Cross-lingual Single-document Abstractive Summarization for Science Journalism

Mehwish Fatima

Department of Computational Linguistics
Heidelberg University

This dissertation is submitted for the degree of
Doctor of Philosophy

Referent: Prof. Dr. Michael Strube
Korreferent: Prof. Dr. Katja Markert
Einreichung: 23.03.2023
Disputation: 05.02.2024

Abstract

We introduce a new task - Cross-lingual Science Journalism as a use case of cross-lingual single-document abstractive summarization. Cross-lingual Science Journalism aims to generate popular science summaries in a local language from scientific articles in a source language for non-expert readers. The cross-lingual popular summaries have distinct properties from regular scientific texts, *e.g.*, conciseness and more readability than source articles and different language from the source. Cross-lingual Science Journalism aims to bridge the gap between the curious local community and scientific research. A real-world example of Cross-lingual Science Journalism is Spektrum der Wissenschaft which converts complex English scientific articles into popular science summaries in German for non-expert audiences.

In this thesis, we focus on (1) curating datasets for summarization in general and science journalism in particular, (2) analyzing the performance of existing summarization models, and (3) developing and evaluating models for Cross-lingual Science Journalism.

For data collection and verification, we create two cross-lingual summarization datasets from online sources by devising systematic methods. Our datasets are collected from Spektrum der Wissenschaft and Wikipedia Science Portal for the English-German language pair. A part of the Spektrum dataset comes from their private domain, so it is only accessible with authorized access. However, the Wikipedia dataset is collected from the public domain and is publicly available to the research community. We perform a thorough analysis based on different statistical and readability features to investigate the linguistic properties of our datasets.

As the second step, we evaluate the collected datasets for the summarization task. For this purpose, we apply several existing summarization models and different training and evaluation strategies. We evaluate the performance of existing summarization models on our datasets with automatic and human evaluation. We also analyze the performance of existing abstractive models to find the limitations of those models.

To address the limitations of existing models, we create a pipeline model - **S**elect, **S**implify and **R**ewrite (SSR). The SSR model combines an extractive summarizer, a simplification model and a cross-lingual abstractive summarizer to generate cross-lingual popular science summaries. We empirically investigate the performance of SSR on our datasets and explore

the contribution of each component to the performance with three different evaluation metrics. The SSR model performs better than the strong baselines with 99% confidence, further suggested by human judgment and readability analysis.

We further investigate Cross-lingual Science Journalism by developing an end-to-end model - joint training of **Sim**plification and **C**ross-lingual **Sum**marization (SimCSum) to improve generated summaries' quality. We empirically evaluate the SimCSum model against several baselines with three evaluation metrics. The SimCSum model outperforms the baselines with 99% confidence, further indicated by human evaluation, readability and error analysis.

To conclude, this work provides a preliminary foundation for a new Cross-lingual Science Journalism task and can help it flourish. In the future, our Wikipedia dataset can help the community to explore and further extend the research in the cross-lingual scientific summarization and Cross-lingual Science Journalism fields. Moreover, our models provide a ground for developing cross-lingual scientific summarization and journalism models. Our models are based on generalized methods, so these models and their derived solutions can be deployed in other domains.

Zusammenfassung

Wir beschreiben einen neuen Aufgabentyp - den sprachübergreifenden Wissenschaftsjournalismus im Bereich der abstraktiven sprachübergreifenden Zusammenfassung von Einzel-Dokumenten. Sprachübergreifender Wissenschaftsjournalismus zielt darauf ab, populärwissenschaftliche Zusammenfassungen in einer lokalen Sprache aus wissenschaftlichen Artikeln in einer Ausgangssprache für nicht fachkundige Leser zu erstellen. Diese sprachübergreifenden populären Zusammenfassungen haben bestimmte Eigenschaften, die sie von wissenschaftlichen Texten unterscheiden, wie etwa deren Prägnanz und bessere Lesbarkeit im Vergleich zum Ausgangsmaterial sowie einer anderen als der Ursprungssprache. Sprachübergreifender Wissenschaftsjournalismus soll die Kluft zwischen einer interessierten lokalen Community und der wissenschaftlichen Forschung überbrücken. Ein Praxisbeispiel für sprachübergreifenden Wissenschaftsjournalismus ist Spektrum der Wissenschaft, das komplexe englische Fachartikel in populärwissenschaftliche Zusammenfassungen in deutscher Sprache für ein nicht fachkundiges Publikum umwandelt.

In dieser Arbeit konzentrieren wir uns auf (1) die Kuratierung von Datensätzen für Zusammenfassungen im Allgemeinen und Journalismus im Besonderen, (2) die Analyse der Leistung bestehender abstraktiver Zusammenfassungsmodelle und (3) die Entwicklung und Evaluierung von Modellen für den sprachübergreifenden Wissenschaftsjournalismus.

Für die Datensammlung und -überprüfung erstellen wir systematisch zwei sprachübergreifende Zusammenfassungsdatensätze aus Online-Quellen. Diese stammen von Spektrum der Wissenschaft und vom Wikipedia Science Portal für das Sprachpaar Englisch-Deutsch. Ein Teil des Spektrum-Datensatzes ist privat und folglich nur mit Autorisierung zugänglich. Der Wikipedia-Datensatz hingegen ist in der Public Domain und somit für die Forschungsgemeinschaft frei verfügbar. Wir führen eine gründliche Analyse auf der Basis verschiedener Lesbarkeits- und statistischer Merkmale durch, um die sprachlichen Eigenschaften unserer Datensätze zu untersuchen.

Im zweiten Schritt evaluieren wir die gesammelten Datensätze für die Zusammenfassungsaufgabe. Dazu verwenden wir verschiedene bestehende Zusammenfassungsmodelle und unterschiedliche Trainings- und Bewertungsstrategien. Wir bewerten die Leistung bestehen-

der Zusammenfassungsmodelle auf unseren Datensätzen sowohl automatisch als auch menschlich. Außerdem analysieren wir die Leistung bestehender abstraktiver Modelle, um die Grenzen dieser Modelle zu ermitteln.

Um die Grenzen bestehender Modelle zu überwinden, haben wir ein Pipeline-Modell entwickelt - **S**elect, **S**implify und **R**ewrite (SSR). Das SSR-Modell kombiniert einen extraktiven Summarizer, ein Vereinfachungsmodell und einen sprachübergreifenden abstraktiven Summarizer, um sprachübergreifende populärwissenschaftliche Zusammenfassungen zu erstellen. Wir untersuchen empirisch die Leistung von SSR auf unseren Datensätzen und erforschen den Beitrag jeder Komponente an der Leistung anhand drei verschiedener Bewertungsmaßstäbe. Das SSR-Modell schneidet mit 99% Konfidenz besser ab als die starken Baselines, was auch durch die menschliche Beurteilung und die Lesbarkeitsanalyse bestätigt wird.

Wir untersuchen den sprachübergreifenden Wissenschaftsjournalismus weiter, indem wir ein End-to-End-Modell entwickeln - gemeinsames Training von **Sim**plification und **C**rosslingual **Sum**marization (SimCSum), um die Qualität der generierten Zusammenfassungen zu verbessern. Wir evaluieren das SimCSum-Modell empirisch gegen mehrere Baselines unter Verwendung dreier Evaluationsmetriken. Das SimCSum-Modell übertrifft die Basismodelle mit 99% Konfidenz, was auch durch die menschliche Bewertung, die Lesbarkeits- und die Fehleranalyse belegt wird.

Zusammenfassend stellt diese Arbeit also eine vorläufige Grundlage für eine neue Aufgabe im Bereich des sprachübergreifenden Wissenschaftsjournalismus dar und kann diesem zum Erfolg verhelfen. In Zukunft kann unser Wikipedia-Datensatz der Community dazu dienen, die Forschung in den Bereichen der sprachübergreifenden wissenschaftlichen Zusammenfassung und des sprachübergreifenden Wissenschaftsjournalismus weiter auszubauen. Darüber hinaus bieten unsere Modelle eine Grundlage für die Entwicklung sprachübergreifender wissenschaftlicher Zusammenfassungs- und Journalismusmodelle. Sie basieren auf verallgemeinerten Methoden, so dass diese Modelle und ihre abgeleiteten Lösungen in anderen Bereichen eingesetzt werden können.

Dedication

To my father, Raja Abdul Hameed, who gave me the wings to fly and explore the world.

Acknowledgements

Thanks go first to my advisor, Michael Strube, who did a great job of pushing, questioning, interpreting and suggesting as the research progressed. I would also like to thank Spektrum der Wissenschaft for providing us with their data. Thanks also go to the Higher Education Commission (HEC) of Pakistan, the Deutscher Akademischer Austausch Dienst (DAAD), Germany and the Heidelberg Institute for Theoretical Studies (HITS) (supported by the Klaus Tschira Foundation), Heidelberg, Germany, for funding and giving me an opportunity to conduct this research.

Thanks to Tim Kolber for helping with this research with some implementation parts, analysis and annotations. Thanks to Nadia Arslan, Fabian Düker and Jason Brockmeyer for participating in data collection and their annotations. I would also like to thank some other university fellows who helped by providing manual annotations. I would also like to pay my gratitude for the friendly growth environment at HITS and their administration for having my back. Thanks to my (former and current) colleagues, Mark-Christoph Müller, Ivan Sekulić, Wei Liu, Haixia Chai and many others at HITS for their beneficial research discussions and feedback. A special thanks to Ghadeer Mobasher, who invested her time in giving me quality feedback for my research and helped me pull through my bad days here as a friend.

I would also like to thank my family for their constant support and prayers, my eldest brother for his continuous help, and for providing his Grammarly account for hassle-free research writing. The most special thanks go to my husband, who is waiting patiently for me on the other side of the world. We got married during my Ph.D., and due to Covid restrictions and some other work and visa conditions, we spent most of our time apart from each other. His unconditional support and gentle demeanor helped me to focus on my research, grow as a better person and survive my toughest days.

Finally, I would like to thank Michael Strube again, who encouraged and pushed me hard to write this thesis while I got trapped in imposter syndrome. Thanks, Michael.

Contents

Part I

Introduction

Chapter 1

Introduction

"I do not read a book; I hold a conversation with the author."

Elbert Hubbard

Summarization is a Natural Language Processing (NLP) task that summarizes a given text by compressing it while preserving its salient points. It can be categorized into different categories, *e.g.*, based on (i) *input type*: single-document or multi-document - if the given text consists of a single document or multiple documents, (ii) *model type*: extractive or abstractive - extractive models rely on selecting important sentences from the given text, while abstractive models additionally rewrite the important information selected from the given text to make summary human-like, and (iii) *domain type* - source texts from blogs, news, tweets, scientific documents, and so on. Figure 1.1 illustrates different categories of summarization depending on the input type, output type, summarization models and domains.

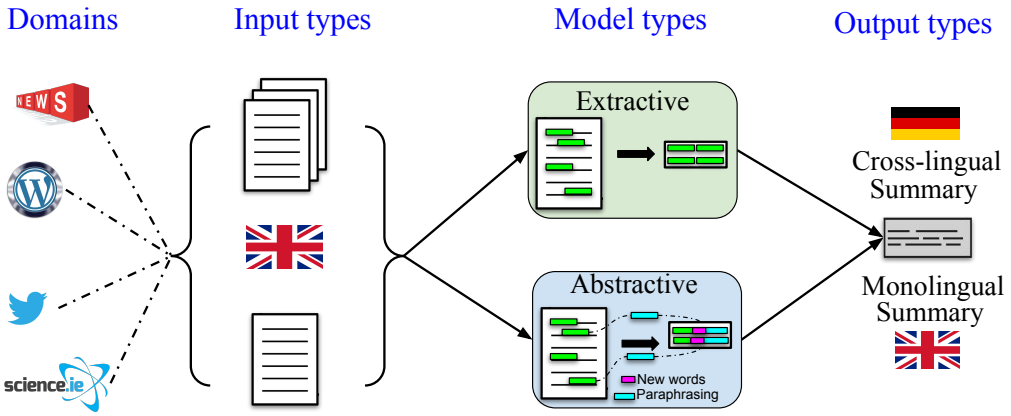

Figure 1.1: Categories of summarization based on domains, input, model, and output types.

