11, pp. 804–825. DOI: 10.1162/tac1_a_00576. URL: <https://aclanthology.org/2023.tac1-1.47> (cit. on pp. 24, 120).
- Chen, Yen-Chun and Mohit Bansal (2018). “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 675–686. DOI: 10.18653/v1/P18-1063. URL: <https://aclanthology.org/P18-1063> (cit. on p. 50).
- Cheng, Jianpeng and Mirella Lapata (2016). “Neural Summarization by Extracting Sentences and Words”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 484–494. DOI: 10.18653/v1/P16-1046. URL: <https://aclanthology.org/P16-1046> (cit. on p. 15).
- Cheng, Myra, Esin Durmus, and Dan Jurafsky (2023). “Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan

- Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1504–1532. DOI: 10.18653/v1/2023.acl-long.84. URL: <https://aclanthology.org/2023.acl-long.84> (cit. on pp. 152, 156, 186).
- Chiang, Cheng-Han and Hung-yi Lee (2023). “Can Large Language Models Be an Alternative to Human Evaluations?” In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 15607–15631. DOI: 10.18653/v1/2023.acl-long.870. URL: <https://aclanthology.org/2023.acl-long.870> (cit. on pp. 25, 178).
- Chieu, Hai Leong and Yoong Keok Lee (2004). “Query based event extraction along a timeline”. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. Sheffield United Kingdom: ACM, pp. 425–432. ISBN: 978-1-58113-881-8. DOI: 10.1145/1008992.1009065. URL: <https://dl.acm.org/doi/10.1145/1008992.1009065> (cit. on p. 13).
- Christensen, Rune Haubo Bojesen (2019). *ordinal—Regression Models for Ordinal Data*. URL: <https://cran.r-project.org/web/packages/ordinal/ordinal.pdf> (visited on 08/26/2024) (cit. on p. 59).
- Chu, Eric and Peter J. Liu (2019). “MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarization”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1223–1232. URL: <http://proceedings.mlr.press/v97/chu19b.html> (cit. on p. 13).
- Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith (2021). “All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

- Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 7282–7296. DOI: 10.18653/v1/2021.acl-long.565. URL: <https://aclanthology.org/2021.acl-long.565> (cit. on p. 193).
- Clark, Elizabeth, Asli Celikyilmaz, and Noah A. Smith (2019). “Sentence Mover’s Similarity: Automatic Evaluation for Multi-Sentence Texts”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2748–2760. DOI: 10.18653/v1/P19-1264. URL: <https://aclanthology.org/P19-1264> (cit. on p. 24).
- Clark, Elizabeth, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roei Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh (2023). “SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 9397–9413. DOI: 10.18653/v1/2023.emnlp-main.584. URL: <https://aclanthology.org/2023.emnlp-main.584> (cit. on pp. 19, 22, 26, 87).
- Cohan, Arman, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian (2018). “A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 615–621. DOI: 10.18653/v1/N18-2097. URL: <http://aclweb.org/anthology/N18-2097> (cit. on p. 255).
- Cohen, Jacob (1960). “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1, pp. 37–46 (cit. on p. 33).
- Cortes, Corinna and Vladimir Vapnik (1995). “Support-Vector Networks”. In: *Machine Learning* 20.3, pp. 273–297. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00994018. URL: <http://link.springer.com/10.1007/BF00994018> (cit. on p. 94).

- Dang, Hoa and Karolina Owczarzak (2009a). *Overview of the TAC 2008 Update Summarization Task*. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=903465 (cit. on pp. 2, 13, 77).
- Dang, Hoa Trang (2005). “Overview of DUC 2005”. In: *Proceedings of the Document Understanding Conference*. Vol. 2005. Citeseer, pp. 1–12. URL: <https://duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf> (cit. on pp. 10, 22, 30, 86, 92, 255).
- (2006). “Overview of DUC 2006”. In: *In Proceedings of HLT-NAACL 2006*. URL: <https://www-nlpir.nist.gov/projects/duc/pubs/2006papers/duc2006.pdf> (cit. on pp. 10, 22, 86, 92, 255).
- Dang, Hoa Trang and Karolina Owczarzak (2009b). *Overview of TAC 2009 Summarization Track*. URL: https://tac.nist.gov/publications/2009/presentations/TAC2009_Summ_overview.pdf (cit. on pp. 37, 77, 86).
- Darrin, Maxime, Philippe Formont, Jackie Cheung, and Pablo Piantanida (2024). “COSMIC: Mutual Information for Task-Agnostic Summarization Evaluation”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 12696–12717. URL: <https://aclanthology.org/2024.acl-long.686> (cit. on p. 25).
- Dash, Abhisek, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty (2019). “Summarizing User-Generated Textual Content: Motivation and Methods for Fairness in Algorithmic Summaries”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW. DOI: 10.1145/3359274. URL: <https://doi.org/10.1145/3359274> (cit. on p. 158).
- Deutsch, Daniel, Tania Bedrax-Weiss, and Dan Roth (2021a). “Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary”. In: *Transactions of the Association for Computational Linguistics* 9, pp. 774–789. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00397. URL: https://doi.org/10.1162/tacl_a_00397 (cit. on p. 24).
- Deutsch, Daniel, Rotem Dror, and Dan Roth (2021b). “A Statistical Analysis of Summarization Evaluation Metrics using Resampling Methods”. In: *Transactions of the Association for Computational Linguistics* 9, pp. 1132–1146. DOI:

- 10.1162/tacl_a_00417. URL: <https://aclanthology.org/2021.tacl-1.67> (cit. on pp. 71, 78).
- Deutsch, Daniel, Rotem Dror, and Dan Roth (2022). “Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 6038–6052. DOI: 10.18653/v1/2022.naacl-main.442. URL: <https://aclanthology.org/2022.naacl-main.442> (cit. on p. 84).
- Deutsch, Daniel, George Foster, and Markus Freitag (2023). “Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12914–12929. DOI: 10.18653/v1/2023.emnlp-main.798. URL: <https://aclanthology.org/2023.emnlp-main.798> (cit. on p. 77).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423> (cit. on pp. 16, 83, 102, 103).
- Dhamala, Jwala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta (2021). “BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, pp. 862–872. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445924. URL: <https://doi.org/10.1145/3442188.3445924> (cit. on p. 156).

- Dinan, Emily, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams (2020). “Multi-Dimensional Gender Bias Classification”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 314–331. DOI: 10.18653/v1/2020.emnlp-main.23. URL: <https://aclanthology.org/2020.emnlp-main.23> (cit. on pp. 152, 153).
- Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston (2019). “Wizard of Wikipedia: Knowledge-Powered Conversational Agents”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=r1173iRqKm> (cit. on p. 134).
- Douglas, Benjamin D., Patrick J. Ewell, and Markus Brauer (2023). “Data Quality in Online Human-Subjects Research: Comparisons Between MTurk, Prolific, CloudResearch, Qualtrics, and SONA”. In: *PLOS ONE* 18.3. Ed. by Jeffrey S. Hallam, e0279720. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0279720. URL: <https://dx.plos.org/10.1371/journal.pone.0279720> (visited on 03/20/2024) (cit. on p. 39).
- Dror, Rotem, Gili Baumer, Segev Shlomov, and Roi Reichart (2018). “The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1383–1392. DOI: 10.18653/v1/P18-1128. URL: <https://aclanthology.org/P18-1128> (cit. on p. 41).
- Durmus, Esin, He He, and Mona Diab (2020). “FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5055–5070. DOI: 10.18653/v1/2020.acl-main.454. URL: <https://www.aclweb.org/anthology/2020.acl-main.454> (cit. on p. 125).
- Durmus, Esin, Faisal Ladhak, and Tatsunori Hashimoto (2022). “Spurious Correlations in Reference-Free Evaluation of Text Generation”. In: *Proceedings of the*

- 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1443–1454. DOI: 10.18653/v1/2022.acl-long.102. URL: <https://aclanthology.org/2022.acl-long.102> (cit. on p. 85).
- Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter (2022). “Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark”. In: *Transactions of the Association for Computational Linguistics* 10, pp. 1066–1083. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00506. URL: https://doi.org/10.1162/tacl_a_00506 (cit. on pp. 122, 126, 127).
- Edmundson, H. P. (1969). “New Methods in Automatic Extracting”. In: *Journal of the ACM* 16.2, pp. 264–285. ISSN: 0004-5411, 1557-735X. DOI: 10.1145/321510.321519. URL: <https://dl.acm.org/doi/10.1145/321510.321519> (visited on 03/01/2024) (cit. on p. 14).
- Elhady, Ahmed, Khaled Elsayed, Eneko Agirre, and Mikel Artetxe (2024). “Improving Factuality in Clinical Abstractive Multi-Document Summarization by Guided Continued Pre-training”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 755–761. DOI: 10.18653/v1/2024.naacl-short.66. URL: <https://aclanthology.org/2024.naacl-short.66> (cit. on p. 68).
- Elsner, Micha and Eugene Charniak (2011). “Extending the Entity Grid with Entity-Specific Features”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 125–129. URL: <https://aclanthology.org/P11-2022> (cit. on pp. 74, 81, 82, 92–94).
- Erkan, Günes and Dragomir R. Radev (2004). “LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization”. In: *Journal of Artificial Intelligence Research* 22.1, pp. 457–479. ISSN: 1076-9757 (cit. on p. 14).
- Fabbri, Alexander, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev (2021a). “ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining”. In: *Proceedings of the 59th Annual Meeting of the Association for*

- Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 6866–6880. DOI: 10.18653/v1/2021.acl-long.535. URL: <https://aclanthology.org/2021.acl-long.535> (cit. on p. 13).
- Fabbri, Alexander R, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev (2021b). “Summeval: Re-evaluating summarization evaluation”. In: *Transactions of the Association for Computational Linguistics* 9, pp. 391–409 (cit. on pp. 19, 22, 37, 38, 73, 74, 84, 87, 126, 127).
- Falke, Tobias, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych (2019). “Ranking generated summaries by correctness: An interesting but challenging application for natural language inference”. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2214–2220. DOI: 10.18653/v1/P19-1213. URL: <https://aclanthology.org/P19-1213> (cit. on pp. 119, 120, 124, 126).
- Fatima, Mehwish and Michael Strube (2023). “Cross-lingual Science Journalism: Select, Simplify and Rewrite Summaries for Non-expert Readers”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1843–1861. DOI: 10.18653/v1/2023.acl-long.103. URL: <https://aclanthology.org/2023.acl-long.103> (cit. on p. 14).
- Fawcett, Tom (2006). “An Introduction to ROC Analysis”. In: *Pattern Recognition Letters* 27.8, pp. 861–874. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2005.10.010>. URL: <https://www.sciencedirect.com/science/article/pii/S016786550500303X> (cit. on p. 145).
- Feng, Vanessa Wei and Graeme Hirst (2012). “Extending the Entity-based Coherence Model with Multiple Ranks”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Walter Daelemans. Avignon, France: Association for Computational Linguistics, pp. 315–324. URL: <https://aclanthology.org/E12-1032> (cit. on p. 81).