A lot of research has been conducted for the news domain summarization (Hermann et al., 2015; Grusky et al., 2018; See et al., 2017; Nallapati et al., 2017; Gehrmann et al., 2018), specifically in monolingual summarization. Recently, the summarization community has shifted its focus from monolingual to cross-lingual summarization (Zhu et al., 2019, 2020; Ouyang et al., 2019). By definition, cross-lingual summarization aims to generate a human-like summary in a target language from a given document in a source language. It is a complex task that combines text summarization (Paice, 1980, 1990) and machine translation (Brown et al., 1990, 1993), both of which have seen significant advances over time, yet are still unsolved. Despite advancements in the summarization field, scientific texts have yet to catch much attention; particularly, cross-lingual scientific summarization is an extremely understudied field with no available resources. Cross-lingual summarization for scientific texts is challenging because it deals with long scientific texts to find the salient information in different sections and summarizes it in a language other than the source.

EXAMPLE 1

Source: "Coronavirus disease 2019 (COVID-19) is a contagious disease caused by a virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease quickly spread worldwide, resulting in the COVID-19 pandemic. The symptoms of COVID-19 are variable but often include fever, cough, headache, fatigue, breathing difficulties, loss of smell, and loss of taste. Symptoms may begin one to fourteen days after exposure to the virus. At least a third of people who are infected do not develop noticeable symptoms. Of those who develop symptoms noticeable enough to be classified as patients, most (81%) develop mild to moderate symptoms (up to mild pneumonia), while 14% develop severe symptoms (dyspnea, hypoxia, or more than 50% lung involvement on imaging), and 5% develop critical symptoms (respiratory failure, shock, or multiorgan dysfunction). Older people are at a higher risk of developing severe symptoms. Some people continue to experience a range of effects (long COVID) for months after recovery, and damage to organs has been observed. Multi-year studies are underway to further investigate the long-term effects of the disease."
German summary: COVID-19 ist eine ansteckende Krankheit, die durch das Virus SARS-CoV-2 ausgelöst wird. Es wurde erstmals im Dezember 2019 in Wuhan, China identifiziert und hat sich schnell auf der ganzen Welt ausgebreitet. Symptome können Fieber, Husten, Kopfschmerzen, Müdigkeit, Atembeschwerden, Verlust des Geruchssinns und des Geschmacksssinns umfassen. Ein Drittel der Infizierten entwickelt keine Symptome, während die anderen Symptome von mild bis schwer haben können. Ältere Menschen sind einem höheren Risiko schwerer Symptome ausgesetzt.

For instance, in the recent Covid-era, people want to know more about Coronavirus. Cross-lingual scientific summarization can help to get concise information in various local languages. Example 1 shows a chunk of a Wikipedia article on COVID-19[1] and its German summary. The

[1] https://en.wikipedia.org/wiki/COVID-19

German summary is generated by a cross-lingual abstractive summarization model. However, such summaries might contain complex scientific information, which could be hard to understand for people with less to no expertise in scientific domains (non-experts). Therefore, these summaries need to be simplified for non-expert readers. Example 2 presents a simplified German summary, unlike Example 1, for non-expert readers. We define the simplified cross-lingual summary generation task as a use case of cross-lingual scientific summarization targetting non-expert readers and introduce it as **Cross-lingual Science Journalism**.

EXAMPLE 2

Source: "Coronavirus disease 2019 (COVID-19) is a contagious disease caused by a virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease quickly spread worldwide, resulting in the COVID-19 pandemic. The symptoms of COVID-19 are variable but often include fever, cough, headache, fatigue, breathing difficulties, loss of smell, and loss of taste. Symptoms may begin one to fourteen days after exposure to the virus. At least a third of people who are infected do not develop noticeable symptoms. Of those who develop symptoms noticeable enough to be classified as patients, most (81%) develop mild to moderate symptoms (up to mild pneumonia), while 14% develop severe symptoms (dyspnea, hypoxia, or more than 50% lung involvement on imaging), and 5% develop critical symptoms (respiratory failure, shock, or multiorgan dysfunction). Older people are at a higher risk of developing severe symptoms. Some people continue to experience a range of effects (long COVID) for months after recovery, and damage to organs has been observed. Multi-year studies are underway to further investigate the long-term effects of the disease."
Simplified German summary: Es wurde erstmals im Dezember 2019 in Wuhan, China, entdeckt und hat sich schnell weltweit verbreitet. Zu den Symptomen gehören Fieber, Husten, Kopfschmerzen, Müdigkeit, Atembeschwerden und Geruchs- und Geschmacksverlust. Einige Menschen mit der Infektion entwickeln keine Symptome, während andere leichte bis schwere Symptome entwickeln.

1.1 Task and Motivation

Cross-lingual science journalism aims to produce popular science summaries from scientific texts in a source language for non-expert readers in their local language. It focuses on simultaneously reducing linguistic complexity and length of the original text. Automating Cross-lingual Science Journalism can facilitate science journalists in their work for writing popular science summaries. It is particularly important for topics that need to be more widely discussed or understood, such as climate change, mental health, and the ethical implications of emerging technologies, to make educated decisions about the future. It can also revolutionize how scientific knowledge is shared by bridging language barriers and allowing a wider range of people to access and understand scientific knowledge.

A real-world instance of Cross-lingual Science Journalism is Spektrum der Wissenschaft (SPEKTRUM)[2]. It is a popular science magazine in Germany that is published monthly. It is the German version of Scientific American and an acclaimed bridge between local readers and the latest scientific research. Figure 1.2 presents a sample of science summaries from a published SPEKTRUM magazine. It is a special section in the magazine - SPEKTROGRAMM that contains popular summaries of the latest scientific articles written by science journalists.

Figure 1.2: A sample of science summaries in the SPEKTROGRAMM section from a published Spektrum magazine.

This research project is initiated upon SPEKTRUM's request to automate a part of their journalists' workflow related to the SPEKTOGRAMM section. The science summaries written by science journalists in the SPEKTOGRAMM section are distinct from regular scientific texts:

- These are *popular science* summaries.
- These are much more *concise* than the original articles.
- These summaries have *less complex* words and technical terms.
- These are written in German from English source articles, making them cross-lingual.

Science journalists at SPEKTRUM have been writing these popular science summaries in German for decades. To write these summaries, SPEKTRUM's journalists first read English

[2] https://www.spektrum.de/

scientific articles and then summarize them into popular science summaries in German that are comprehensible by local non-expert readers. However, it is closer to the cross-lingual summarization task with additional simplification.

Automating a new task requires a sufficient amount of data for an extensive investigation of the data properties and behavior of models applied to that data. Initially, we experience some challenges in obtaining sufficient data size and getting legal consent for this task. SPEKTRUM releases a part of their data that only contains the German summaries, and their original articles need to be found and scrapped from online sources such as Springer, Nature, and so on. Moreover, SPEKTRUM makes its data release exclusively for us and could not agree to share it publically with the research community. The size of the raw data obtained from SPEKTRUM is 20K, and it is a private source, so it cannot be released to the research community. To overcome this and increase the amount of data for a thorough investigation, we decide to collect a similar dataset from the public domain - Wikipedia Science Portal (WIKIPEDIA). It is a dedicated portal in Wikipedia, which contains science articles in different languages. Based on the literature review, we find that Wikipedia has been widely used in the summarization community for data collection. Moreover, Wikipedia allows researchers to collect their data and share the resources with the research community. All these attributes make WIKIPEDIA a proper fit for our project.

1.2 Research Objectives

In this thesis, we introduce a new task of Cross-lingual Science Journalism as a use case of cross-lingual scientific summarization. Based upon the properties of popular science summaries and the nature of the task, we aim to investigate the task first by cross-lingual summarization models and in the latter chapters by fusing simplification and cross-lingual summarization.

We first review the related literature in science journalism, summarization and simplification. After a thorough literature review, we find the following research gaps. Due to its complex nature, the science journalism task has not gained much attention and has yet to be thoroughly explored in the NLP community. Moreover, Cross-lingual Science Journalism has not been investigated before, and there are no available resources for Cross-lingual Science Journalism. Furthermore, plenty of studies are present in the simplification and summarization tasks, still, scientific simplification and cross-lingual scientific summarization have not gained much attention. Based on these findings, we define our research objectives as follows:

- Collection of cross-lingual scientific resources that can be used for cross-lingual summarization and science journalism.

- Analysis of collected resources based on statistical and readability features.
- Evaluation of existing summarization models on collected resources.
- Development and evaluation of new models to provide a basis for Cross-lingual Science Journalism.

1.3 Research Questions and Thesis Contributions

We investigate and analyze the following research questions based on our objectives. We list each research question and our contributions as the answer to them.

1. **How can we obtain SPEKTRUM source articles and similar data from WIKIPEDIA and measure similarities and differences in both datasets?**

   To answer this research question, we first inspect and analyze SPEKTRUM raw data manually. Based on this analysis, we devise a structured procedure to collect and verify source articles. We also perform manual filtering and cleaning to ensure the high quality of the SPEKTRUM dataset. For WIKIPEDIA, we develop a methodology to collect data simultaneously from the German and English Wikipedia Science Portals. We also perform manual verification on a subset of the WIKIPEDIA dataset. The WIKIPEDIA dataset has two parts - one for monolingual summarization and one for cross-lingual summarization. Then we perform a thorough analysis based on different statistical and linguistic features to investigate the properties of our datasets.

2. **Can we train Cross-lingual Science Journalism models with WIKIPEDIA science articles and evaluate them on science summaries from SPEKTRUM?**

   To investigate this research question, we find suitable existing summarization models and training strategies for our task. We evaluate our datasets for two tasks - monolingual summarization and cross-lingual summarization. In monolingual summarization, we apply several extractive and abstractive summarization models on the WIKIPEDIA monolingual dataset. For cross-lingual summarization, we train different pipelines and abstractive models on the WIKIPEDIA cross-lingual dataset. We then evaluate these trained models on the WIKIPEDIA test set and the Spketrum dataset. Furthermore, we perform human evaluation and manual analysis to find the limitations of existing models.

3. **Can a combination of summarization and simplification models perform better than the existing summarization models for Cross-lingual Science Journalism?**

From the analysis of existing models, we find that the existing summarization models struggle with the scientific documents length. Moreover, our task demands simplification. To address these, we combine an extractive summarizer, a simplification model and a cross-lingual abstractive summarizer in a pipeline - SELECT, SIMPLIFY and REWRITE (SSR). The SSR model first addresses long scientific documents with SELECT, then simplifies the content with SIMPLIFY and generates a cross-lingual summary with REWRITE. We compare the performance of SSR with several existing models. We also investigate the performance of ablated models. The SSR model performs better than the strong baselines with 99% confidence, further suggested by human judgment and readability analysis. We further conduct an analysis to find the readability of output summaries.

4. **Can joint training of cross-lingual summarization and simplification help to improve the quality of generated summaries for Cross-lingual Science Journalism?**

   From the good performance of SSR and its analysis, we find that the simplification model improves the SSR's performance. Moreover, the simplification model is similar to cross-lingual abstractive summarization models, so have a possibility for joint training. To further explore it, we develop an end-to-end model - SIMCSUM based on joint learning of simplification and cross-lingual summarization to improve the quality of output summaries. We construct a synthetic simplification dataset from our WIKIPEDIA dataset and train SIMCSUM jointly on simplification and summarization data. We compare the performance of SIMCSUM against several baselines. The SIMCSUM model outperforms the baseline models with 99% confidence, further indicated by human evaluation. We then analyze the quality of output summaries with readability and error analysis.

1.4 Thesis Structure

The first part of this thesis contains two more chapters. In Chapter 2, we discuss the background concepts of summarization, including summarization categories, the basics of neural network-based summarization models and evaluation strategies. Chapter 3 reviews the existing work on science journalism, summarization and simplification. This part helps us to find the research gap and to establish well-defined steps for data collection and well-structured experimental designs.

The second part includes Chapter 4. In this chapter, we introduce two newly curated datasets, a systematic process for data collection and its verification for each dataset. We thoroughly explore various properties of our datasets with statistical and linguistic analysis.

Then we perform an empirical evaluation of these datasets for the summarization task. Based on these findings, we establish our directions for developing models.

The third part contains two chapters - 5 and 6, in which we introduce and evaluate two models for Cross-lingual Science Journalism. We assess these models against several state-of-the-art baselines and with different evaluation metrics. We also conduct an in-depth investigation into the linguistic properties of generated summaries and an error analysis. Our models demonstrate better performance from the baselines. Finally, in the last part of the thesis - Chapter 7, we summarize the contributions and conclusions of this thesis and discuss some future directions.

1.5 Publications and Resources

Most of the work presented in this thesis is an extension of the published work by the author of this thesis. Chapter 4 is an extension of Fatima and Strube (2021), and Chapters 5 and 6 are extensions of Fatima and Strube (2023); Fatima et al. (2023). The generated resources during this work can be accessed as follows:

1. WIKIPEDIA Monolingual Dataset - https://doi.org/10.11588/data/GGWW2J
2. WIKIPEDIA Cross-lingual Dataset - https://doi.org/10.11588/data/PK3FWS
3. SPEKTRUM Dataset - Contact person: Michael Strube - Michael.Strube@h-its.org
4. SSR Model - https://doi.org/10.11588/data/5VHCBV
5. SIMCSUM Model - https://doi.org/10.11588/data/U9FWOU

Chapter 2

Background

> "It is the theory which decides what can be observed."
>
> Albert Einstein

In this chapter, we focus on the relevant background concepts of summarization. We present a brief overview of different dimensions in the summarization field, the basics of deep learning summarization models and evaluation strategies.

2.1 Summarization Classification

Text summarization can be classified in different ways depending on the nature and criteria of the given task. Here we classify it into five categories that can be overlapped.

2.1.1 Input Size

Input size refers to how many input documents are given to generate a summary. Single document summarization accepts a single document to generate a summary (Kim et al., 2016a; Li et al., 2017; See et al., 2017; Nallapati et al., 2017; Cachola et al., 2020; Takeshita et al., 2022). In contrast, multi-document requires more than one document to generate a summary (Zopf et al., 2016; Hättasch et al., 2020; Gholipour Ghalandari et al., 2020).

2.1.2 Input Domain

Different text genres demand domain-specific summaries. Therefore, the input domain heavily influences the summarization model design. As our focus is the scientific domain and most

summarization models have been developed with the news domain (§ 3.2.1), we take these two domains as an example to highlight the differences in their structure. Figure 2.1 presents a visual demonstration of how important information is structured in a news text (right) and in a scientific text (left). In the news text, the most important information is placed at the top of the inverted pyramid, known as "lede" or "lead" (Grusky et al., 2018). In contrast, the scientific text contains multiple sections and the salient information is scattered at different points in those sections.

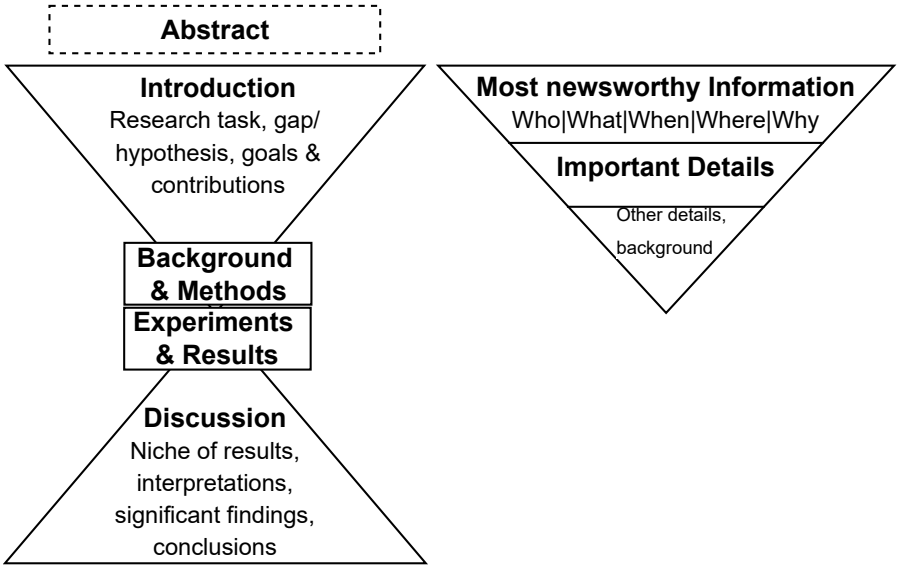

Figure 2.1: Scientific text discourse (left) vs. news text discourse (right).

The summarization field has an ample amount of domains, *e.g.*, news (See et al., 2017; Ouyang et al., 2019; Nallapati et al., 2017; Gehrmann et al., 2018), blogs (Hu et al., 2007; Joshi et al., 2019), meetings (Zhong et al., 2021; Li et al., 2019), opinions or reviews (Bražinskas et al., 2019, 2020), scientific texts (Kim et al., 2016a; Nikolov et al., 2018; Cohan et al., 2018), legal documents (Sharma et al., 2019; Casola and Lavelli, 2022) and medical documents (Mishra et al., 2014; Kieuvongngam et al., 2020).

2.1.3 Output Size

Output size refers to the generated summary length, which can be a headline (Vadapalli et al., 2018), highlights (Hermann et al., 2015), a sentence (Narayan et al., 2018a), TLDR (too long; didn't read) (Cachola et al., 2020; Takeshita et al., 2022) or a summary(Kim et al., 2016a; Li et al., 2017; See et al., 2017; Nallapati et al., 2017; Gehrmann et al., 2018; Cohan et al., 2018; Nikolov et al., 2018).

2.1.4 Output Language

It refers to the relationship between source documents and target summaries language. It is called monolingual summarization if both source and target are in the same language (Cohan et al., 2018; Nikolov et al., 2018; Rush et al., 2015; Grusky et al., 2018). In cross-lingual summarization, source and target languages are different (Ouyang et al., 2019; Zhu et al., 2019, 2020; Ladhak et al., 2020; Takeshita et al., 2022).

2.1.5 Model Type

The model type defines the nature of the design, algorithm and/or architecture and training approach used for the summarization system. Based on the nature of the design, we can classify it as generic (Haghighi and Vanderwende, 2009) or query-based (Sun et al., 2016) summarization.

Classifying it by output type, it can be extractive or abstractive. Extractive summarization relies on selecting and ranking sentences from the source and combining them as a summary (Luhn, 1958; Mihalcea and Tarau, 2004; Kenton and Toutanova, 2019; Zhang et al., 2020b). Whereas abstractive summarization models select the salient content from the source, however, generate sentences different from the source text with paraphrasing (Gehring et al., 2017; Liu et al., 2020; Lewis et al., 2019). A third type is a hybrid which combines extractive and abstractive algorithms (Kipf and Welling, 2016; See et al., 2017).

It can also be classified as the algorithm type such as sentence centrality (Erkan and Radev, 2004; Radev et al., 2004), graph-based (Mihalcea and Tarau, 2004; Dong et al., 2021), semantic-based (Steinberger et al., 2004), neural networks-based (Liu et al., 2020; Lewis et al., 2019; Zhang et al., 2020b), *etc*. The algorithms can be used for both extractive and abstractive summarization.

Based on the training approach, summarization models can be classified as supervised (See et al., 2017; Kenton and Toutanova, 2019), unsupervised (Mihalcea and Tarau, 2004; Gillick et al., 2009; Dong et al., 2021), reinforcement (Narayan et al., 2018b; Scialom et al., 2019) and transfer learning (zero-shot, few-shots, multi-task learning) (Zhu et al., 2019; Takase and Okazaki, 2020).

2.2 Model Architecture

In this section, we provide a brief introduction to deep learning-based summarization architectures.

2.2.1 Sequence-to-Sequence

A sequence-to-sequence (S2S) model accepts a text sequence and generates another text sequence. The S2S model consists of two neural networks - an encoder and a decoder. The encoder receives the input text sequentially, captures the contextual representations of the given text and sends it to the decoder. The decoder generates the output text sequentially based on the given representations. Figure 2.2 presents a black box representation of the S2S model.

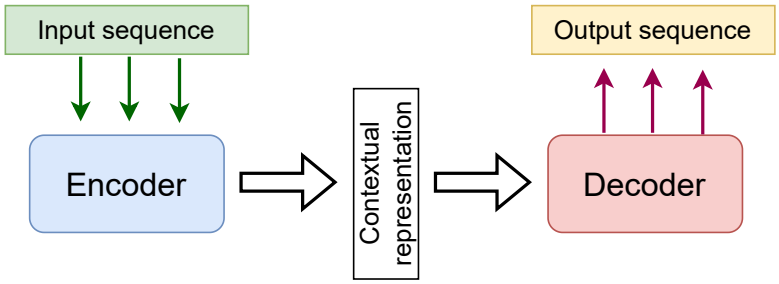

Figure 2.2: A black box illustration of an S2S summarization model.

Mathematically, the encoder takes the input sequence of tokens x of length n and converts it into continuous (contextual) representations c received by the decoder for generating the target summary y of length m. The decoder is a conditional language model $p(y|x)$. The S2S model can be built with a Recurrent neural network (RNN) or a transformer (Vaswani et al., 2017). In the next sections, we first discuss basic building blocks with RNN and later discuss transformer architecture.

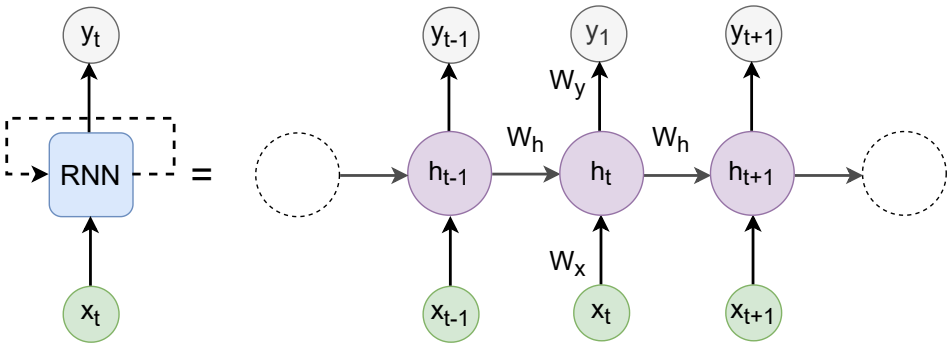

Figure 2.3: An illustration of a rolled-out RNN.

2.2.1.1 Recurrent Neural Network

Recurrent neural network (RNN) is a general term for neural networks with multiple self-copies to process the sequence of data at different time steps. Figure 2.3 illustrates a rolled-out RNN where x is the input sequence, y is the output sequence, h represents a hidden state, and W

represents the weight vectors. The types of nodes in an RNN can be an RNN (Kawakami, 2008), LSTM (Goodfellow et al., 2013) or GRU (Cho et al., 2014).