- Gal, Yarin and Zoubin Ghahramani (2016). “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1050–1059. URL: <https://proceedings.mlr.press/v48/gal16.html> (cit. on p. 130).
- Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed (2024). “Bias and Fairness in Large Language Models: A Survey”. In: *Computational Linguistics*, pp. 1–83. ISSN: 0891-2017. DOI: 10.1162/coli_a_00524. URL: https://doi.org/10.1162/coli_a_00524 (visited on 08/30/2024) (cit. on p. 154).
- Ganesan, Kavita, ChengXiang Zhai, and Jiawei Han (2010). “Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Ed. by Chu-Ren Huang and Dan Jurafsky. Beijing, China: Coling 2010 Organizing Committee, pp. 340–348. URL: <https://aclanthology.org/C10-1039> (cit. on p. 13).
- Gao, Yang, Wei Zhao, and Steffen Eger (2020). “SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1347–1354. DOI: 10.18653/v1/2020.acl-main.124. URL: <https://www.aclweb.org/anthology/2020.acl-main.124> (cit. on p. 25).
- Gao, Yanjun, Chen Sun, and Rebecca J. Passonneau (2019). “Automated Pyramid Summarization Evaluation”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 404–418. DOI: 10.18653/v1/K19-1038. URL: <https://www.aclweb.org/anthology/K19-1038> (cit. on p. 24).
- Gehrmann, Sebastian, Elizabeth Clark, and Thibault Sellam (2023). “Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text”. In: *Journal of Artificial Intelligence Research* 77, pp. 103–166. ISSN: 1076-9757. DOI: 10.1613/jair.1.13715. URL: <https://www.jair.org>

- org/index.php/jair/article/view/13715 (visited on 10/16/2023) (cit. on pp. 2, 37).
- Gholipour Ghalandari, Demian and Georgiana Ifrim (2020). “Examining the State-of-the-Art in News Timeline Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 1322–1334. DOI: 10.18653/v1/2020.acl-main.122. URL: <https://aclanthology.org/2020.acl-main.122> (cit. on p. 13).
- Gillick, Dan and Yang Liu (2010). “Non-Expert Evaluation of Summarization Systems is Risky”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Los Angeles: Association for Computational Linguistics, pp. 148–151. URL: <https://www.aclweb.org/anthology/W10-0722> (cit. on pp. 37, 46, 53).
- Gliwa, Bogdan, Iwona Mochol, Maciej Biesek, and Aleksander Wawer (2019). “SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization”. In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, pp. 70–79. DOI: 10.18653/v1/D19-5409. URL: <https://www.aclweb.org/anthology/D19-5409> (cit. on p. 255).
- Goel, Purvi and Li Chen (2021). “On the Robustness of Monte Carlo Dropout Trained With Noisy Labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2219–2228 (cit. on pp. 121, 130).
- Goldstein, Jade and Jaime Carbonell (1998). “Summarization: (1) Using MMR for Diversity- Based Reranking and (2) Evaluating Summaries”. In: *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 181–195. DOI: 10.3115/1119089.1119120. URL: <https://aclanthology.org/X98-1025> (cit. on p. 15).
- Goyal, Tanya and Greg Durrett (2020). “Evaluating Factuality in Generation with Dependency-level Entailment”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang

- Liu. Online: Association for Computational Linguistics, pp. 3592–3603. DOI: 10.18653/v1/2020.findings-emnlp.322. URL: <https://aclanthology.org/2020.findings-emnlp.322> (cit. on p. 123).
- Goyal, Tanya, Junyi Jessy Li, and Greg Durrett (2022). *News Summarization and Evaluation in the Era of GPT-3*. URL: <http://arxiv.org/abs/2209.12356> (visited on 11/10/2022) (cit. on pp. 1, 17, 19, 21, 194).
- Graff, David and Christopher Cieri (2003). *English Gigaword*. DOI: 10.35111/0Z6Y-Q265. URL: <https://catalog.ldc.upenn.edu/LDC2003T05> (visited on 03/11/2024) (cit. on p. 255).
- Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel (2017). “Can machine translation systems be evaluated by the crowd alone”. In: *Natural Language Engineering* 23.1, pp. 3–30. ISSN: 1351-3249, 1469-8110. DOI: 10.1017/S1351324915000339. URL: https://www.cambridge.org/core/product/identifier/S1351324915000339/type/journal_article (visited on 03/20/2024) (cit. on p. 37).
- Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein (1995). “Centering: A Framework for Modeling the Local Coherence of Discourse”. In: *Computational Linguistics* 21.2. Ed. by Julia Hirschberg, pp. 203–225. URL: <https://aclanthology.org/J95-2003> (cit. on p. 80).
- Grusky, Max, Mor Naaman, and Yoav Artzi (2018). “Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: <https://aclanthology.org/N18-1065> (cit. on pp. 19, 22, 87, 255).
- Guerreiro, Nuno M., Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins (2023). “Hallucinations in Large Multilingual Translation Models”. In: *Transactions of the Association for Computational Linguistics* 11, pp. 1500–1517. DOI: 10.1162/tac1_a_00615. URL: <https://aclanthology.org/2023.tac1-1.85> (cit. on p. 123).
- Guinaudeau, Camille and Michael Strube (2013). “Graph-Based Local Coherence Modeling”. In: *Proceedings of the 51st Annual Meeting of the Association for*

- Computational Linguistics (volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 93–103. URL: <https://aclanthology.org/P13-1010> (cit. on pp. 82, 93, 95, 96).
- Gupta, Prakhar, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong (2022). “Dial-Fact: A Benchmark for Fact-Checking in Dialogue”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3785–3801. DOI: 10.18653/v1/2022.acl-long.263. URL: <https://aclanthology.org/2022.acl-long.263> (cit. on pp. 122, 126, 127).
- Halliday, M.A.K. and Ruqaiya Hasan (2013). *Cohesion in English*. 1st ed. Routledge. ISBN: 978-1-315-83601-0. DOI: 10.4324/9781315836010. URL: <https://www.taylorfrancis.com/books/9781317869603> (visited on 06/15/2024) (cit. on pp. 79, 80).
- Halteren, Hans van and Simone Teufel (2003). “Examining the Consensus Between Human Summaries: Initial Experiments With Factoid Analysis”. In: *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pp. 57–64. URL: <https://aclanthology.org/W03-0508> (cit. on pp. 19, 21).
- Hardt, Moritz, Eric Price, and Nathan Srebro (2016). “Equality of Opportunity in Supervised Learning”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Red Hook, NY, USA: Curran Associates Inc., pp. 3323–3331. ISBN: 978-1-5108-3881-9 (cit. on p. 155).
- He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2020). “DeBERTa: Decoding-Enhanced BERT with Disentangled Attention”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=XPZiaotutsD> (cit. on p. 131).
- Herbrich, Ralf, Tom Minka, and Thore Graepel (2006). “TrueSkill™: a Bayesian skill rating system”. In: *Advances in Neural Information Processing Systems* 19. URL: <http://papers.neurips.cc/paper/3079-trueskilltm-a-bayesian-skill-rating-system.pdf> (cit. on p. 41).
- Hermann, Karl Moritz, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). “Teaching Machines to Read and Comprehend”. In: *Proceedings of the 28th International Conference*

- on Neural Information Processing Systems - Volume 1*. NIPS'15. Cambridge, MA, USA: MIT Press, pp. 1693–1701 (cit. on pp. 11, 13, 16, 50, 87, 133, 164, 255).
- Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long Short-Term Memory”. In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: <https://doi.org/10.1162/neco.1997.9.8.1735> (cit. on p. 98).
- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2020). “The Curious Case of Neural Text Degeneration”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=rygGQyrFvH> (cit. on p. 193).
- Honnayalli, Samhita, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang (2022). “Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 1665–1670. URL: <https://aclanthology.org/2022.lrec-1.177> (cit. on p. 153).
- Honour, David (2016). *What is Krippendorff's Alpha?* Tech. rep. URL: https://github.com/foolswood/krippendorffs_alpha/raw/master/krippendorff.pdf (visited on 08/04/2024) (cit. on p. 35).
- Honovich, Or, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansky, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias (2022). “TRUE: Re-evaluating Factual Consistency Evaluation”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 3905–3920. DOI: 10.18653/v1/2022.naacl-main.287. URL: <https://aclanthology.org/2022.naacl-main.287> (cit. on pp. 78, 119, 120, 127, 131, 132).
- Honovich, Or, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend (2021). “Q²: Evaluating Factual Consistency in Knowledge-Grounded

- Dialogues via Question Generation and Question Answering”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7856–7870. DOI: 10.18653/v1/2021.emnlp-main.619. URL: <https://aclanthology.org/2021.emnlp-main.619> (cit. on pp. 122, 126–128, 131).
- Howcroft, David M., Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Saeed A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser (2020). “Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions”. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Ed. by Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada. Dublin, Ireland: Association for Computational Linguistics, pp. 169–182. DOI: 10.18653/v1/2020.inlg-1.23. URL: <https://aclanthology.org/2020.inlg-1.23> (cit. on p. 29).
- Huang, Luyang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang (2021). “Efficient Attentions for Long Document Summarization”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 1419–1436. DOI: 10.18653/v1/2021.naacl-main.112. URL: <https://aclanthology.org/2021.naacl-main.112> (cit. on p. 255).
- Huang, Nannan, Lin Tian, Haytham Fayek, and Xiuzhen Zhang (2023). “Examining Bias in Opinion Summarisation through the Perspective of Opinion Diversity”. In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Ed. by Jeremy Barnes, Orphée De Clercq, and Roman Klinger. Toronto, Canada: Association for Computational Linguistics, pp. 149–161. DOI: 10.18653/v1/2023.wassa-1.14. URL: <https://aclanthology.org/2023.wassa-1.14> (cit. on p. 158).

- Inouye, David and Jugal K. Kalita (2011). “Comparing Twitter Summarization Algorithms for Multiple Post Summaries”. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 298–306. DOI: 10.1109/PASSAT/SocialCom.2011.31 (cit. on p. 13).
- Iskender, Neslihan, Tim Polzehl, and Sebastian Möller (2021). “Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead”. In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Online: Association for Computational Linguistics, pp. 86–96. URL: <https://www.aclweb.org/anthology/2021.humeval-1.10> (cit. on p. 37).
- Jeon, Sungho and Michael Strube (2020). “Incremental Neural Lexical Coherence Modeling”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6752–6758. DOI: 10.18653/v1/2020.coling-main.594. URL: <https://aclanthology.org/2020.coling-main.594> (cit. on p. 81).
- Jones, Karen Sparck (1999). “Automatic summarising: factors and directions”. In: *Advances in Automatic Text Summarisation*. Ed. by Inderjeet Mani and Mark Maybury. Cambridge MA: MIT Press, pp. 1–12. URL: <https://www.cl.cam.ac.uk/archive/ksj21/ksjdigipapers/summbok99.pdf> (cit. on pp. 9, 10).
- Joty, Shafiq, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen (2018). “Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 558–568. DOI: 10.18653/v1/P18-1052. URL: <https://aclanthology.org/P18-1052> (cit. on p. 82).
- Judd, Charles M, Jacob Westfall, and David A Kenny (2017). “Experiments with more than one random factor: Designs, analytic models, and statistical power”. In: *Annual Review of Psychology* 68, pp. 601–625 (cit. on pp. 63, 65).
- Jung, Taehee, Dongyeop Kang, Lucas Mentch, and Eduard Hovy (2019). “Earlier Isn’t Always Better: Sub-aspect Analysis on Corpus and System Biases

- in Summarization”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3324–3335. DOI: 10.18653/v1/D19-1327. URL: <https://aclanthology.org/D19-1327> (cit. on p. 189).
- Jwalapuram, Prathyusha, Shafiq Joty, and Xiang Lin (2022). “Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6044–6059. DOI: 10.18653/v1/2022.acl-long.418. URL: <https://aclanthology.org/2022.acl-long.418> (cit. on pp. 73, 117).
- Keswani, Vijay and L. Elisa Celis (2021). “Dialect Diversity in Text Summarization on Twitter”. In: *Proceedings of the Web Conference 2021*. WWW ’21. New York, NY, USA: Association for Computing Machinery, pp. 3802–3814. ISBN: 978-1-4503-8312-7. DOI: 10.1145/3442381.3450108. URL: <https://doi.org/10.1145/3442381.3450108> (cit. on p. 158).
- Kiritchenko, Svetlana and Saif Mohammad (2017). “Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 465–470. DOI: 10.18653/v1/P17-2074. URL: <http://aclweb.org/anthology/P17-2074> (cit. on pp. 36, 40).
- (2018). “Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems”. In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Ed. by Malvina Nissim, Jonathan Berant, and Alessandro Lenci. New Orleans, Louisiana: Association for Computational Linguistics, pp. 43–53. DOI: 10.18653/v1/S18-2005. URL: <https://aclanthology.org/S18-2005> (cit. on p. 156).
- Koo, Ryan, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang (2023). *Benchmarking Cognitive Biases in Large Language Models as Evaluators*. URL: <http://arxiv.org/abs/2309.17012> (visited on 07/27/2024) (cit. on p. 25).