2.2.1.1.1 Encoder The encoder is responsible for processing the input sequence x. Encoders can be unidirectional or bi-directional (forward and backward passes) to handle the context present in the textual sequence. Formally, an encoder hidden state is represented as:

$$h_t = \begin{cases} 0, & t = 0 \\ \phi(W_x \cdot x_t, W_h \cdot h_{t-1}), & otherwise \end{cases} \tag{2.1}$$

where x_t is the input token at time step t with its weight W_x, h_{t-1} refers to previous hidden state, W_h is hidden state's weight and ϕ is a nonlinear function with affine transformations.

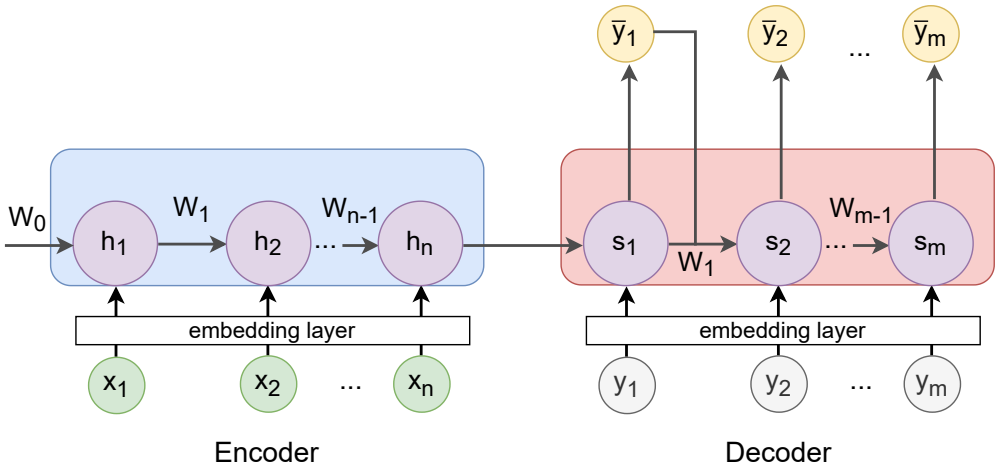

Figure 2.4: An illustration of the RNN-based S2S model.

2.2.1.1.2 Word Embeddings Figure 2.4 presents a detailed view of the models' architecture where $\bar{y}$ denotes the generated output. An embedding layer converts the word sequence to numeric vectors. Word embeddings map each word to one vector in a predefined vector space. The vector values can be learned with different techniques, *e.g.*, word2vec (Chelba et al., 2013), glove (Pennington et al., 2014), *etc*.

2.2.1.1.3 Context Vector The encoder and decoder are connected with contextual representations in the S2S model. The contextual representation or context vector is computed as:

$$c = \phi(h_1, \cdots, h_n), \tag{2.2}$$

where h_t is the hidden state at time step t, n is the length of input sequence and ϕ is a function.

2.2.1.1.4 Decoder The decoder is trained to predict the output with a given context vector. Mostly, decoders are unidirectional as at step t, y_{t+1} can not be estimated. The decoder predicts the word y_t with the context vector c and all previously predicted tokens $y_{<t}$. The conditional probability of the decoder is:

$$p(y) = \prod_{t=1}^{m} p(y_t|y_1, \cdots, y_{t-1}, c) \tag{2.3}$$

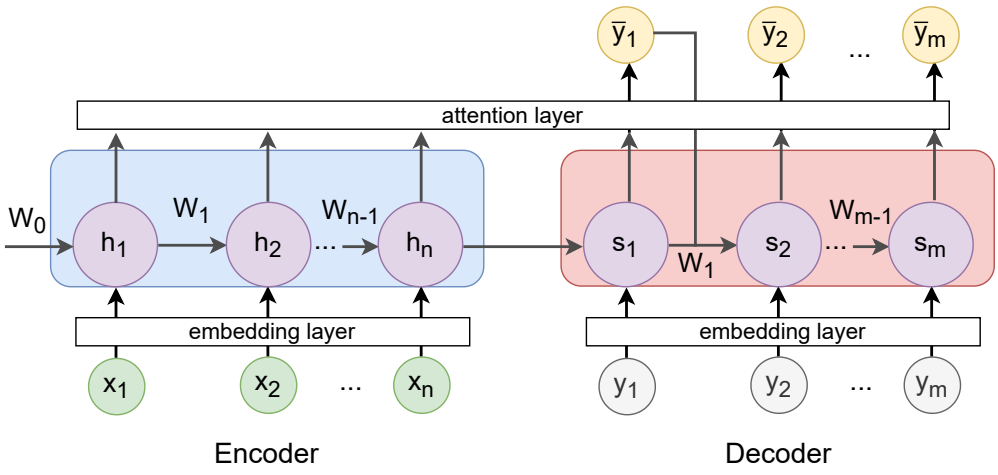

Figure 2.5: The S2S model with attention.

2.2.1.1.5 Attention The RNN models struggle to keep all hidden states, and usually, the last hidden state of the encoder is used as the context vector for decoding. Due to this, these models cannot retain long-term dependencies with a long input sequence. An attention mechanism is introduced by Bahdanau et al. (2015) and Luong et al. (2015) to solve this problem. Attention allows the decoder to focus on different parts of the input sequence at every decoding step (accessing multiple hidden states). It is similar to how the human brain focuses on important information and ignores the unimportant part. Figure 2.5 illustrates the attention layer in the S2S model. The attention mechanism works on the idea of calculating and distributing the attention weight, and the focus is placed on the important token by increasing its weight. Here we discuss the key concepts of Bahdanau et al. (2015) attention mechanism. With the attention layer, the decoder probability is redefined as:

$$p(y_t|y_1, \cdots, y_{t-1}, x) = \phi(y_{t-1}, s_t, c_t) \tag{2.4}$$

where s_t is a decoder hidden state for time t, computed by $s_t = \phi(s_{t-1}, y_{t-1}, c_t)$. The context vector is computed as:

$$c_t = \sum_{j=1}^{n} \alpha_{tj} \cdot h_j \tag{2.5}$$

where h_j is the encoder hidden state and n is the length of the input sequence. The weight α_{tj} is computed as:

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{n} exp(e_{tk})} \tag{2.6}$$

where $e_{tj} = a(s_{t-1}, h_j)$ and a is an alignment model to compute scores of how well the inputs around position j and the output at position t match.

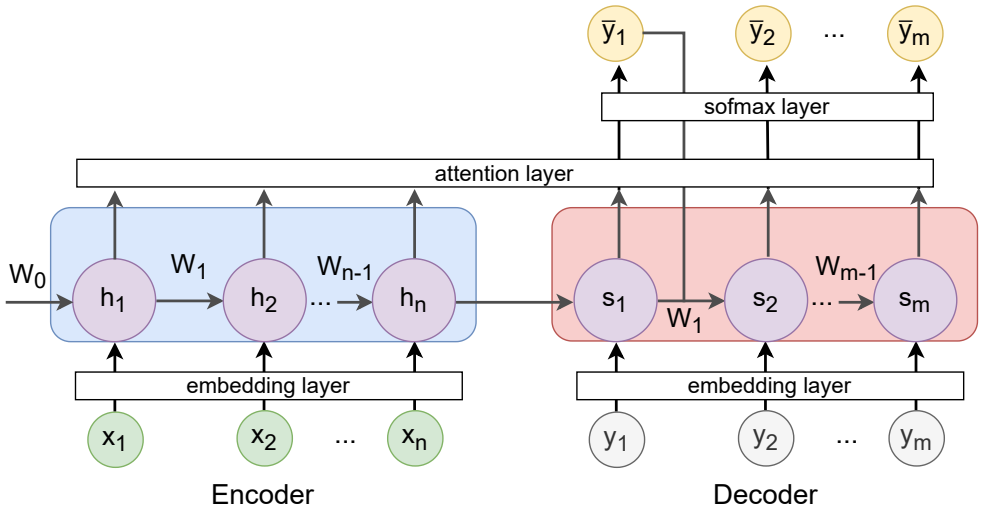

Figure 2.6: The S2S model with attention and a softmax layer.

2.2.1.1.6 Softmax As previously discussed, the embedding layer converts the text into numeric vectors. So, decoder outputs are basically numeric vectors. The softmax layer (normalized exponential function) converts a numeric vector into a probability distribution over a vocabulary. A word-based vocabulary contains all unique words present in a set of inputs and outputs. Usually, the vocabulary size in the S2S models for summarization is huge and calculating probability distributions at each decoding step ($\bar{y}$) is computationally expensive. So, Top-K sampling and beam search are introduced for decoding to reduce the candidate sample space.

2.2.1.2 Transformer

Vaswani et al. (2017) has introduced transformer architecture that also follows the S2S model, however, it has stacked encoders and decoders with self-attention and encoders, decoders layers are fully connected. Figure 2.7 shows a transformer S2S model with 6 stacked encoders and 6 stacked decoders. In contrast to RNN models, there is no concept of time steps because input and output sequences are encoded positionally to preserve the sequence information.

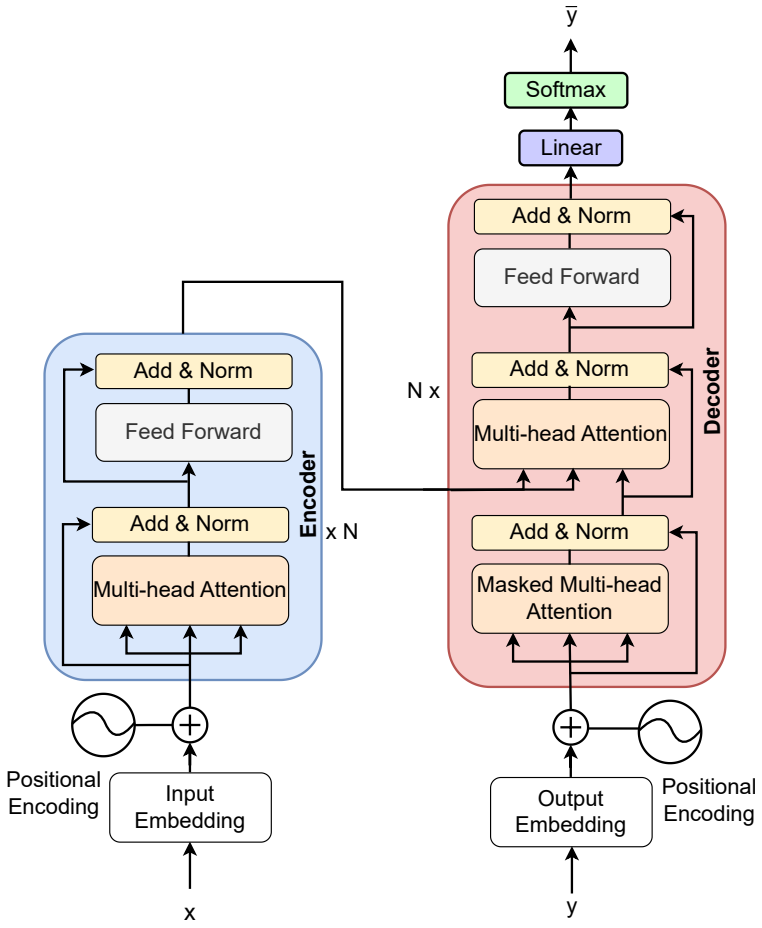

Figure 2.7: A transformer-based S2S model where N represents the stack (N=6).

2.2.1.2.1 Positional Encodings Positional encoding describes the location or position of an entity in a sequence so that each position is assigned a unique representation. Vaswani et al. (2017) use sine and cosine functions for positional encodings.

$$\begin{aligned} PE_{(pos,2i)} &= sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{2.7}$$

where pos is the position, i represents the dimension, and each dimension's positional encoding is represented with a sinusoid.

2.2.1.2.2 Feed forward network Each layer of the encoder and decoder contains fully connected feed-forward networks (FFNs). The main difference between RNN and FFN is directionality, an RNN can be bi-directional, while an FFN is always unidirectional. The FFNs are connected position-wise. Two linear transformations with a ReLU activation in between the

layers are performed to connect them position-wise.

$$FFN(x) = max(0, x \cdot W_1 + b_1) \cdot W_2 + b2 \tag{2.8}$$

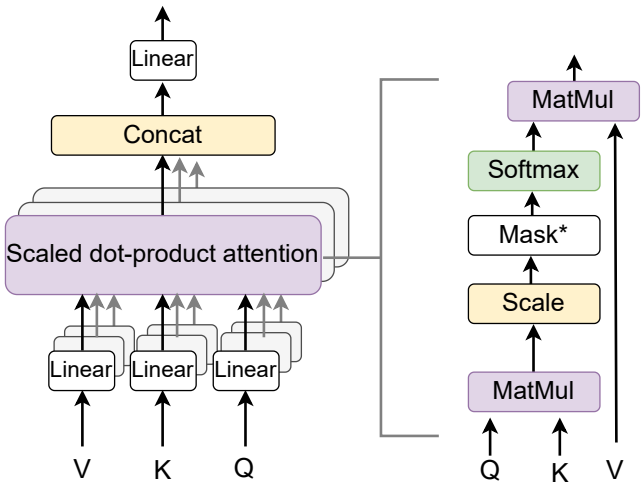

Figure 2.8: Multi-head attention in the transformer.

2.2.1.2.3 Self-Attention In transformer architecture, attention is calculated as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. Figure 2.8 illustrates self-attention and multi-head attention operations, where "*" denotes the optional operation. The self-attention (scaled dot-product) is computed on a set of keys, values, and queries generated from the same sequence, where the key is a label of a word that is used to distinguish between different words. The query represents an active request for specific information, checks all available keys, and selects the best-matched one. A value is information that a word contains.

$$Attention(Q, K, V) = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{2.9}$$

where Q is a query, K^T is transposed K (key) and V is the value, "$\cdot$" represents the dot products and $Softmax$ is the scaling function over the result vector.

2.2.1.2.4 Multi-head Attention Each encoder and decoder layer contains multiple attention heads to jointly attend to the given sequence from different representation spaces at different positions. All parallel self-attentions are concatenated to generate multi-head attention scaled with a weight matrix W.

$$MultiHead(Q, K, V) = Concat(Attention_1, \cdots, Attention_h) \cdot W^O \tag{2.10}$$

2.2.1.2.5 Add and Norm This layer refers to two functions - add is a residual connection for adding the input of each layer to the output, and the norm is layer normalization.

2.2.1.2.6 Variations Based on transformer architecture, a lot of pre-trained language models has been developed in the past few years. Some examples of only encoder models include BERT, ROBERTa and ELECTRA, some examples of only decoder models are GPT, GPT-2 and CTRL, and some instances of encoder-decoder (S2S) models are BART, T5 and Marian.

2.2.2 Multitask Learning

Multitask learning (MTL) is introduced by Caruana (1998) as an inductive transfer learning. The core idea of multitask learning comes from how humans learn new tasks by transferring the knowledge of a learned task. In conventional transfer learning, the knowledge or model of one task is applied to another with the hypothesis the source task help to solve the target task. However, in multitask learning, multiple tasks are trained simultaneously to help each other to learn generalized representations.

2.2.2.1 Definition

Multitask learning is defined mathematically as follows: given k learning tasks $\{\mathcal{T}_i\}_{i=1}^{k}$ where all tasks are related but not identical, multitask learning aims to improve the learning of a model for $\mathcal{T}_i$ by applying the knowledge present in the k tasks (Zhang and Yang, 2018). Multitask Learning relies on inductive transfer, more specifically on inductive bias provided by the auxiliary tasks, which causes the model to learn hypotheses that generalize over the given tasks.

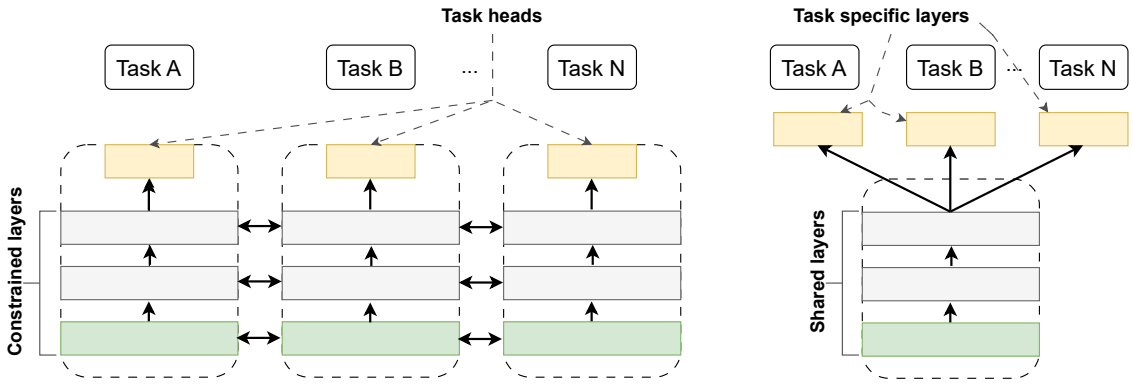

Figure 2.9: Soft (left) vs. hard (right) parameter sharing.

2.2.2.2 Types

There are two types of multitask learning based on parameter sharing: hard parameter sharing and soft parameter sharing (Ruder, 2017). In hard parameter sharing, the hidden layers of the model are shared between all tasks, and the output layers are task-specific. Hard parameter sharing is useful when the tasks are related in their nature; it improves the positive transfer and reduces the risk of overfitting. In soft parameter sharing, each task has its own model and parameters. These models are trained together with some constraints on layers for regularizing the parameters among the tasks. Soft parameter sharing is useful when tasks are not highly related. A lot of synthetic cross-lingual summarization models have been developed with multitask learning (Zhu et al., 2019; Cao et al., 2020b; Takase and Okazaki, 2020; Bai et al., 2022).

2.3 Summarization Evaluation

There are plenty of evaluation metrics and linguistic features considered over the period of time for summarization. Here we discuss the most common and well-established metrics and features in the summarization community.

2.3.1 Automatic Metrics

2.3.1.1 Rouge

ROUGE is a recall-based standard summarization metric for evaluating generated summaries (Lin, 2004). It has been proven to be an effective metric for measuring the qualities of generated summaries and correlates well with human judgments (Sun et al., 2016). It computes the overlap between system-generated and reference summaries. The ROUGE score has various types the basis of overlap:

- ROUGE N-gram measures the overlap of n-grams between a system summary and a reference summary, where the value of n can be in the range 1-4.
- ROUGE LCS computes the longest common subsequences (LCS) overlap between a system summary and a reference summary.
- ROUGE Skip-grams computes the overlap ratio of skip-uni/bi-grams between a system summary and a reference summary.

These scores are reported with precision, recall and F-score. Precision refers to the proportion of words/sentences present in the system summary that is actually present in the reference

summary.

$$Precision = \frac{W_{ref} \cap W_{sys}}{W_{sys}} \tag{2.11}$$

Recall refers to the proportion of words/sentences in the reference summary captured by the system summary.

$$Recall = \frac{W_{ref} \cap W_{sys}}{W_{ref}} \tag{2.12}$$

where W denotes words/sentences, ref is the reference summary, sys is the system summary and $W_{ref} \cap W_{sys}$ refers to total number of captured n-grams.

F-score denotes the harmonic mean of precision and recall.

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.13}$$

For instance, we calculate ROUGE-1 for a system summary. Example 3 shows the tokenized reference and system sentences. We compute the total number of tokens in the reference summary (n_ref), the total number of tokens in the system summary (n_sys), and the total number of tokens that have been captured by the system summary (n_cap). From these values, we compute the precision, recall and F-score for the system summary. The F-score of ROUGE-1, ROUGE-2 and ROUGE-L is the most commonly used in the recent summarization studies (Cohan et al., 2018; Ouyang et al., 2019; Takeshita et al., 2022).

EXAMPLE 3

Reference Tokens: [“a”, “cat”, “sits”, “on”, “the”, “mat”] (n_ref=6)
System Tokens: [“cat”, “perches”, “on”, “the”, “mat”] (n_sys=5)
Captured Tokens: [“cat”, “on”, “the”, “mat”] (n_cap=4)
ROUGE-1 Recall = n_cap/n_ref = 4/6 = 0.67
ROUGE-1 Precision = n_cap/n_sys = 4/5 = 0.80
ROUGE-1 Precision = $2 \times \frac{0.80 \times 0.67}{0.80 + 0.67}$ = 0.72

2.3.1.2 Bert-Score

BERT-SCORE (Zhang et al., 2020b) is a recent metric for summarization, translation and simplification tasks as an alternative to n-gram-based metrics. The n-gram metrics depend on the syntactic overlap (unigrams, bigrams, *etc*.) between the system and reference summaries. However, these metrics cannot capture the semantic similarities of the similar words in the system and reference summaries. BERT-SCORE calculates the semantic similarities with pairwise cosine similarity. It also captures faraway dependencies using contextual embeddings.

Precision and Recall of BERT-SCORE are calculated as follows:

$$P_{Bert} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \tag{2.14}$$

$$R_{Bert} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{2.15}$$

where x is the reference summary and $\hat{x}$ is the system summary. Figure 2.10 shows the maximum similarity score matrix between the reference and system summary ("a" is removed for the sake of simplicity) of Example 3.

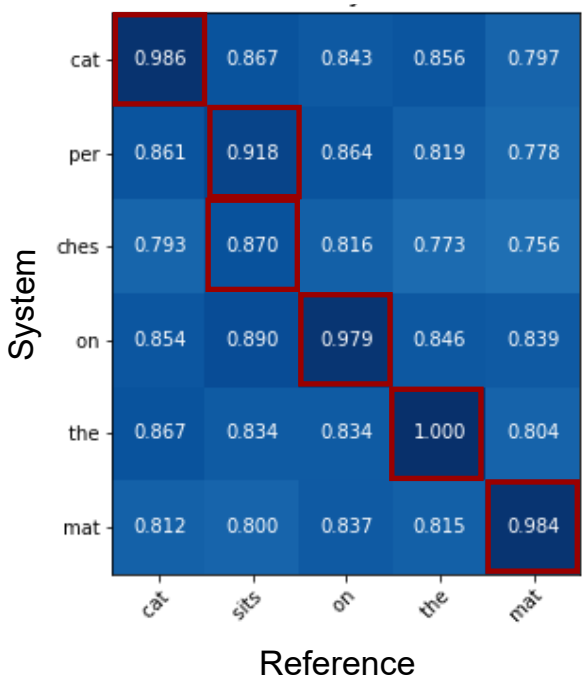

Figure 2.10: An example of similarity scores between a reference and a system summary.

2.3.2 Human Evaluation

In general, qualitative evaluation (readability, coherence and fluency) of generated summaries is trickier than the quantitative analysis. It requires expensive human annotations, specifically in the scientific domain with the cross-lingual setup. Some of the important linguistic properties are:

- **Relevance** - A summary that delivers adequate information about the original text. Relevance determines the content relevancy of the summary.

- **Fluency** - The words and phrases fit together within a sentence, and so do the sentences. Fluency determines the structural and grammatical properties of a summary.
- **Readability** - It refers to sentences' lexical and syntactic simplicity. A readable summary should have minimal use of complex words or phrases and sentence structure.
- **Overall Ranking** - How coherent is the system summary?