- Koo, Ryan, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang (2024). “Benchmarking Cognitive Biases in Large Language Models as Evaluators”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 517–545. URL: <https://aclanthology.org/2024.findings-acl.29> (cit. on pp. 25, 68).
- Kornilova, Anastassia and Vladimir Eidelman (2019). “BillSum: A Corpus for Automatic Summarization of US Legislation”. In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Ed. by Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu. Hong Kong, China: Association for Computational Linguistics, pp. 48–56. DOI: 10.18653/v1/D19-5406. URL: <https://aclanthology.org/D19-5406> (cit. on p. 255).
- Koto, Fajri, Timothy Baldwin, and Jey Han Lau (2022). “FFCI: A Framework for Interpretable Automatic Evaluation of Summarization”. In: *Journal of Artificial Intelligence Research* 73, pp. 1553–1607 (cit. on pp. 73, 87).
- Krippendorff, Klaus (1970). “Estimating the Reliability, Systematic Error and Random Error of Interval Data”. In: *Educational and Psychological Measurement* 30.1, pp. 61–70. DOI: 10.1177/001316447003000105. URL: <https://doi.org/10.1177/001316447003000105> (cit. on pp. 34, 35).
- (2011). *Computing Krippendorff’s Alpha-Reliability*. Tech. rep. URL: <https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf> (cit. on p. 35).
- Kryscinski, Wojciech, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher (2019). “Neural Text Summarization: A Critical Evaluation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 540–551. DOI: 10.18653/v1/D19-1051. URL: <https://www.aclweb.org/anthology/D19-1051> (cit. on p. 44).
- Kryscinski, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher (2020). “Evaluating the Factual Consistency of Abstractive Text Summarization”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

- Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 9332–9346. DOI: 10.18653/v1/2020.emnlp-main.750. URL: <https://aclanthology.org/2020.emnlp-main.750> (cit. on pp. 119, 120, 123, 126, 160).
- Kryscinski, Wojciech, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev (2022). “BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6536–6558. DOI: 10.18653/v1/2022.findings-emnlp.488. URL: <https://aclanthology.org/2022.findings-emnlp.488> (cit. on p. 255).
- Kulkarni, Sayali, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie (2020). *AQuaMuSe: Automatically Generating Datasets for Query-Based Multi-Document Summarization*. URL: <http://arxiv.org/abs/2010.12694> (visited on 03/12/2024) (cit. on p. 10).
- Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva (2017). “Counterfactual Fairness”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf (cit. on p. 155).
- Kusner, Matt J., Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger (2015). “From Word Embeddings to Document Distances”. In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. JMLR.org, pp. 957–966. URL: <http://dl.acm.org/citation.cfm?id=3045118.3045221> (cit. on p. 104).
- Laban, Philippe, Luke Dai, Lucas Bandarkar, and Marti A. Hearst (2021). “Can Transformer Models Measure Coherence in Text: Re-Thinking the Shuffle Test”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 1058–1064. DOI: 10.18653/v1/2021.acl-short.134.

- URL: <https://aclanthology.org/2021.acl-short.134> (cit. on pp. 93, 101, 263).
- Laban, Philippe, Andrew Hsi, John Canny, and Marti A. Hearst (2020). “The Summary Loop: Learning to Write Abstractive Summaries Without Examples”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5135–5150. DOI: 10.18653/v1/2020.acl-main.460. URL: <https://www.aclweb.org/anthology/2020.acl-main.460> (cit. on p. 83).
- Laban, Philippe, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst (2022). “SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization”. In: *Transactions of the Association for Computational Linguistics* 10, pp. 163–177. DOI: 10.1162/tacl_a_00453. URL: <https://aclanthology.org/2022.tacl-1.10> (cit. on pp. 124, 125, 127, 128, 131).
- Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown (2020). “WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 4034–4048. DOI: 10.18653/v1/2020.findings-emnlp.360. URL: <https://aclanthology.org/2020.findings-emnlp.360> (cit. on p. 14).
- Ladhak, Faisal, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto (2023). “When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3206–3219. DOI: 10.18653/v1/2023.eacl-main.234. URL: <https://aclanthology.org/2023.eacl-main.234> (cit. on p. 158).
- Landis, J. Richard and Gary G. Koch (1977). “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1, p. 159. ISSN: 0006341X. DOI: 10.2307/2529310. URL: <https://www.jstor.org/stable/2529310?origin=crossref> (visited on 03/21/2024) (cit. on p. 35).

- Laurer, Moritz, W v Atteveldt, Andreu Casas, and Kasper Welbers (2022). *Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI* (cit. on pp. 131, 265).
- Lee, Chris van der, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer (2019). “Best Practices for the Human Evaluation of Automatically Generated Text”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 355–368. DOI: 10.18653/v1/W19-8643. URL: <https://www.aclweb.org/anthology/W19-8643> (cit. on p. 29).
- Lenth, Russell, Henrik Singmann, Jonathon Love, Paul Buerkner, and Maxime Herve (2018). “Emmeans: Estimated Marginal Means, aka Least-Squares Means”. In: *R package version 1.1*, p. 3 (cit. on p. 59).
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020). “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: <https://aclanthology.org/2020.acl-main.703> (cit. on pp. 1, 16, 50, 152, 172, 193).
- Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda (2023). “Holistic Evaluation of Language Models”. In: *Transactions on Machine Learning Research*. ISSN:

- 2835-8856. URL: <https://openreview.net/forum?id=i04LZibEqW> (cit. on pp. 156, 157, 159, 160, 167, 180).
- Lin, Chin-Yew (2001). *Summary Evaluation Environment User's Guide*. Tech. rep. URL: <http://www1.cs.columbia.edu/nlp/tides/SEEManual.pdf> (cit. on p. 19).
- (2004a). “Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?” In: *NTCIR*. URL: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/ntcir4.pdf> (cit. on p. 76).
- (2004b). “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: <https://www.aclweb.org/anthology/W04-1013> (cit. on pp. 1, 13, 22, 44, 71, 86, 123).
- Lin, Chin-Yew and Eduard Hovy (2002). “Manual and Automatic Evaluation of Summaries”. In: *Proceedings of the ACL-02 Workshop on Automatic Summarization*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 45–51. DOI: 10.3115/1118162.1118168. URL: <https://aclanthology.org/W02-0406> (cit. on p. 19).
- Lin, Hui and Jeff Bilmes (2011). “A Class of Submodular Functions for Document Summarization”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, pp. 510–520. URL: <https://aclanthology.org/P11-1052> (cit. on p. 16).
- Liu, Alisa, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi (2022). “WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6826–6847. DOI: 10.18653/v1/2022.findings-emnlp.508. URL: <https://aclanthology.org/2022.findings-emnlp.508> (cit. on p. 131).
- Liu, Nelson F., Roy Schwartz, and Noah A. Smith (2019a). “Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets”. In: *Proceedings of the*

- 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2171–2179. DOI: 10.18653/v1/N19-1225. URL: <https://aclanthology.org/N19-1225> (cit. on p. 143).
- Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu (2023a). “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 2511–2522. URL: <https://aclanthology.org/2023.emnlp-main.153> (cit. on pp. 25, 178).
- Liu, Yang and Mirella Lapata (2019). “Text Summarization with Pretrained Encoders”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: <https://aclanthology.org/D19-1387> (cit. on pp. 15, 16).
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019b). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. URL: <http://arxiv.org/abs/1907.11692> (visited on 11/15/2021) (cit. on pp. 83, 93, 101).
- Liu, Yixin, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev (2023b). “Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 4140–4170. DOI: 10.18653/v1/2023.acl-long.228. URL: <https://aclanthology.org/2023.acl-long.228> (cit. on p. 20).

- Louviere, Jordan, Terry Flynn, and A. A. J. Marley (2015). *Best-Worst Scaling: Theory, Methods and Applications*. ISBN: 978-1-107-04315-2. DOI: 10.1017/CB09781107337855 (cit. on p. 40).
- Luhn, H. P. (1958). “The Automatic Creation of Literature Abstracts”. In: *IBM Journal of Research and Development* 2.2, pp. 159–165. ISSN: 0018-8646, 0018-8646. DOI: 10.1147/rd.22.0159. URL: <http://ieeexplore.ieee.org/document/5392672/> (visited on 03/01/2024) (cit. on p. 14).
- Mani, Inderjeet, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim (1999). “The TIPSTER SUMMAC Text Summarization Evaluation”. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Henry S. Thompson and Alex Lascarides. Bergen, Norway: Association for Computational Linguistics, pp. 77–85. URL: <https://aclanthology.org/E99-1011> (cit. on pp. 17, 194).
- Mann, William C. and Sandra A. Thompson (1988). “Rhetorical Structure Theory: Toward a functional theory of text organization”. In: *Text - Interdisciplinary Journal for the Study of Discourse* 8.3, pp. 243–281. DOI: doi:10.1515/text.1.1988.8.3.243. URL: <https://doi.org/10.1515/text.1.1988.8.3.243> (visited on 06/15/2024) (cit. on p. 79).
- Martschat, Sebastian and Katja Markert (2017). “Improving ROUGE for Timeline Summarization”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 285–290. URL: <https://aclanthology.org/E17-2046> (cit. on p. 13).
- (2018). “A Temporally Sensitive Submodularity Framework for Timeline Summarization”. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Ed. by Anna Korhonen and Ivan Titov. Brussels, Belgium: Association for Computational Linguistics, pp. 230–240. DOI: 10.18653/v1/K18-1023. URL: <https://aclanthology.org/K18-1023> (cit. on p. 13).
- Mathur, Nitika, Timothy Baldwin, and Trevor Cohn (2017). “Sequence Effects in Crowdsourced Annotations”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2860–2865. DOI: 10.18653/v1/D17-

1306. URL: <https://www.aclweb.org/anthology/D17-1306> (cit. on pp. 38, 39).
- (2020). “Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4984–4997. DOI: 10.18653/v1/2020.acl-main.448. URL: <https://aclanthology.org/2020.acl-main.448> (cit. on p. 76).
- Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (2020). “On Faithfulness and Factuality in Abstractive Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: <https://aclanthology.org/2020.acl-main.173> (cit. on pp. 2, 11, 21, 119, 121–123, 127).
- McCoy, Tom, Ellie Pavlick, and Tal Linzen (2019). “Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: <https://aclanthology.org/P19-1334> (cit. on p. 140).
- McCullagh, P. and J.A. Nelder (2019). *Generalized Linear Models*. 2nd ed. Routledge. ISBN: 978-0-203-75373-6. DOI: 10.1201/9780203753736. URL: <https://www.taylorfrancis.com/books/9781351445856> (visited on 07/03/2024) (cit. on p. 57).
- McDonald, Ryan (2007). “A Study of Global Inference Algorithms in Multi-document Summarization”. In: *Advances in Information Retrieval*. Ed. by Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 557–564. ISBN: 978-3-540-71496-5. URL: https://link.springer.com/chapter/10.1007/978-3-540-71496-5_51 (cit. on p. 15).
- Mesgar, Mohsen, Leonardo F. R. Ribeiro, and Iryna Gurevych (2021). “A Neural Graph-Based Local Coherence Model”. In: *Findings of the Association for*

- Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2316–2321. URL: <https://aclanthology.org/2021.findings-emnlp.199> (cit. on pp. 73, 74, 81–83, 93, 100).
- Mesgar, Mohsen and Michael Strube (2015). “Graph-based Coherence Modeling For Assessing Readability”. In: *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. Ed. by Martha Palmer, Gemma Boleda, and Paolo Rosso. Denver, Colorado: Association for Computational Linguistics, pp. 309–318. DOI: 10.18653/v1/S15-1036. URL: <https://aclanthology.org/S15-1036> (cit. on pp. 81, 82).
- (2016). “Lexical Coherence Graph Modeling Using Word Embeddings”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, pp. 1414–1423. DOI: 10.18653/v1/N16-1167. URL: <https://aclanthology.org/N16-1167> (cit. on pp. 82, 114).
- (2018). “A Neural Local Coherence Model for Text Quality Assessment”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 4328–4339. DOI: 10.18653/v1/D18-1464. URL: <https://aclanthology.org/D18-1464> (cit. on p. 82).
- Mihalcea, Rada and Paul Tarau (2004). “TextRank: Bringing Order into Text”. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pp. 404–411. URL: <https://aclanthology.org/W04-3252> (cit. on p. 14).
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_

files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (cit. on p. 151).

- Mohankumar, Akash Kumar and Mitesh Khapra (2022). “Active Evaluation: Efficient NLG Evaluation with Few Pairwise Comparisons”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 8761–8781. DOI: 10.18653/v1/2022.acl-long.600. URL: <https://aclanthology.org/2022.acl-long.600> (cit. on p. 195).
- Mohiuddin, Tasnim, Prathyusha Jwalapuram, Xiang Lin, and Shafiq Joty (2021). “Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3528–3539. DOI: 10.18653/v1/2021.eacl-main.308. URL: <https://aclanthology.org/2021.eacl-main.308> (cit. on pp. 73, 74, 84, 93, 112).
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2017). “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict”. In: *Political Analysis* 16.4, pp. 372–403. ISSN: 1476-4989. DOI: 10.1093/pan/mpn018. URL: <https://www.cambridge.org/core/article/fightin-words-lexical-feature-selection-and-evaluation-for-identifying-the-content-of-political-conflict/81B3703230D21620B81EB6E2266C7A66> (cit. on pp. 164, 165).
- Montani, Ines, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, Henning Peters, Paul O’Leary McCann, Jim Geovedi, Jim O’Regan, Maxim Samsonov, Daniël De Kok, György Orosz, Marcus Blättermann, Duygu Altinok, Madeesh Kannan, Raphael Mitsch, Søren Lind Kristiansen, Edward, Lj Miranda, Peter Baumgartner, Raphaël Bournhonesque, Richard Hudson, Explosion Bot, Roman, Leander Fiedler, Ryn Daniels, Kadarakos, Wannaphong Phatthiyaphaibun, and Schero1994 (2023). *explosion/spaCy: v3.6.1: Support for Pydantic v2, find-function CLI and more*. DOI: 10.5281/ZENODO.1212303. URL: <https://zenodo.org/record/1212303> (visited on 09/13/2023) (cit. on p. 170).

- Moon, Han Cheol, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu (2019). “A Unified Neural Coherence Model”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2262–2272. DOI: 10.18653/v1/D19-1231. URL: <https://aclanthology.org/D19-1231> (cit. on pp. 82, 83, 93, 98).
- Naik, Aakanksha, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig (2018). “Stress Test Evaluation for Natural Language Inference”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2340–2353. URL: <https://aclanthology.org/C18-1198> (cit. on pp. 140, 143).
- Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2016a). “SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents”. In: abs/1611.04230. URL: <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14636> (cit. on p. 15).
- Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang (2016b). “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond”. In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Ed. by Stefan Riezler and Yoav Goldberg. Berlin, Germany: Association for Computational Linguistics, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: <https://aclanthology.org/K16-1028> (cit. on pp. 16, 23, 255).
- Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman (2020). “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 1953–1967. DOI: 10.18653/v1/2020.emnlp-main.154. URL: <https://aclanthology.org/2020.emnlp-main.154> (cit. on p. 156).
- Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018a). “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for

- Extreme Summarization”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: <https://aclanthology.org/D18-1206> (cit. on pp. 11, 164, 255).
- (2018b). “Ranking Sentences for Extractive Summarization with Reinforcement Learning”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1747–1759. DOI: 10.18653/v1/N18-1158. URL: <https://aclanthology.org/N18-1158> (cit. on pp. 15, 46).
- Nawrath, Marcel, Agnieszka Nowak, Tristan Ratz, Danilo Walenta, Juri Opitz, Leonardo Ribeiro, João Sedoc, Daniel Deutsch, Simon Mille, Yixin Liu, Sebastian Gehrmann, Lining Zhang, Saad Mahamood, Miruna Clinciu, Khyathi Chandu, and Yufang Hou (2024). “On the Role of Summary Content Units in Text Summarization Evaluation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 272–281. DOI: 10.18653/v1/2024.naacl-short.25. URL: <https://aclanthology.org/2024.naacl-short.25> (cit. on p. 24).
- Nenkova, Ani and Kathleen McKeown (2012). “A Survey of Text Summarization Techniques”. In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Boston, MA: Springer US, pp. 43–76. ISBN: 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_3. URL: https://doi.org/10.1007/978-1-4614-3223-4_3 (cit. on p. 14).
- Nenkova, Ani and Rebecca Passonneau (2004). “Evaluating Content Selection in Summarization: The Pyramid Method”. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA:

- Association for Computational Linguistics, pp. 145–152. URL: <https://aclanthology.org/N04-1019> (cit. on pp. 1, 19).
- Ng, Jun-Ping and Viktoria Abrecht (2015). “Better summarization evaluation with word embeddings for ROUGE”. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1925–1930. DOI: 10.18653/v1/D15-1222. URL: <https://aclanthology.org/D15-1222> (cit. on p. 23).
- Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela (2020). “Adversarial NLI: A New Benchmark for Natural Language Understanding”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4885–4901. DOI: 10.18653/v1/2020.acl-main.441. URL: <https://aclanthology.org/2020.acl-main.441> (cit. on pp. 124, 129, 131).
- NIST (2007). *DUC 2007: Task, Documents, and Measures*. URL: <https://duc.nist.gov/duc2007/tasks.html> (visited on 03/11/2024) (cit. on pp. 10, 13, 22, 86, 92, 255).
- Noreen, Eric W. (1989). *Computer-Intensive Methods for Testing Hypotheses: an Introduction*. New York: Wiley. ISBN: 978-0-471-61136-3 (cit. on pp. 42, 60).
- Novikova, Jekaterina, Ondřej Dušek, and Verena Rieser (2018). “RankME: Reliable Human Ratings for Natural Language Generation”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 72–78. DOI: 10.18653/v1/N18-2012. URL: <https://aclanthology.org/N18-2012> (cit. on pp. 31, 41).
- Olabisi, Olubusayo, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal (2022). “Analyzing the Dialect Diversity in Multi-document Summaries”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and

- Seung-Hoon Na. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6208–6221. URL: <https://aclanthology.org/2022.coling-1.542> (cit. on p. 158).
- Opitz, Juri (2024). “Schroedinger’s Threshold: When the AUC Doesn’t Predict Accuracy”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 14400–14406. URL: <https://aclanthology.org/2024.lrec-main.1255> (cit. on p. 128).
- Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko (2009). “Instructional manipulation checks: Detecting satisficing to increase statistical power”. In: *Journal of Experimental Social Psychology* 45.4, pp. 867–872. ISSN: 00221031. DOI: 10.1016/j.jesp.2009.03.009. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0022103109000766> (visited on 03/20/2024) (cit. on p. 38).
- Over, Paul (2001). *Introduction to DUC-2001: an Intrinsic Evaluation of Generic News Text Summarization Systems*. URL: https://www-nlpir.nist.gov/projects/duc/pubs/2001slides/pauls_slides/ (cit. on p. 21).
- Over, Paul, Hoa Dang, and Donna Harman (2007). “DUC in Context”. In: *Information Processing & Management* 43.6, pp. 1506–1520. ISSN: 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2007.01.019>. URL: <https://www.sciencedirect.com/science/article/pii/S0306457307000404> (cit. on pp. 10, 19).
- Over, Paul and Walter Liggett (2002). *Introduction to DUC-2002: an Intrinsic Evaluation of Generic News Text Summarization Systems*. URL: <https://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf> (cit. on pp. 21, 40).
- Over, Paul and James Yen (2003). *An Introduction to DUC-2003*. URL: <https://duc.nist.gov/pubs/2003slides/duc2003intro.pdf> (cit. on pp. 18, 19, 21, 86).
- (2004). *An Introduction to DUC-2004*. URL: <https://duc.nist.gov/pubs/2004slides/duc2004.intro.pdf> (cit. on p. 18).