2.4 Summary

This chapter presents an overview of summarization, a brief introduction to neural network architectures and common practices for quantitative and qualitative analysis of output summaries.

Chapter 3

Literature Review

> "Research is to see what everybody else has seen and to think what nobody else has thought."
>
> Albert Szent-Gyorgyi

In this chapter, we discuss all the relevant scientific contributions that have been made in the past. We divide this chapter into three sections - science journalism, summarization and text simplification. We provide the details of existing datasets and state-of-the-art (SOTA) models in these fields.

3.1 Science Journalism

3.1.1 Writing Quality

Louis and Nenkova (2013a) investigate the writing quality of science journalism articles. They develop an annotated dataset from New York Times articles between 1999 and 2007. These articles are divided into three categories: (i) great, (ii) very good and (iii) typical. The great articles are selected from the "Best American Science Writing" annual anthologies. The stories are selected by prominent science journalists and editors for best writings. There are a total of 52 articles in the great category coming from various topics. The very good category articles are selected from the great category sub-corpus. The articles in the typical category are collected that were published around the same time, however, not selected for the anthologies. The authors filter the articles for science journalism containing at least 1000 words and science tags in meta-data. The very good category contains 2243 articles and the typical category contains 11223 articles. Then the authors annotate the sentences of a subset of articles for gen-

eral and specific content to find the relationship between the writing quality of those articles. Louis and Nenkova (2013b) then extend their work for topic classification of the very good category. They implement several features to capture the textual properties such as surprising, visual and emotional content. They also investigate general features of discourse organization and sentence structure. We find that this dataset is suitable for classification tasks, however, it cannot be used for cross-lingual science journalism.

Barel-Ben David et al. (2020) investigate a different dimension of science journalism, how readers react to science news written by scientists and science journalists. The authors use two online resources - Mako and Ynet, from 2015 to 2018. They collect a dataset of 150 science news written by 50 early career scientists trained for publishing popular science stories in the Davidson Institute of Science Education reporters' program. Each story is paired with another science story written by the website's science journalists and published on the same website. For estimating users' behavior, the authors collect data from the websites for the number of clicks, likes, comments and average time spent on a page. Based on these scores, the authors perform statistical analysis to find readers' likeness to those stories. We find that this dataset is also unsuitable for the generation of cross-lingual science summaries. We find that all these studies investigate the quality of writing in the monolingual science journalism domain and are unsuitable for automating cross-lingual science journalism.

3.1.2 Summary Generation

In this section, we report the papers that created models for science journalism along with dataset creation. Vadapalli et al. (2018) collect a parallel corpus of 87K pairs of research paper titles, abstracts and corresponding blog titles for title generation. They extract their data from two science news websites - Science Daily and Phys with manual rules for extraction. The authors develop a two-stage pipeline system. The first stage is a TF-IDF-based heuristic function to extract relevant keywords from the title and abstract. The output of this stage is given to the Pointer Generator Network (PGN) (See et al., 2017) model to generate the blog title.

Zaman et al. (2020) collect a customized dataset containing simplified summaries from the Eureka Alert science news website. Eureka alert is a science website where researchers, bloggers and experts publish scientific articles in a summarised and easy-to-read version. They harvest a parallel corpus of summaries and online scientific journal articles. They first collect 227,590 simplified summaries from the website. Each summary is linked to its source article via DOI. With DOIs, PDFs of source articles are obtained and then parsed, however, this resulted in significant cohesion errors. So the authors decide to filter the summaries for only

those articles with an XML version. It greatly shrinks the size of the dataset, yielding 5,204 articles in total, mainly from PLOS-ONE, Nature Communication, and Scientific Reports. Each instance in the dataset consists of six attributes: (i) "Eureka Title Simplified", (ii) "Eureka Text Simplified", (iii) "Paper Title", (iv) "Full Paper XML", (v) "Paper Journal" and (vi) "Paper DOI". Zaman et al. (2020) develop an extension of the PGN (See et al., 2017) for generating monolingual science simplified summaries. The PGN model is an encoder-decoder S2S network. The encoder consists of a single layer of bi-directional LSTM and the decoder is a single layer of unidirectional LSTM with attention and coverage mechanism. The decoder side also contains a soft switch to decide whether to generate a new token or copying (point) it from the input. Zaman et al. (2020) modifies the PGN model for implementing a loss function for joint simplification and summarization. The authors also introduce an evaluation metric based on the F-SCORE of ROUGE-1 and SARI, named as "Combined Summarisation and Simplification score" (CSS). They select two baselines. The first baseline is a Neural Text Simplification (NTS) model consisting of an encoder-decoder S2S model with beam search. The encoder has two LSTM layers and 500 hidden units and the decoder has two layers of LSTM with global attention. The second baseline is an Abstractive Text Summarisation (ATS) model that consists of two layers of LSTM encoders with 1000 hidden units and 500 dimensions of word embedding. The modified PGN network performs better than the baselines.

Dangovski et al. (2021) introduce monolingual science journalism as a downstream task of abstractive summarization and story generation. They create two versions of a previously created dataset (Dangovski et al., 2019) containing 60K press releases from Science Daily paired with scientific papers. Their first version is for long summarization and it contains pairs of full-text scientific documents and corresponding press releases as summaries. The second version consists of the first 400 words from papers paired with corresponding short highlights of press releases. Dangovski et al. (2021) apply several existing baselines to investigate their performance on the Science Daily dataset. They use BertSumAbs, SciBertSumAbs, CNN-based S2S and story generation models. BertSumAbs (Liu and Lapata, 2019) is a pre-trained model consisting of a BERT-based (Kenton and Toutanova, 2019) encoder and a Transformer-based (Vaswani et al., 2017) decoder. SciBertSumAbs is a variation of BertSumAbs by replacing the BERT encoder with SciBERT (Beltagy et al., 2019) that is fine-tuned on scientific papers. The CNN-based S2S architecture is a vanilla convolutional model from the FairSeq library (Ott et al., 2019). For story generation models, the authors select two neural story generation models (Fan et al., 2018, 2019). The authors find that scientific pre-trained and story-generation models usually overfit the data and need further investigation.

3.1.3 Observations

Although, Vadapalli et al. (2018) collected their datasets from science news websites to automate science journalism. We find that this work is limited to title generation only, making it unsuitable for generating science summaries. Both Zaman et al. (2020) and Dangovski et al. (2021) present a general approach to investigating science journalism as a downstream task of monolingual abstractive summarization. None of these papers describes their data collection process in detail. Moreover, there is no manual inspection of at least a subset to find the anomalies of collected data from online sources. Furthermore, these studies investigate the task with existing models with minor modifications. To conclude, we find no similarity between these studies and our work except for the overlap of abstractive summarization. However, these studies have been investigated on monolingual datasets, while our focus domain is cross-lingual science journalism.

3.2 Summarization

Text summarization is a well-established broad field with manifold directions. In the following section, we discuss the scientific advent in summarization by dividing it into three text domains and then further dividing each domain into monolingual and cross-lingual datasets and models.

3.2.1 News Articles

Here we only focus on single-document summarization.

3.2.1.1 Monolingual

Sandhaus (2008) create a dataset of 8 million articles from the New York Times (NYT) website and newsroom. They extract the articles and their metadata from 1987 to 2007 provided by the New York Times Newsroom. Each article is paired with the journalist's written summary (bullet points). Hermann et al. (2015) collect a corpus from CNN and Daily Mail (DM) online news articles. They collect 220K articles from the Daily Mail and 93K from the CNN websites. The news articles are paired with highlights (bullet points) as target summaries. Napoles et al. (2012) create a large and diversified dataset of 10 million documents - Gigaword. They collect this data from the Associated Press, New York Times Newswire Service, and Washington Post Newswire Service. However, this dataset had not been annotated for the summarization task. Later Rush et al. (2015) annotate this dataset for summarization. Grusky et al. (2018) collect newsroom data from online resources. They scrap 1,321,995 newswire

articles and summaries from HTML metadata. Narayan et al. (2018a) construct a dataset based on BBC articles paired with single-sentence summaries for the extreme summarization task. The total size of their dataset is 226,711 and it contains various topic news from 2010 to 2017. These datasets facilitate the summarization community to grow rapidly with neural-based extractive and abstractive models (Liu et al., 2015; Cao et al., 2016; Nallapati et al., 2016; Tan et al., 2017; Li et al., 2017; See et al., 2017; Nallapati et al., 2017; Gehrmann et al., 2018) and also multitask learning summarization models which we discuss here in detail.

(Guo et al., 2018) introduce a multitask summarization model paired with question and entailment generation as auxiliary tasks. They modify the PGN (See et al., 2017) model by adding auxiliary task encoders and decoders. The auxiliary encoders and decoders share the high-level layers and attention layers with the main encoder and decoder. The model shares parameters with soft-sharing so the tasks are coupled loosely in the shared contextual representation. The model is pre-trained for summarization and during training, the model is more focused on the primary task (summarization) while learning one auxiliary task at a time. The model is evaluated on CNN-DM and Gigaword datasets. (Liu et al., 2022) also develop a multitask summarization model to improve the quality of abstractive summarization. The model consists of one abstractive fine-tuned transformer, whether BART (Lewis et al., 2019) or PEGASUS (Zhang et al., 2020a) and one non-deterministic method for assigning probabilities to candidate outputs during evaluation by contrastive learning. The model is evaluated on CNN-DM, XSUM and NYT datasets.

3.2.1.2 Cross-lingual

3.2.1.2.1 Synthetic During the emerging era of the cross-lingual summarization task, most studies rely on synthetic cross-lingual datasets developed with Round Trip Translation (RTT) and two popular pipelines: (i) Translate then Summarize (Trans-Sum) and (ii) Summarize then Translate (Sum-Trans) (Zhu et al., 2019, 2020). The RTT and pipelines rely on machine translation systems for generating cross-lingual text. Ouyang et al. (2019) propose an RTT-based cross-lingual model for the NYT dataset. They use RTT to convert the English dataset into three low-resource languages - Somali, Swahili, and Tagalog and then back into noisy English. They use an on-shelf neural machine translation (NMT) system, *i.e.*, Marian Framework[1] for RTT, and create four noisy English datasets named Somali, Swahili, Tagalog and a mix (even mix of all three). After creating these synthetic cross-lingual datasets, they train the PGN (See et al., 2017) model for each dataset. They use noisy English documents as inputs and the original English summaries as targets. Each trained model is evaluated with all four test sets to investigate the performance of trained models. Duan et al. (2019) investigate the

[1] https://marian-nmt.github.io/

cross-lingual abstractive sentence summarization task for Chinese-English. They also use an on-shelf NMT system for RTT to create a synthetic Gigaword dataset. For the summarization model, they modify the vanilla transformer Vaswani et al. (2017) with a teacher-student setting for generation and attention. They train the models with different settings: (i) Trans-Sum, (ii) Sum-Trans, (iii-iv) few-shots synthetic Chinese input and summary and (v) variations of teaching generation and attention.