- Owczarzak, Karolina, John M. Conroy, Hoa Trang Dang, and Ani Nenkova (2012). “An Assessment of the Accuracy of Automatic Evaluation in Summarization”. In: *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Ed. by John M. Conroy, Hoa Trang Dang, Ani Nenkova, and Karolina Owczarzak. Montréal, Canada: Association for Computational Linguistics, pp. 1–9. URL: <https://aclanthology.org/W12-2601> (cit. on p. 77).
- Owczarzak, Karolina and Hoa Trang Dang (2010). *Overview of TAC 2010 Summarization Track*. URL: https://tac.nist.gov/publications/2010/presentations/TAC2010_Summ_Overview.pdf (cit. on pp. 77, 86).
- (2011). *Overview of TAC 2011 Summarization Track*. URL: https://tac.nist.gov/publications/2011/presentations/Summarization2011_overview.presentation.pdf (cit. on pp. 77, 86).
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd (1999). “The PageRank Citation Ranking : Bringing Order to the Web”. In: *The Web Conference*. URL: <https://api.semanticscholar.org/CorpusID:1508503> (cit. on p. 14).
- Pagnoni, Artidoro, Vidhisha Balachandran, and Yulia Tsvetkov (2021). “Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 4812–4829. DOI: 10.18653/v1/2021.naacl-main.383. URL: <https://aclanthology.org/2021.naacl-main.383> (cit. on pp. 85, 126, 127).
- Panickssery, Arjun, Samuel R. Bowman, and Shi Feng (2024). *LLM Evaluators Recognize and Favor Their Own Generations*. URL: <http://arxiv.org/abs/2404.13076> (visited on 07/27/2024) (cit. on pp. 25, 68).
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

- Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: <https://aclanthology.org/P02-1040> (cit. on pp. 23, 71).
- Parrish, Alicia, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman (2022). “BBQ: A hand-built bias benchmark for question answering”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 2086–2105. DOI: 10.18653/v1/2022.findings-acl.165. URL: <https://aclanthology.org/2022.findings-acl.165> (cit. on pp. 152, 168, 184, 187).
- Parrish, Alicia, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman (2021). “Does Putting a Linguist in the Loop Improve NLU Data Collection?” In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4886–4901. DOI: 10.18653/v1/2021.findings-emnlp.421. URL: <https://aclanthology.org/2021.findings-emnlp.421> (cit. on p. 131).
- Parveen, Daraksha, Hans-Martin Ramsel, and Michael Strube (2015). “Topical Coherence for Graph-based Extractive Summarization”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 1949–1954. DOI: 10.18653/v1/D15-1226. URL: <https://aclanthology.org/D15-1226> (cit. on p. 16).
- Pearl, Judea (2009). *Causality: Models, Reasoning and Inference*. 2nd ed. USA: Cambridge University Press. ISBN: 0-521-89560-X (cit. on p. 155).
- Perrella, Stefano, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli (2024). “Guardians of the Machine Translation Meta-Evaluation: Sentinel Metrics Fall In!” In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for

- Computational Linguistics, pp. 16216–16244. URL: <https://aclanthology.org/2024.acl-long.856> (cit. on p. 77).
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: <https://aclanthology.org/N18-1202> (cit. on p. 263).
- Peyrard, Maxime (2019). “Studying Summarization Evaluation Metrics in the Appropriate Scoring Range”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5093–5100. DOI: 10.18653/v1/P19-1502. URL: <https://www.aclweb.org/anthology/P19-1502> (cit. on pp. 44, 86).
- Peyrard, Maxime and Judith Eckle-Kohler (2017). “A Principled Framework for Evaluating Summarizers: Comparing Models of Summary Quality against Human Judgments”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 26–31. DOI: 10.18653/v1/P17-2005. URL: <http://www.aclweb.org/anthology/P17-2005> (cit. on p. 9).
- Pitler, Emily, Annie Louis, and Ani Nenkova (2010). “Automatic Evaluation of Linguistic Quality in Multi-Document Summarization”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 544–554. URL: <https://aclanthology.org/P10-1056> (cit. on pp. 26, 73, 83).
- Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman (2004). “Centering: A Parametric Theory and Its Instantiations”. In: *Computational Linguistics* 30.3, pp. 309–363. DOI: 10.1162/0891201041850911. URL: <https://aclanthology.org/J04-3003> (cit. on pp. 79, 80).
- Popović, Maja (2017). “chrF++: Words Helping Character N-Grams”. In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 612–618. DOI: 10.18653/v1/W17-4770. URL: <https://aclanthology.org/W17-4770> (cit. on p. 91).

- Price, Paul C, Rajiv Jhangiani, I-Chant A Chiang, et al. (2015). *Research Methods in Psychology*. BCCampus (cit. on p. 32).
- Pu, Xiao, Mingqi Gao, and Xiaojun Wan (2024). “Is Summary Useful or Not? An Extrinsic Human Evaluation of Text Summaries on Downstream Tasks”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 9389–9404. URL: <https://aclanthology.org/2024.lrec-main.821> (cit. on pp. 18, 194).
- Qiu, Haoyi, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng (2024). “AMR-Fact: Enhancing Summarization Factuality Evaluation with AMR-Driven Negative Samples Generation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 594–608. DOI: 10.18653/v1/2024.naacl-long.33. URL: <https://aclanthology.org/2024.naacl-long.33> (cit. on p. 124).
- Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska (2000). “Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies”. In: *NAACL-ANLP 2000 Workshop: Automatic Summarization*. URL: <https://aclanthology.org/W00-0403> (cit. on p. 14).
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8, p. 9 (cit. on pp. 134, 158).
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: <http://jmlr.org/papers/v21/20-074.html> (cit. on pp. 131, 134).
- Rashkin, Hannah, David Reitter, Gaurav Singh Tomar, and Dipanjan Das (2021). “Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features”. In: *Proceedings of the 59th Annual Meeting of the Association for*

- Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 704–718. DOI: 10.18653/v1/2021.acl-long.58. URL: <https://aclanthology.org/2021.acl-long.58> (cit. on p. 267).
- Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (2020). “COMET: A Neural Framework for MT Evaluation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: <https://aclanthology.org/2020.emnlp-main.213> (cit. on p. 26).
- Reimers, Nils and Iryna Gurevych (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: <https://aclanthology.org/D19-1410> (cit. on pp. 25, 163).
- Riezler, Stefan and Michael Haggmann (2024). *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science - Second Edition*. Ed. by Graeme Hirst. Synthesis Lectures on Human Language Technologies. Springer. ISBN: 978-3-031-57064-3. DOI: <https://doi.org/10.1007/978-3-031-57065-0>. URL: <https://doi.org/10.1007/978-3-031-57065-0> (cit. on pp. 2, 56).
- Ruan, Jie, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu (2024). “Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.17, pp. 18915–18923. DOI: 10.1609/aaai.v38i17.29857. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/29857> (visited on 06/17/2024) (cit. on p. 195).
- Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme (2018). “Gender Bias in Coreference Resolution”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

- Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 8–14. DOI: 10.18653/v1/N18-2002. URL: <https://aclanthology.org/N18-2002> (cit. on pp. 151, 152, 156, 168).
- Rush, Alexander M., Sumit Chopra, and Jason Weston (2015). “A Neural Attention Model for Abstractive Sentence Summarization”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 379–389. DOI: 10.18653/v1/D15-1044. URL: <http://www.aclweb.org/anthology/D15-1044> (cit. on pp. 16, 255).
- Sadeqi Azer, Erfan, Daniel Khashabi, Ashish Sabharwal, and Dan Roth (2020). “Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5715–5725. DOI: 10.18653/v1/2020.acl-main.506. URL: <https://aclanthology.org/2020.acl-main.506> (cit. on pp. 41, 43).
- Sakaguchi, Keisuke, Matt Post, and Benjamin Van Durme (2014). “Efficient Elicitation of Annotations for Human Evaluation of Machine Translation”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 1–11. DOI: 10.3115/v1/W14-3301. URL: <https://www.aclweb.org/anthology/W14-3301> (cit. on p. 41).
- Sakaguchi, Keisuke and Benjamin Van Durme (2018). “Efficient Online Scalar Annotation with Bounded Support”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 208–218. DOI: 10.18653/v1/P18-1020. URL: <https://www.aclweb.org/anthology/P18-1020> (cit. on pp. 31, 41).
- Sandhaus, Evan (2008). “The New York Times Annotated Corpus”. In: *Linguistic Data Consortium, Philadelphia* 6.12, e26752. URL: https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf (cit. on p. 255).

- Saunders, Danielle and Bill Byrne (2020). “Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 7724–7736. DOI: 10.18653/v1/2020.acl-main.690. URL: <https://aclanthology.org/2020.acl-main.690> (cit. on p. 153).
- Saxena, Rohit and Frank Keller (2024). “Select and Summarize: Scene Saliency for Movie Script Summarization”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 3439–3455. DOI: 10.18653/v1/2024.findings-naacl.218. URL: <https://aclanthology.org/2024.findings-naacl.218> (cit. on p. 68).
- Schlichtkrull, Michael, Thomas N. Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling (2018). “Modeling Relational Data with Graph Convolutional Networks”. In: *The Semantic Web*. Ed. by Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam. Vol. 10843. Cham: Springer International Publishing, pp. 593–607. ISBN: 978-3-319-93417-4. DOI: 10.1007/978-3-319-93417-4_38. URL: https://link.springer.com/10.1007/978-3-319-93417-4_38 (visited on 06/15/2024) (cit. on p. 100).
- Schluter, Natalie (2017). “The Limits of Automatic Summarisation According to ROUGE”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 41–45. URL: <https://www.aclweb.org/anthology/E17-2007> (cit. on p. 44).
- Schoch, Stephanie, Diyi Yang, and Yangfeng Ji (2020). ““This is a Problem, Don’t You Agree?” Framing and Bias in Human Evaluation for Natural Language Generation”. In: *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*. Ed. by Shubham Agarwal, Ondřej Dušek, Sebastian Gehrmann, Dimitra Gkatzia, Ioannis Konstas, Emiel Van Miltenburg, and Sashank Santhanam. Online (Dublin, Ireland): Association for Computational Linguistics, pp. 10–16. URL: <https://aclanthology.org/2020.evalnlgeval-1.2> (cit. on p. 39).

- See, Abigail, Peter J. Liu, and Christopher D. Manning (2017). “Get To The Point: Summarization with Pointer-Generator Networks”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: <http://www.aclweb.org/anthology/P17-1099> (cit. on pp. 16, 23, 44, 48, 255).
- Seshadri, Preethi, Pouya Pezeshkpour, and Sameer Singh (2022). “Quantifying Social Biases Using Templates is Unreliable”. In: *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*. URL: <https://openreview.net/forum?id=rIhzjia7SLa> (cit. on p. 156).
- Shandilya, Anurag, Kripabandhu Ghosh, and Saptarshi Ghosh (2018). “Fairness of Extractive Text Summarization”. In: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. Lyon, France: ACM Press, pp. 97–98. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3186947. URL: <http://dl.acm.org/citation.cfm?doid=3184558.3186947> (visited on 09/05/2023) (cit. on p. 158).
- Shapira, Ori, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan (2019). “Crowdsourcing Lightweight Pyramids for Manual Summary Evaluation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 682–687. DOI: 10.18653/v1/N19-1072. URL: <https://www.aclweb.org/anthology/N19-1072> (cit. on pp. 20, 87).
- Shapira, Ori, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer (2022). *Interactive Query-Assisted Summarization via Deep Reinforcement Learning*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States. DOI: 10.18653/v1/2022.naacl-main.184. URL: <https://aclanthology.org/2022.naacl-main.184> (cit. on p. 194).

- Sharma, Eva, Luyang Huang, Zhe Hu, and Lu Wang (2019a). “An Entity-Driven Framework for Abstractive Summarization”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3271–3282. URL: <https://aclanthology.org/D19-1323> (cit. on p. 50).
- Sharma, Eva, Chen Li, and Lu Wang (2019b). “BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2204–2213. URL: <https://www.aclweb.org/anthology/P19-1212> (cit. on p. 255).
- Shen, Chenhui, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing (2023). “Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 4215–4233. DOI: 10.18653/v1/2023.findings-emnlp.278. URL: <https://aclanthology.org/2023.findings-emnlp.278> (cit. on pp. 25, 178).
- Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng (2019). “The Woman Worked as a Babysitter: On Biases in Language Generation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3407–3412. DOI: 10.18653/v1/D19-1339. URL: <https://aclanthology.org/D19-1339> (cit. on p. 152).
- Shi, Weijia, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih (2024). “Trusting Your Evidence: Hallucinate Less with Context-aware Decoding”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 783–791. DOI: 10.18653/v1/2024.naacl-short.69. URL: <https://aclanthology.org/2024.naacl-short.69> (cit. on p. 68).

- Siddharthan, Advaith and Napoleon Katsos (2012). “Offline Sentence Processing Measures for testing Readability with Users”. In: *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Ed. by Sandra Williams, Advaith Siddharthan, and Ani Nenkova. Montréal, Canada: Association for Computational Linguistics, pp. 17–24. URL: <https://aclanthology.org/W12-2203> (cit. on p. 40).
- Singh, Sameer, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum (2011). “Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, pp. 793–803. URL: <https://aclanthology.org/P11-1080> (cit. on p. 170).
- Srivastava, Aarohi et al. (2023). “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=uyTL5Bvosj> (cit. on p. 156).
- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html> (cit. on pp. 120, 130).
- Stanczak, Karolina and Isabelle Augenstein (2021). *A Survey on Gender Bias in Natural Language Processing*. URL: <http://arxiv.org/abs/2112.14168> (visited on 08/30/2024) (cit. on p. 154).
- Steen, Julius and Katja Markert (2021). “How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 1861–1875. DOI: 10.18653/v1/2021.eacl-main.160. URL: <https://aclanthology.org/2021.eacl-main.160> (cit. on pp. 6, 32).

- Steen, Julius and Katja Markert (2022). “How to Find Strong Summary Coherence Measures? A Toolbox and a Comparative Study for Summary Coherence Measure Evaluation”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6035–6049. URL: <https://aclanthology.org/2022.coling-1.527> (cit. on pp. 6, 74).
- (2024). “Bias in News Summarization: Measures, Pitfalls and Corpora”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 5962–5983. URL: <https://aclanthology.org/2024.findings-acl.356> (cit. on pp. 6, 154).
- Steen, Julius, Juri Opitz, Anette Frank, and Katja Markert (2023). “With a Little Push, NLI Models can Robustly and Efficiently Predict Faithfulness”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 914–924. DOI: 10.18653/v1/2023.acl-short.79. URL: <https://aclanthology.org/2023.acl-short.79> (cit. on pp. 6, 121).
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). “Energy and Policy Considerations for Deep Learning in NLP”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: <https://aclanthology.org/P19-1355> (cit. on p. 148).
- Subbiah, Melanie, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown (2024). *Reading Subtext: Evaluating Large Language Models on Short Story Summarization with Writers*. URL: <http://arxiv.org/abs/2403.01061> (visited on 03/18/2024) (cit. on p. 193).

- Sun, Simeng, Ori Shapira, Ido Dagan, and Ani Nenkova (2019a). “How to Compare Summarizers without Target Length? Pitfalls, Solutions and Re-Examination of the Neural Summarization Literature”. In: *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 21–29. URL: <https://www.aclweb.org/anthology/W19-2303> (cit. on p. 23).
- Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang (2019b). “Mitigating Gender Bias in Natural Language Processing: Literature Review”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 1630–1640. DOI: 10.18653/v1/P19-1159. URL: <https://aclanthology.org/P19-1159> (cit. on p. 156).
- Takeshita, S., T. Green, N. Friedrich, K. Eckert, and S. Ponzetto (2022). “X-SCITLDR: Cross-Lingual Extreme Summarization of Scholarly Documents”. In: *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1–12. URL: <https://doi.ieeecomputersociety.org/> (cit. on p. 14).
- Tang, Liyan, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett (2023). “Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 11626–11644. DOI: 10.18653/v1/2023.acl-long.650. URL: <https://aclanthology.org/2023.acl-long.650> (cit. on p. 127).
- Tang, Xiangru, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev (2022). “Investigating Crowdsourcing Protocols for Evaluating the Factual Consistency of Summaries”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 5680–5692. DOI: 10.18653/v1/2022.naacl-main.417. URL: <https://aclanthology.org/2022.naacl-main.417> (cit. on p. 126).
- Tavakol, Mohsen and Reg Dennick (2011). “Making Sense of Cronbach’s Alpha”. In: *International Journal of Medical Education* 2, pp. 53–55. ISSN: 20426372. DOI: 10.5116/ijme.4dfb.8dfd. URL: <http://www.ijme.net/archive/2/cronbachs-alpha/> (visited on 09/01/2024) (cit. on p. 36).
- Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal (2018). “FEVER: a Large-scale Dataset for Fact Extraction and VERification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: <https://aclanthology.org/N18-1074> (cit. on p. 131).
- Tien Nguyen, Dat and Shafiq Joty (2017). “A Neural Local Coherence Model”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1320–1330. DOI: 10.18653/v1/P17-1121. URL: <https://aclanthology.org/P17-1121> (cit. on pp. 73, 74, 81–83, 93, 96).
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. (2023). *Llama 2: Open Foundation and Fine-tuned Chat Models*. URL: <https://arxiv.org/abs/2307.09288> (visited on 09/09/2024) (cit. on p. 172).
- Tran, Giang, Mohammad Alrifai, and Eelco Herder (2015). “Timeline Summarization from Relevant Headlines”. In: *Advances in Information Retrieval*. Ed. by Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr. Vol. 9022. Cham: Springer International Publishing, pp. 245–256. ISBN: 978-3-319-16354-3. DOI: 10.1007/978-3-319-16354-3_26. URL: http://link.springer.com/10.1007/978-3-319-16354-3_26 (visited on 03/13/2024) (cit. on p. 13).

- United States Census Bureau (1990). *Frequently Occurring Surnames from Census 1990 - names file*. URL: https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html (cit. on p. 168).
- Utama, Prasetya, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych (2022). “Falsesum: Generating Document-level NLI Examples for Recognizing Factual Inconsistency in Summarization”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2763–2776. DOI: 10.18653/v1/2022.naacl-main.199. URL: <https://aclanthology.org/2022.naacl-main.199> (cit. on p. 124).
- Vasilyev, Oleg, Vedant Dharnidharka, and John Bohannon (2020). “Fill in the BLANC: Human-free quality estimation of document summaries”. In: *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Ed. by Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy. Online: Association for Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/2020.eval4nlp-1.2. URL: <https://aclanthology.org/2020.eval4nlp-1.2> (cit. on p. 25).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on pp. 50, 100).
- Verma, Yash, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha (2023). “Large Scale Multi-Lingual Multi-Modal Summarization Dataset”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3620–3632. DOI: 10.18653/v1/2023.eacl-main.263. URL: <https://aclanthology.org/2023.eacl-main.263> (cit. on p. 13).
- Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West (2023). *Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language*

- Models for Text Production Tasks*. URL: <http://arxiv.org/abs/2306.07899> (visited on 03/20/2024) (cit. on p. 39).
- Vinyals, Oriol, Samy Bengio, and Manjunath Kudlur (2015). “Order Matters: Sequence to Sequence for Sets”. In: *Proceedings of ICLR 2016*. San Juan, Puerto Rico. URL: <http://arxiv.org/abs/1511.06391> (cit. on p. 50).
- Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors (2020). “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17, pp. 261–272. DOI: <https://doi.org/10.1038/s41592-019-0686-2> (cit. on p. 53).
- Wagner, Claudia, David Garcia, Mohsen Jadidi, and Markus Strohmaier (2015). “It’s a Man’s Wikipedia? Assessing Gender Inequality in an Online Encyclopedia”. In: *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI), pp. 454–463. URL: <https://kops.uni-konstanz.de/entities/publication/2a845524-e4ed-48a0-a69d-b0a67d719034> (cit. on p. 172).
- Wang, Alex and Kyunghyun Cho (2019). “BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model”. In: *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Ed. by Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 30–36. DOI: 10.18653/v1/W19-2304. URL: <https://aclanthology.org/W19-2304> (cit. on p. 103).
- Wang, Alex, Kyunghyun Cho, and Mike Lewis (2020). “Asking and Answering Questions to Evaluate the Factual Consistency of Summaries”. In: *Proceedings*

- of the 58th Annual Meeting of the Association for Computational Linguistics. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5008–5020. DOI: 10.18653/v1/2020.acl-main.450. URL: <https://aclanthology.org/2020.acl-main.450> (cit. on pp. 119, 125–127).
- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2019). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *Advances in Neural Information Processing Systems* 32 (cit. on p. 131).
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2018). “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: <https://aclanthology.org/W18-5446> (cit. on p. 131).
- Wang, Yike, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov (2023). *Resolving Knowledge Conflicts in Large Language Models*. URL: <http://arxiv.org/abs/2310.00935> (visited on 01/18/2024) (cit. on p. 183).
- Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman (2019). “Neural Network Acceptability Judgments”. In: *Transactions of the Association for Computational Linguistics* 7. Ed. by Lillian Lee, Mark Johnson, Brian Roark, and Ani Nenkova, pp. 625–641. DOI: 10.1162/tac1_a_00290. URL: <https://aclanthology.org/Q19-1040> (cit. on p. 103).
- Waugh, Linda R. (1982). “Marked and Unmarked: A Choice between Unequals in Semiotic Structure”. In: *Semiotica* 38.3-4, pp. 299–318. DOI: doi:10.1515/semi.1982.38.3-4.299. URL: <https://doi.org/10.1515/semi.1982.38.3-4.299> (visited on 09/05/2023) (cit. on p. 186).
- Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann (2013). *OntoNotes Release 5.0*. DOI: 10.35111/XMHB-2B84. URL: [https:](https://)

- `//catalog.ldc.upenn.edu/LDC2013T19` (visited on 09/05/2023) (cit. on p. 168).
- Weng, Rongxiang, Heng Yu, Xiangpeng Wei, and Weihua Luo (2020). “Towards Enhancing Faithfulness for Neural Machine Translation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2675–2684. DOI: 10.18653/v1/2020.emnlp-main.212. URL: <https://aclanthology.org/2020.emnlp-main.212> (cit. on p. 123).
- Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: <https://aclanthology.org/N18-1101> (cit. on p. 131).
- Williams, Edward J. (1959). *Regression Analysis*. Vol. 14. New York, USA: Wiley (cit. on p. 142).
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: <https://aclanthology.org/2020.emnlp-demos.6> (cit. on pp. 263, 266).
- Wolhandler, Ruben, Arie Cattan, Ori Ernst, and Ido Dagan (2022). “How “Multi” is Multi-Document Summarization?” In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg,

- Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5761–5769. DOI: 10.18653/v1/2022.emnlp-main.389. URL: <https://aclanthology.org/2022.emnlp-main.389> (cit. on p. 10).
- Wu, Felix, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli (2019). “Pay Less Attention with Lightweight and Dynamic Convolutions”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=SkVhlh09tX> (cit. on p. 98).
- Xenouelas, Stratos, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos (2019). “SUM-QE: A BERT-based Summary Quality Estimation Model”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 6005–6011. DOI: 10.18653/v1/D19-1618. URL: <https://aclanthology.org/D19-1618> (cit. on pp. 26, 73, 83, 93, 102).
- Xie, Jian, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su (2024). “Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts”. In: *The Twelfth International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=auKAUJZM06> (cit. on p. 183).
- Xu, Weijia, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat (2023). “Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection”. In: *Transactions of the Association for Computational Linguistics* 11, pp. 546–564. DOI: 10.1162/tac1_a_00563. URL: <https://aclanthology.org/2023.tac1-1.32> (cit. on p. 123).
- Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-