Zhu et al. (2019) introduce a multitask learning model for English-Chinese cross-lingual summarization. They develop two variations of the transformer model (Vaswani et al., 2017) with one shared encoder and two independent decoders. They create synthetic cross-lingual datasets by RTT with an on-shelf service[2]. For experiments, they create two joint training models: (i) cross-lingual summarization + machine translation, (ii) cross-lingual summarization + monolingual summarization, and two variations of the training strategy for baselines: (i) translate early, and (ii) translate late. Cao et al. (2020b) present a multitask learning model for cross-lingual summarization by joint learning of alignment and summarization. Their model consists of two encoders and two decoders, each dedicated to one task while sharing contextual representations. The authors evaluate their model on synthetic cross-lingual datasets for the English-Chinese language pairs. Takase and Okazaki (2020) introduce a multitask learning framework for cross-lingual abstractive summarization by augmenting monolingual training data with translations for three pairs: (i) Chinese-English, (ii) Arabic-English, and (iii) English-Japanese. The model consists of a transformer encoder-decoder model with prompt-based learning in which each training instance is affixed with a special prompt-to-signal example type. Bai et al. (2021) develop a variation of multi-lingual BERT for English-Chinese cross-lingual abstractive summarization. The model is trained with a few shots of monolingual and cross-lingual examples. Bai et al. (2022) extend their work by introducing a multitask learning model to improve cross-lingual summaries by combining cross-lingual summarization and translation rates. They create synthetic cross-lingual datasets by RTT. They add a compression scoring method at the encoder and decoder of their model. They augment their datasets for different compression levels of summaries. One variation of their model comprises cross-lingual and monolingual summarization decoders, while the other comprises cross-lingual and translation decoders.

3.2.1.2.2 Non-synthetic Nguyen and Daumé III (2019) collect a dataset from descriptions of 4,573 news articles from the Global Voices[3] website in 15 languages - English, Spanish, French, Italian, Portuguese, Russian, Greek, German, Dutch, Malagasy, Bengali, Arabic,

[2] http://www.anylangtech.com/
[3] https://globalvoices.org/

Macedonian, Swahili and Japenese. The translations of the news articles are performed by their Lingua team. The English summaries are created with clickbait descriptions on social media via crowd-sourcing, so these can be argued as not true summaries as these summaries might not contain sufficient salient information.

3.2.2 Wikipedia Articles

Wikipedia is considered a feasible and reliable source for data collection in the summarization community. Here we discuss summarization datasets collected from either Wikipedia or other encyclopedias.

3.2.2.1 Monolingual

Zopf et al. (2016) create an English dataset for multi-document summarization. The dataset contains 91 Wikipedia articles paired with 1,265 source documents. The reference summaries are generated by human annotators by using an automatic extraction technique from source documents. They evaluate their dataset with various extractive summarizers - Lead, Random, optimal, TF-IDF (Luhn, 1958), LexRank (Erkan and Radev, 2004), latent semantic analysis (LSA) (Steinberger et al., 2004) and ICSI (Gillick et al., 2009). Antognini and Faltings (2020) construct an English multi-document summarization dataset containing 14,652 game reviews from Wikipedia. They investigate the performance of various strong extractive and abstractive baselines - Lead, TextRank (Mihalcea and Tarau, 2004), LexRank, SumBasic (Nenkova and Vanderwende, 2005), C-Skip (Rossiello et al., 2017), Conv2Conv (Gehring et al., 2017), the vanilla transformers (Vaswani et al., 2017) and its variations. Gholipour Ghalandari et al. (2020) also collect English multi-document summarization dataset from the Wikipedia Current Events Portal (WCEP). The WCEP is a portal for international news events. They collect source news articles from the Internet Archive Wayback Machine. The dataset contains around 2.39 million news articles divided into 10,200 clusters. They evaluate their dataset with various extractive baselines - Lead, Oracale-multi, Oracale-single, Lead-Oracale (first sentences up to 40 words), Random-Lead (lead of randomly selected article), TextRank, Centroid (Radev et al., 2004), regression-based sentence ranking (Ren et al., 2016) and extractive BERT (Kenton and Toutanova, 2019).

Hättasch et al. (2020) collect the Fandom wikis dataset in English and German for monolingual single and multi-document summarization. The dataset consists of fictional narratives and descriptions of movies, television and book series. They divide their dataset into two parts - Harry Potter and Star Wars. The English Harry Potter set contains 15,993 articles and 1,466 candidate summaries. The English Star Wars set has 148,348 articles with 5,659

candidate summaries. The German Star Wars set contains 39,356 articles with 999 candidate summaries. However, this dataset can be used in monolingual summarization only. They investigate the performance of several extractive baselines - Luhn (Luhn, 1958), LexRank, LSA, KL-Greedy (Haghighi and Vanderwende, 2009), ICSI and S2S extractor (Kedzie et al., 2018). Frefel (2020) collect a monolingual dataset for single-document summarization in the German language. They extract 240K Wikipedia articles from various categories such as people, science, sports, history, politics, art and culture. They evaluate their datasets with extractive baselines - Random-3, Lead-3 and TextRank. Aumiller and Gertz (2022) create a German dataset for joint summarization and simplification for children or dyslexic readers from the German children's encyclopedia "Klexikon". The authors investigate the performance of various extractive baselines on their dataset - Lead-3, Lead-k, ROUGE-2 Oracle, Luhn and LexRank.

3.2.2.2 Cross-lingual

During 2011-2015, Text Analysis Conference (TAC) introduce a shared task - MultiLing for multi-lingual multi-document summarization. They also develop datasets collected from English Wikinews (Giannakopoulos et al., 2011; Giannakopoulos, 2013; Giannakopoulos et al., 2015). These articles are translated with a sentence-by-sentence approach into 9 other languages: Arabic, Czech, French, Greek, Hebrew, Hindi, Chinese, Romanian and Spanish. The final dataset - MultiLing'15 contains a total of 1500 documents in 10 languages and 15 topics per language and 10 texts per topic. This dataset is mostly used with extractive methods.

Ladhak et al. (2020) create a cross-lingual dataset - WikiLingua by collecting parallel data in 12 languages from the WikiHow website[4]. The articles on Wikihow describe the instructions to solve a specific problem or task. The English parallel article-summary set is the biggest in the size, with a total of 141,457 articles, while other languages contain much fewer articles. They evaluate their dataset for English summaries generation from other languages with existing extractive and abstractive baselines - Lead-3, Trans-Sum, Sum-Trans, RTT and multilingual BART (mBART) (Liu et al., 2020).

3.2.3 Scientific Articles

In this section, we focus on scientific abstractive summarization, which is a quite challenging task. The scientific text has a well-defined structure, including an abstract, introduction, methods, experiments, and conclusions. Due to this standard structure, salient information that should be present in a summary spreads over the document in different sections. So the

[4] https://www.wikihow.com/Main-Page

abstractive models for scientific text summarization require longer inputs than the newswire to generate a meaningful summary.

3.2.3.1 Monolingual

Kim et al. (2016a) create a dataset of scientific papers from ArXiv for scientific abstractive summarization. They collect LaTeX source files of articles from the computer science domain. Then they extract introductions and abstracts from the source files. The training set consists of 10k paragraph-sentence pairs. The authors also develop an MTGRU-based (Multiple Time Scale Gated Recurrent Unit) encoder-decoder model consisting of 4 layers, 1792 hidden units and an embedding size of 512. The model is trained with the paragraphs and their target sentences. For each introduction, the generated sentences are concatenated to create a summary that is then evaluated against the gold abstract. The MTGRU-based model performs better than the conventional RNN model, however, it also struggles with empty and/or unknown tokens (UNK) as the vocabulary size is limited in the conventional neural models.

Nikolov et al. (2018) create two datasets from scientific articles from MEDLINE and PubMed. They extract 5M abstract-title pairs from MEDLINE for the title generation task and 900K paper-abstract pairs from PubMed for the abstract generation task. The authors parse metadata of MEDLINE and PubMed that is present in XML format. They skip all figures, tables, and section headings in the papers. They apply tokenization, lowercase conversion, and numeral and URL removal from the MOSES statistical machine translation pipeline. They also remove all the pairs where the abstract length is less than 150 tokens or greater than 370 tokens, the title length is less than 6 tokens or greater than 25 tokens and the body length is less than 700 tokens or greater than 10,000 tokens. They investigate the performance of several existing extractive and abstractive systems, including Lead, LexRank, TF-IDF-Emb, RWMD-Rank, LSTM, FConv and C2C. They struggle with memory issues for abstract generation, and the neural models did not show promising performance for generating abstracts.

Cohan et al. (2018) collect a scientific dataset consisting of 215K scientific articles from ArXiv and 133K from PubMed. Each article is paired with the corresponding abstract for scientific abstractive summarization. They remove the documents that are either extremely long - theses or extremely short - tutorial announcements and missing abstract or discourse structure. They use the level-1 section header as the discourse marker. For ArXiv, they process the latex source files with PanDoc to convert them into plain text with preserved discourse markers. They remove figures and tables and add special tokens for citations and mathematical formulas. The authors also remove the sections after the conclusion section. They develop a discourse-aware summarization model. The encoder is a hierarchical RNN encoder (bi-directional LSTM) for word and sentence levels. The decoder is a uni-directional LSTM with a

copying and coverage mechanism.

Yasunaga et al. (2019) create a scientific dataset consisting of 1,000 most cited papers in the ACL Anthology Network (AAN) (Radev et al., 2013). These articles are paired with their citation information and gold summaries annotated by experts. These articles have a citation range of 21-298. For each reference paper (RP), the authors collect 20 clean citation sentences assuming them as summaries and keep only 15 citation sentences on average after cleaning. They manually annotate their dataset for salient citation sentences and human written summaries based on abstract and these salient sentences. The authors create a hybrid summarization model based on the Graph convolutional network (GCN) (Kipf and Welling, 2016) and LSTM-based sentence encoder. The summary is generated with two greedy heuristics to select sentences.

Cachola et al. (2020) introduce a new task TLDR (too long; didn't read) generation from scientific papers as a downstream task of extreme summarization. TLDRs focus on the key aspects of the paper, like contributions. The authors collect 3,230 scientific abstracts from the computer science domain and paired them with their corresponding TLDRs. The training set consists of a single TLDR per paper, however, validation and testing sets might contain multiple instances of TLDR per paper. They collect pairs of papers and author-written TLDRs from the Open Review website. The papers are presented in PDF format, then processed with S2ORC pipeline (Lo et al., 2019). They also collect peer-review summaries to create manual TLDRs. The authors develop a model - CATTS (Controlled Abstraction for TLDRs with Title Scaffolding). They compare the performance of CATTS with extractive and abstractive methods such as PACSUM (Zhang et al., 2020b), BERT (Kenton and Toutanova, 2019), BART (Lewis et al., 2019) and their variants.

3.2.3.2 Cross-lingual

Cross-lingual scientific summarization is an understudied area due to its challenging nature. We only find one study for extreme cross-lingual summarization of scientific texts. Recently, Takeshita et al. (2022) follow the steps of Cachola et al. (2020) for the TLDR task. However, they construct a synthetic dataset for cross-lingual-TLDR (CL-TLDR) or TLDR-like extreme summary of scientific texts. They take the original TLDR dataset (Cachola et al., 2020) and translate them into four languages - German, Italian, Chinese and Japanese. They first create synthetic cross-lingual summaries with DeepL[5] and then manually correct a subset of those translations. Due to this reason, the dataset size is quite small. The total size of the English, German, Italian and Chinese sets is 3229 documents, while the size of the Japanese set is 2004 articles. They evaluate their datasets with existing abstractive pre-trained models - BART and

[5]https://www.deepl.com/pro-api?cta=header-pro-api

mBART. To the best of our knowledge, there is no previous work on cross-lingual scientific summary generation.

3.2.4 Observations

We find that the majority of the studies focus on monolingual or synthetic cross-lingual summarization. Most of the abstractive summarization models have been developed for newswire as they contain short articles and summaries compared to Wikipedia and Scientific texts. The RNN-based abstractive models have struggled with text length and adequate vocabulary size. With the RNN models, a word-based fix-sized vocabulary has been used, due to which models tend to produce UNK tokens more frequently. However, the breakthrough invention of the transformer model and its successors - BERT, BART, PEGASUS, *etc*., has rapidly reshaped the abstractive summarization field.