- main.41. URL: <https://aclanthology.org/2021.naacl-main.41> (cit. on p. 26).
- Yan, Yiming, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang (2023). “BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 5428–5443. DOI: 10.18653/v1/2023.acl-long.297. URL: <https://aclanthology.org/2023.acl-long.297> (cit. on p. 25).
- Yuan, Weizhe, Graham Neubig, and Pengfei Liu (2021). “BARTScore: Evaluating generated text as text generation”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 27263–27277. URL: <https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf> (cit. on pp. 24, 73, 83, 93, 103, 106).
- Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu (2020a). “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization”. In: *International Conference on Machine Learning*. PMLR, pp. 11328–11339. URL: <https://dl.acm.org/doi/abs/10.5555/3524938.3525989> (cit. on pp. 1, 16, 172, 193).
- Zhang, Lining, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc (2023a). “A Needle in a Haystack: An Analysis of High-Agreement Workers on MTurk for Summarization”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 14944–14982. DOI: 10.18653/v1/2023.acl-long.835. URL: <https://aclanthology.org/2023.acl-long.835> (cit. on p. 38).
- Zhang, Shiyue and Mohit Bansal (2021). “Finding a Balanced Degree of Automation for Summary Evaluation”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine

- Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6617–6632. DOI: 10.18653/v1/2021.emnlp-main.531. URL: <https://aclanthology.org/2021.emnlp-main.531> (cit. on pp. 20, 24).
- Zhang, Shiyue, David Wan, and Mohit Bansal (2023b). “Extractive is not Faithful: An Investigation of Broad Unfaithfulness Problems in Extractive Summarization”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 2153–2174. DOI: 10.18653/v1/2023.acl-long.120. URL: <https://aclanthology.org/2023.acl-long.120> (cit. on p. 21).
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020b). “BERTScore: Evaluating Text Generation with BERT”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=SkeHuCVFDr> (cit. on pp. 24, 71).
- Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto (2024). “Benchmarking Large Language Models for News Summarization”. In: *Transactions of the Association for Computational Linguistics* 12, pp. 39–57. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00632. URL: https://doi.org/10.1162/tacl_a_00632 (visited on 05/03/2024) (cit. on p. 22).
- Zhang, Yuan, Jason Baldridge, and Luheng He (2019). “PAWS: Paraphrase Adversaries from Word Scrambling”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Tamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1298–1308. DOI: 10.18653/v1/N19-1131. URL: <https://aclanthology.org/N19-1131> (cit. on pp. 122, 127).
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (2018). “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and

- Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 15–20. DOI: 10.18653/v1/N18-2003. URL: <https://aclanthology.org/N18-2003> (cit. on pp. 153, 156).
- Zhao, Wei, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger (2019). “MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 563–578. DOI: 10.18653/v1/D19-1053. URL: <https://aclanthology.org/D19-1053> (cit. on pp. 24, 71).
- Zhao, Wei, Michael Strube, and Steffen Eger (2023). “DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3865–3883. DOI: 10.18653/v1/2023.eacl-main.278. URL: <https://aclanthology.org/2023.eacl-main.278> (cit. on p. 84).
- Zheng, Hao and Mirella Lapata (2019). “Sentence Centrality Revisited for Unsupervised Summarization”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6236–6247. URL: <https://www.aclweb.org/anthology/P19-1628> (cit. on p. 15).
- Zhong, Ming, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang (2020). “Extractive Summarization as Text Matching”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 6197–6208. DOI: 10.18653/v1/2020.acl-main.552. URL: <https://aclanthology.org/2020.acl-main.552> (cit. on p. 16).
- Zhong, Ming, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev (2021). “QMSum: A New Benchmark for Query-based Multi-domain

- Meeting Summarization”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 5905–5921. DOI: 10.18653/v1/2021.naacl-main.472. URL: <https://aclanthology.org/2021.naacl-main.472> (cit. on p. 10).
- Zhou, Karen and Chenhao Tan (2023). “Entity-Based Evaluation of Political Bias in Automatic Summarization”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 10374–10386. DOI: 10.18653/v1/2023.findings-emnlp.696. URL: <https://aclanthology.org/2023.findings-emnlp.696> (cit. on p. 158).
- Zhu, Junnan, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong (2018). “MSMO: Multimodal Summarization with Multimodal Output”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 4154–4164. URL: <http://aclweb.org/anthology/D18-1448> (cit. on p. 13).
- Zhu, Wanzheng and Suma Bhat (2020). “GRUEN for Evaluating Linguistic Quality of Generated Text”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 94–108. DOI: 10.18653/v1/2020.findings-emnlp.9. URL: <https://aclanthology.org/2020.findings-emnlp.9> (cit. on pp. 25, 73, 83, 93, 103–105).

Appendix A

Summarization Dataset References

Table A.1 relates the dataset names in Table 2.1 to the corresponding references.

Corpus	Reference
Arxiv	Cohan et al. (2018)
BigPatent	Sharma et al. (2019b)
BillSum	Kornilova and Eidelman (2019)
BookSum Full	Kryscinski et al. (2022)
BookSum Chapter	Kryscinski et al. (2022)
BookSum Paragraph	Kryscinski et al. (2022)
CNN/DM	Hermann et al. (2015); Nallapati et al. (2016b); See et al. (2017)
DUC 2005	Dang (2005)
DUC 2006	Dang (2006)
DUC 2007	NIST (2007)
Gigaword	Graff and Cieri (2003); Rush et al. (2015)
GovReport	Huang et al. (2021)
Multinews	Afli et al. (2017)
Newsroom	Grusky et al. (2018)
NYT	Sandhaus (2008)
PubMed	Cohan et al. (2018)
SamSUM	Gliwa et al. (2019)
XSum	Narayan et al. (2018a)

TABLE A.1: References for datasets mentioned in Table 2.1. Where multiple references are given, the first introduces the corpus, whereas the second pioneers its use as a summarization dataset. For CNN/DM, Nallapati et al. (2016b) introduce the dataset but use an anonymized version, where named entities are replaced with placeholders. See et al. (2017) are the first to use a non-anonymized version.

Appendix B

Survey

B.1 Categories

While most categories are self-explanatory, we elaborate on some of the decisions we made during the survey in Chapter 3 in this section.

Evaluation Questions. We allow a single study to include multiple evaluation questions, as long as all questions are answered by the same annotators and use the same method. We make no distinction between informativeness, coverage, focus, and relevance and summarize them under *Content*. Similarly, we summarize fluency, grammaticality, and readability under *Fluency*. *Other* includes:

- One study with a specialized set of evaluation questions evaluating the usefulness of a generated related work summary
- One study of *polarity* in a sentiment summarization context
- One study where annotators were asked to identify the aspect a summary covers in the context of review summarization
- Two studies evaluating formality and *meaning similarity* of reference and system summary
- One study evaluating diversity
- One study conducting a Turing test

- One study asking paper authors whether they would consider a sentence part of a summary of their own paper.
- One study evaluating structure and topic diversity.

Evaluation Method. *Binary* includes any task with a yes/no style decision, while *pairwise* includes any method in which two systems are ranked against each other. *Other* includes

- The aspect identification task mentioned above
- One study in which participants selected a single best summary out of a set of summaries.

Annotator Recruitment. *Other* includes any recruitment strategy that does not rely on crowdsourcing. This includes cases in which the recruitment was not specified, students, experts, the authors themselves, and various kinds of volunteers.

Statistical Evaluation. *Other/unspecified* includes

- Four studies which reported statistical significance without reporting the test used
- Two studies using the approximate randomization test
- One study using the chi-square test
- One study using a Tukey test without prior ANOVA.

B.2 Survey Files

All papers we considered for the survey are listed in the supplementary material in the file `all_papers.yaml` by their id in the ACL anthology bib-file. The file can be found in the repository for Chapter 3 (see Section 1.6). The 58 SDS/MDS

system papers that contain new human evaluation studies and are thus included in the survey are listed in the category `with_human_eval`.

For the sake of completeness, we further list summarization papers we did not include in our survey. We separate them into the following categories:

no_human_eval 47 SDS/MDS system papers without human evaluation

sentsum 27 Sentence summarization and headline generation papers

non_system 34 summarization papers that do not introduce new systems, like surveys, opinion pieces and evaluation studies

other 10 Papers that conduct summarization with either non-textual input or non-textual output

We give a full list of the survey results for all papers with human evaluation studies in the file `survey_details.csv`. The file has the following columns:

paper Id of the paper in the ACL anthology

eval_id Id of the evaluation study to differentiate them in papers with multiple studies

task Summarization task of the paper: SDS vs. MDS

genre Genre of the summarized documents

#docs Number of documents in the evaluation

#systems Number of systems in the evaluation

includes_reference Whether the reference summary is included in the human evaluation

#ann_total Total number of annotators in the study

#ann_item Number of annotators per summary

content, fluency, repetition, coherence, referential_clarity, other, overall
Binary columns indicating evaluation questions in the paper

measure Annotation method used in the study

anntype Annotator recruitment strategy

statetest Statistical test used

design_specified Indicates whether it is possible to determine the full design from the information given about the study in the paper

comments Comments column. This column describes the use of *other* where present.

B.3 Files for the Repeat of the Survey

Since our second survey in Section 3.8 is smaller in scope, we slightly simplify our reporting. `all_papers_new.yaml` lists all papers we considered as system papers. Unlike in `all_papers.yaml`, we do not list sentence summarization or non-system papers.

`survey_papers_new.csv` contains the detailed evaluation for each paper with at least one human evaluation study.

The one instance of *Other* in evaluation questions is one instance of annotators being asked to rate the fairness of opinion summaries.

Appendix C

Coherence Measures: Implementation Details

C.1 Extended Entity Grid (EEG)

We use the original implementation that is part of the Brown Coherence Toolkit.¹ For preprocessing, we use the Stanford parser.² We identify entities using OpenNLP as suggested in the README.

For WSJ, we used the pretrained `f-wsj` model provided in the toolkit. For CNN/DM, we trained our own model. We found that the implementation ran out of memory on the 287,011 instances in CNN/DM on our machine with 32GB of RAM. We thus limited the instances considered for CNN/DM to 10% of the original dataset (28,701).

C.2 Entity Graph (EGR)

Since there is no reference implementation of the Entity Graph, we implement our own version based on the grid created by the Brown Coherence Toolkit. We use the P_{Acc} measure with distance penalty which performed best in the original paper.

¹<https://web.archive.org/web/20200505174052/https://bitbucket.org/melsner/browncoherence>

²<https://nlp.stanford.edu/software/lex-parser.shtml>

Embedding Size	100
Batch Size	64
Pool Length	6
Window Size	6
Number of Filters	150
Hidden Size	250

TABLE C.1: Best hyper-parameters for the neural entity grid on DUC 03.

C.3 Neural Entity Grid (NEG)

Since no models are publicly available, we train new models for all settings using the reference implementation.³

For DUC 03 and WSJ, we use the entity grids and training pairs provided by the authors in the repository. These were also created using the Brown Coherence Toolkit. For CNN/DM, we create our own samples following the original settings. We found that the original implementation of the shuffling procedure leaves artifacts in the data, since the row order is unchanged between shuffled and unshuffled documents. However, for unshuffled documents, the order of rows in the entity grid roughly corresponds to the order of entities in the sentences, whereas for shuffled documents this is not the case. Since this can be picked up by the convolutional network for short documents, we modify the input data to randomly shuffle the row order for each instance.

For the shuffling tasks on WSJ, we use the hyper-parameters reported in the original paper. We also use these hyper-parameters for CNN/DM. For DUC, no hyper-parameters were reported, so we use the built-in hyper-parameter search. We achieve the best results using the parameters reported in Table C.1.

C.4 Graph-based Model (GRA)

We use the original implementation.⁴ For WSJ, we use the provided pretrained model. For DUC and CNN/DM, we train the model using default settings, which

³https://github.com/datienguyen/cnn_coherence

⁴<https://github.com/UKPLab/emnlp2021-neural-graph-based-coherence-model>

includes an ELMo (Peters et al., 2018) embedding layer. The graph representation is created from an entity grid representation as provided by the Brown Coherence Toolkit.

C.5 Unified Coherence Model (UNF)

We use the original implementation.⁵ We train new models for CNN/DM and WSJ using default settings. In the original implementation, scores are computed using a sum over coherence scores for windows of three sentences each, since in their pairwise evaluation, samples always have the same length. In our experiments, we use the mean over the windows instead to normalize for length. For completeness, we also conducted experiments using the original setting, i.e. the sum instead of the mean, which did not lead to any improvement.

C.6 Coherence Classifier (CCL)

We originally experimented with the pretrained WSJ model provided by the authors of Laban et al. (2021).⁶ However, we found that the model achieved near-random scores when evaluated on SummEval for reasons that are difficult to ascertain as the original training code is unavailable. We thus train our own coherence classifier models for both CNN/DM and WSJ. We use the **roberta-large** model as implemented in the huggingface library (Wolf et al., 2020) in a sequence classification setup. We use a learning rate of $2e - 6$ and train for a maximum of six epochs. We select the best model using F1-score on the validation set.

C.7 BARTScore (BAS)

We reimplement the fine-tuned BARTScore variant using the **bart-large-cnn** checkpoint from the huggingface library. Since the original model is evaluated using Spearman’s ρ , we separately verified that it exactly reproduces the reported results.

⁵<https://github.com/taasnim/unified-coherence-model>

⁶https://github.com/tingofurro/shuffle_test

C.8 GRUEN (GRN)

We use the scores provided by the official reference implementation.⁷

C.9 SumQE (SQE)

We use the scores provided by the official reference implementation.⁸ We use the Q5 head of the model jointly trained on all three DUC datasets.⁹

⁷<https://github.com/WanzhengZhu/GRUEN>

⁸<https://github.com/nlpauieb/SumQE>

⁹https://archive.org/download/sum-qe/BERT_DUC_all_Q5_Multi%20Task-5.h5

Appendix D

NLI Model Augmentation Training Details

D.1 Hyper-Parameters

Table D.1 lists the hyper-parameter settings for our model. We use the same optimizer hyper-parameters as Laurer et al. (2022) except for an increased batch size and the learning rate. For the latter, we tested three learning rates ($5e - 6$, $5e - 2$, $5e - 1$) and select the one that provided the best loss on the augmented ANLI validation set. We initially ran models for 10,000 steps with a checkpoint every 1,000 steps and selected the checkpoint with the lowest loss on the augmented ANLI validation set. Later, we reduced the number of training steps to 2,000 since we found we would usually select an early checkpoint as validation loss increased later in training, likely related to overfitting on the augmented data.

Parameter	Val.
Warmup Ratio	0.06
Weight Decay	0.01
Effective Batch Size	64

TABLE D.1: Hyper-parameters for training models with augmentations.

D.2 Training

We use the DeBERTa implementation in the huggingface transformers library (Wolf et al., 2020) and trained our model on a single node using two RX6800 GPUs, with one training run taking about three hours. Later experiments with fewer steps cut that time by 80%.

Appendix E

Dataset Bias in BEGIN-v2

While BEGIN-v2 contains a different selection of models and employs a different annotation methodology, we find that it contains a similar form of bias as BEGIN-v1.

BEGIN-v2 uses data from four dialogue systems, but a majority of faithful generations is produced by a single system called CTRL-DIALOG (Rashkin et al., 2021). CTRL-DIALOG is specifically trained to generate less subjective text, which we hypothesize might result in fewer first-person pronouns. Since CTRL-DIALOG also produces more faithful texts, this would lead to a negative correlation between faithfulness and first-person pronouns, similar to what we observe on BEGIN-v1.

We verify this assumption by computing the correlation of a binary variable indicating an instance has been generated by CTRL-DIALOG with a) the faithfulness labels on BEGIN-v2 and b) first-person pronoun occurrence. We find that an instance being generated by CTRL-DIALOG is positively correlated with it having a *faithful* label (Kendall τ w.r.t. faithfulness: 0.48, $p < 0.001$) while being negatively correlated with the number of pronouns (Kendall τ w.r.t. *I*-pronoun occurrence: -0.34, $p < 0.001$). This suggests future evaluations on the BEGIN-v2 might run into similar bias issues.

Appendix F

Bias Experiment Topic Assignment Heuristic

For our demonstration of the effect of input bias in Section 6.6, we require a transparent way to assign a topic to an input document. Following the observations on gender/topic association in Table 6.2, we manually select a small number of tokens that we identify as sport- or family-related. A text is classified by counting the number of occurrences for each word list and selecting the majority class. A tie is classified as *unknown*. We list tokens for both categories in Table F.1. This allows us to create a deterministic, easy-to-verify topic assignment. Note that this assignment is purposefully artificial and non-general. It is not intended as a realistic topic classifier but as a tool to demonstrate how summarizers *might* behave and how this influences bias scores.

Sport	Family
league	family
season	husband
club	wife
game	father
win	mother
team	children
shot	boys
	girls
	baby

TABLE F.1: Words used for topic identification.

Appendix G

Dataset Statistics for Intersectional Biases

We report dataset statistics for the intersectional biases in Tables G.1 to G.4. We find no significant differences between settings.

Corpus	C_{loc}			C_{glob}		
	Avg. Tok.	Avg. Ent.	% Hal.	Avg. Tok.	Avg. Ent.	% Hal.
BART CNN/DM	60.92 σ : 8.67	0.86 σ : 1.26	3.94	60.93 σ : 8.69	0.90 σ : 1.30	4.77
BART XSum	23.19 σ : 6.43	0.22 σ : 0.49	42.77	23.49 σ : 6.57	0.25 σ : 0.53	44.99
Pegasus CNN/DM	56.41 σ : 16.99	0.77 σ : 1.17	3.21	56.51 σ : 16.72	0.81 σ : 1.22	4.97
Pegasus XSum	24.83 σ : 12.81	0.18 σ : 0.46	39.48	24.85 σ : 14.07	0.20 σ : 0.49	39.67
LLama2 7b	172.83 σ : 37.48	0.87 σ : 1.52	2.02	171.82 σ : 38.26	0.85 σ : 1.53	1.90
LLama2 13b	161.37 σ : 39.93	1.28 σ : 1.75	1.28	162.12 σ : 39.17	1.39 σ : 1.89	1.40
LLama2 70b	147.01 σ : 42.44	1.51 σ : 1.91	0.52	148.22 σ : 41.95	1.66 σ : 2.10	0.47

TABLE G.1: Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black male, white female). σ indicates standard deviation.

Corpus	C_{loc}			C_{glob}		
	Avg. Tok.	Avg. Ent.	% Hal.	Avg. Tok.	Avg. Ent.	% Hal.
BART CNN/DM	61.04 σ : 8.64	0.86 σ : 1.26	3.76	61.00 σ : 8.64	0.90 σ : 1.32	4.67
BART XSum	23.24 σ : 6.47	0.22 σ : 0.49	41.76	23.49 σ : 6.61	0.24 σ : 0.53	44.77
Pegasus CNN/DM	56.55 σ : 17.06	0.78 σ : 1.17	3.16	56.58 σ : 16.82	0.81 σ : 1.22	4.69
Pegasus XSum	24.93 σ : 12.91	0.19 σ : 0.47	37.05	24.71 σ : 13.20	0.20 σ : 0.51	38.60
LLama2 7b	172.61 σ : 37.72	0.85 σ : 1.51	1.57	173.18 σ : 37.46	0.87 σ : 1.57	2.04
LLama2 13b	161.44 σ : 39.82	1.29 σ : 1.78	1.39	162.53 σ : 39.09	1.40 σ : 1.92	1.70
LLama2 70b	147.39 σ : 42.71	1.52 σ : 1.91	0.47	148.15 σ : 41.75	1.66 σ : 2.11	0.53

TABLE G.2: Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black male, white male). σ indicates standard deviation.

Corpus	C_{loc}			C_{glob}		
	Avg. Tok.	Avg. Ent.	% Hal.	Avg. Tok.	Avg. Ent.	% Hal.
BART CNN/DM	60.92 σ : 8.63	0.85 σ : 1.27	3.21	60.86 σ : 8.57	0.90 σ : 1.32	4.28
BART XSum	23.19 σ : 6.41	0.22 σ : 0.49	42.68	23.46 σ : 6.51	0.24 σ : 0.52	46.81
Pegasus CNN/DM	56.42 σ : 16.90	0.78 σ : 1.16	3.49	56.77 σ : 16.92	0.82 σ : 1.21	4.87
Pegasus XSum	24.87 σ : 12.77	0.19 σ : 0.49	38.46	24.78 σ : 14.01	0.21 σ : 0.50	39.10
LLama2 7b	172.98 σ : 37.46	0.88 σ : 1.54	1.80	173.24 σ : 37.24	0.89 σ : 1.58	1.57
LLama2 13b	161.62 σ : 39.56	1.29 σ : 1.76	1.60	162.44 σ : 39.14	1.40 σ : 1.91	1.23
LLama2 70b	147.21 σ : 42.73	1.53 σ : 1.92	0.37	147.96 σ : 41.93	1.65 σ : 2.09	0.43

TABLE G.3: Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black female, white female). σ indicates standard deviation.

Corpus	C_{loc}			C_{glob}		
	Avg. Tok.	Avg. Ent.	% Hal.	Avg. Tok.	Avg. Ent.	% Hal.
BART CNN/DM	60.96 σ : 8.60	0.85 σ : 1.26	3.30	61.07 σ : 8.63	0.90 σ : 1.31	4.24
BART XSum	23.22 σ : 6.45	0.22 σ : 0.50	41.99	23.44 σ : 6.50	0.24 σ : 0.52	44.22
Pegasus CNN/DM	56.32 σ : 16.99	0.78 σ : 1.17	3.11	56.71 σ : 16.81	0.82 σ : 1.20	4.43
Pegasus XSum	24.86 σ : 12.65	0.19 σ : 0.48	36.29	24.65 σ : 13.82	0.21 σ : 0.51	38.73
LLama2 7b	173.26 σ : 37.28	0.87 σ : 1.50	1.79	172.71 σ : 37.73	0.86 σ : 1.55	1.82
LLama2 13b	161.31 σ : 39.48	1.29 σ : 1.78	1.34	162.80 σ : 38.97	1.40 σ : 1.93	1.52
LLama2 70b	146.97 σ : 42.49	1.54 σ : 1.94	0.35	148.16 σ : 41.92	1.66 σ : 2.09	0.41

TABLE G.4: Average number of tokens and entities, and percentage of all entities tagged as hallucinated for summaries generated on racial bias data (black female, white male). σ indicates standard deviation.