3.3 Simplification

Text simplification has been explored in diversified directions and languages. Most of the work has been investigated for the mentally challenged people, children, and foreign language learners as targeted audience (Rello et al., 2013; Klaper et al., 2013; Cholakov et al.; Saggion et al., 2015; Suter et al., 2016; Štajner, 2021), expert to layman simplification for medical domain (Van den Bercken et al., 2019; Cao et al., 2020a; Devaraj et al., 2021; Phatak et al., 2022) and legal domain (Sheremetyeva, 2014; Collantes et al., 2015; Garimella et al., 2022). However, there are only studies present that target generic simplification without considering the special target audience (Al-Thanyyan and Azmi, 2021). In this section, we focus on the simplification studies for a generic targeted audience and divide it into two categories.

3.3.1 Non-Scientific

Zhu et al. (2010) construct a parallel dataset for sentence simplification from English Wikipedia and Simple English Wikipedia. The targeted audience of Simple Wikipedia includes children and English-learner adults. They collect 65,133 parallel articles from English Wikipedia and Simple Wikipedia for their dataset - PWKP. They tokenize the PWKP for sentence boundaries and use TF-IDF for sentence alignment. The total size of PWKP is 108K sentences. They evaluate their dataset for sentence simplification and substitution tasks with probabilistic models. Coster and Kauchak (2011) also create a parallel dataset from Wikipedia and Simple Wikipedia for sentence-level simplification. They generate 137K simplified and unsimplified aligned sentence pairs from English Wikipedia and Simple English Wikipedia. Simple

English Wikipedia contains similar content to English Wikipedia but with simpler vocabulary and grammar. (Kajiwara and Komachi, 2016) harvest a parallel dataset from English Wikipedia and Simple English Wikipedia. They collect 492,993 sentence pairs from 126,725 article pairs. Then they evaluate their dataset with existing Statistical Machine Translation (SMT)-based text simplification models.

Vajjala and Lučić (2018) construct a dataset from the OneStopEnglish website by extracting their articles from 2013 to 2016. The website has English language learning resources for adult ESL learners (elementary, intermediate, and advanced). Their source articles are collected from The Guardian newspaper and rewritten by teachers for three different levels of learners. Gonzales et al. (2021) collect a dataset from a Swiss magazine, 20-Minuten, in the German language for document-level simplification. They collect around 18K articles published on the magazine websites. However, this magazine publishes articles only related to lifestyle.

Xu et al. (2015) collect a simplified news corpus - Newsela that contains 1130 news articles paired with four simplified versions by trained professionals to create different reading levels. Pavlick and Callison-Burch (2016) collect a dataset of 1,000 phrases from the vocabulary of the PPDB (Paraphrase Database), also present in the Newsela dataset. Paetzold and Specia (2017) develop a lexical simplification model based on neural ranking to evaluate it on the Newsela dataset. Laban et al. (2021) develop a reinforcement learning model for multi-sentence text simplification without the need for parallel corpora. They use part of the Newsela dataset to create their model.

3.3.2 Scientific

Kim et al. (2016b) develop a lexical simplification approach to simplify scientific terminology. The authors collect a parallel corpus based on scientific publications and Simple Wikipedia. They develop a simplification pipeline - SimpleScience and apply it to around 500K articles of PLOS (Public Library of Science) and PubMed Central (PMC). They pair these articles with Simple Wikipedia. They also annotate a subset of scientific terminologies via crowd-sourcing.

To the best of our knowledge, there is no other previous study for scientific text simplification. Recently, Ermakova et al. (2021, 2022) have initiated a workshop investigating science simplification in CLEF2021 and CLEF2022. They provide participants with the abstract of scientific articles and aim to generate a simplified version of these abstracts. Monteiro et al. (2022); Menta and Garcia-Serrano (2022) solve this task with pre-trained T5 models.

3.3.3 Observations

We conclude this section with the findings that a lot of research has been conducted in the simplification field, however, mostly for a specific targeted audience and non-scientific texts. Therefore, the scientific simplification task needs a thorough investigation.

3.4 Summary

This chapter discusses the relevant previous work from science journalism, summarization and simplification. We find that monolingual science journalism is an understudied area to date, and only a few studies focus on it. To the best of our knowledge, no prior work (dataset, model) exists on Cross-lingual Science Journalism. In the summarization field, most of the work has been conducted on monolingual or synthetic cross-lingual summarization. Furthermore, traditional abstractive models struggle with long inputs. However, transformer-based summarization models show promising results. In simplification, we find that it is a diverse field and scientific text simplification has not gained much attention. We conclude this section with these findings: (1) Cross-lingual Science Journalism has not been explored before this work, and (2) cross-lingual scientific summarization and scientific simplification are extremely understudied fields.

Part II

Datasets for Summarization and Science Journalism

Chapter 4

Data Collection and Evaluation

> "Data is a precious thing and will last longer than the systems themselves."
>
> Tim Berners-Lee

In the previous chapter, we found that there are no existing Cross-lingual Science Journalism and cross-lingual scientific summarization datasets and models. We also found that Wikipedia has been a source of multilingual summarization datasets. In this chapter, we focus on our first two research questions:

1. **How can we obtain SPEKTRUM source articles and similar data from WIKIPEDIA and measure similarities and differences in both datasets?** We divide it further into the following questions:
   (a) How can we obtain online data with minimum anomalies?
   (b) What kind of statistical features help to find intrinsic properties of collected datasets?
2. **Can we train Cross-lingual Science Journalism models with WIKIPEDIA science articles and evaluate them on science summaries from SPEKTRUM?**

This chapter provides detailed answers to these questions. We discuss Spektrum and Wikipedia data collection methodologies in Sections 4.1 and 4.2. Section 4.3 presents a statistical overview of collected datasets, and Section 4.4 provides a detailed readability analysis based on various linguistic features. Section 4.5 presents the experimental setup and evaluation for monolingual and cross-lingual summarization tasks.

SPEKTROGRAMM

INFORMATIK
KLIMABILANZ VON BITCOINS

Chinesische Bitcoin-Miner könnten 2024 so viel Energie verbrauchen wie ganz Italien. Das ist das Ergebnis eines chinesischen Forscherteams, das die weitere Entwicklung der Kryptowährung mit ausgeklügelten Modellen simuliert hat. Demnach wird sich der derzeitige Run auf Bitcoins noch Jahre fortsetzen. Und wenn die Zentralregierung in Peking nicht gegensteuere, könne der Trend gar die Klimaschutzbemühungen des Landes unterlaufen, warnen die Wissenschaftler.

Bitcoins basieren auf der Blockchain-Technologie. Jeder Mensch kann damit im Prinzip neue digitale Münzen herstellen, die dann bares Geld wert sind. Dazu muss er oder sie aber einen Computer spezielle Rechenaufgaben lösen lassen. Das lohnt sich in Ländern mit niedrigen Stromkosten sowie mit leistungsfähiger Hardware, welche die Rechnungen möglichst energieeffizient durchführt.

Schon länger ist bekannt, dass die wachsende Bitcoin-Branche einen gewaltigen Energiehunger hat. Das Team um Shangrong Jiang von der Universität der Chinesischen Akademie der Wissenschaften wollte nun ergründen, wie stark dieser noch anwachsen könnte. Die Forschenden entwickelten dazu vier verschiedene Szenarien, die nicht nur marktwirtschaftliche Überlegungen miteinbeziehen, sondern auch mögliche Antworten der Politik.

Ohne politische Intervention oder unvorhergesehene Kursabstürze wird der Bitcoin-Hype demnach erst 2024 seinen Höhepunkt erreichen: In China könnte das Schürfen der Kryptowährung dann pro Jahr 297 Terawattstunden Strom verbrauchen – stolze fünf Prozent des landesweiten Strombedarfs – und dabei 130 Millionen Tonnen CO_2 freisetzen. Langfristig werde das Geschäft zwar unattraktiver, da die Kryptowährung so angelegt ist, dass die für eine Münze erforderlichen Rechenaufgaben mit der Zeit immer anspruchsvoller werden. Doch bis dahin sei der Run auf Kryptowährungen ein Hemmschuh für den Klimaschutz.

Die Autoren plädieren dafür, das Schürfen mit zielgerichteten Gesetzen umweltverträglicher zu machen, etwa mit einer CO_2-Steuer oder Mindestanforderungen an die Energieeffizienz der Schürfer. Noch besser wäre es dem Team zufolge, wenn das Mining nur in Regionen mit einem hohen Anteil an Wasserkraft erlaubt würde, und nicht dort, wo es zurzeit oft stattfinde: in Gegenden, die ihren Strom vorrangig aus Kohlekraftwerken beziehen.

Nature Communications, 10.1038/s41467-021-22256-3, 2021

SONNENSYSTEM
EISWOLKE ÜBER MARSVULKAN

Die ESA-Sonde Mars Express hat ein besonderes Phänomen beobachtet: eine Wolke aus Wassereis, die sich 1800 Kilometer über die Oberfläche des Roten Planeten erstreckt. Sie bildet sich offenbar stets während der Frühlings- und Sommersaison über der Tharsis-Region, wenn Winde die dünne Marsluft am Hang von Arsia Mons emporblasen, dem zweitgrößten Vulkan unseres Nachbarplaneten.

Schon länger ist bekannt, dass die Atmosphäre des Mars in winzigen Mengen Wasserdampf enthält. Die H_2O-Moleküle haben sich aus Eis an der Oberfläche gelöst, wie man es an den Polen findet. Gelangt das Wasser in eine Höhe von einigen dutzend Kilometern, erstarrt es zu Eiskristallen und bildet Wolken wie die auf dem Bild von Mars Express.

Forscher haben immer wieder ähnliche Gebilde beobachtet, etwa über dem Grabensystem Valles Marineris. Und auch über Arsia Mons ist das Phänomen nicht neu. Doch bisher ist es nicht gelungen, die Wolke in ihrer ganzen Ausdehnung zu erfassen, da die meisten Marssonden die Tharsis-Region nur während des Nachmittags gut im Blick haben.

Bei Mars Express machte dies erst die Umwidmung einer Weitwinkelkamera möglich, die eigentlich für die Sichtkontrolle des 2003 abgekoppelten Beagle-2-Landemoduls gedacht war. Mittlerweile nutzt die ESA das Instrument für Forschungszwecke und konnte so die Entwicklung der Wolke über Arsia Mons vollständig erfassen.

Die Wolke bildet sich demnach immer in den Morgenstunden, steigt auf bis zu 40 Kilometer Höhe und breitet sich mit 600 Kilometern pro Stunde aus, wobei die Winde sie bevorzugt nach Norden blasen. Gegen Nachmittag löst sie sich dann auf – bis zum nächsten Morgen, wenn erneut Luft an der Flanke von Arsia Mons aufsteigt.

Journal of Geophysical Research Planets, 10.1029/2020JE006517, 2020

MARSWETTER Im Frühling und Sommer bildet sich täglich eine Wolke über dem Vulkan Arsia Mons.

ESA/DLR/FU BERLIN

Figure 4.1: A sample of science summaries in the SPEKTROGRAMM section from an online published Spektrum magazine.

4.1 Spektrum Data Collection

Spektrum der Wissenschaft (SPEKTRUM) is the German equivalent of the "Scientific American", which began publishing in 1978. The SPEKTRUM magazine is one of the divisions of the Springer Nature publishing group. It is published monthly and covers many core areas of science, such as archaeology, astronomy, biology, chemistry, *etc*. It has a special section - SPEKTROGRAMM dedicated to summaries of the latest research articles. In this section, SPEKTRUM science journalists present complex English scientific research to non-expert readers in their local language (German). Figure 4.1 shows a couple of summaries from the SPEKTRO-

GRAMM section of an online published issue.

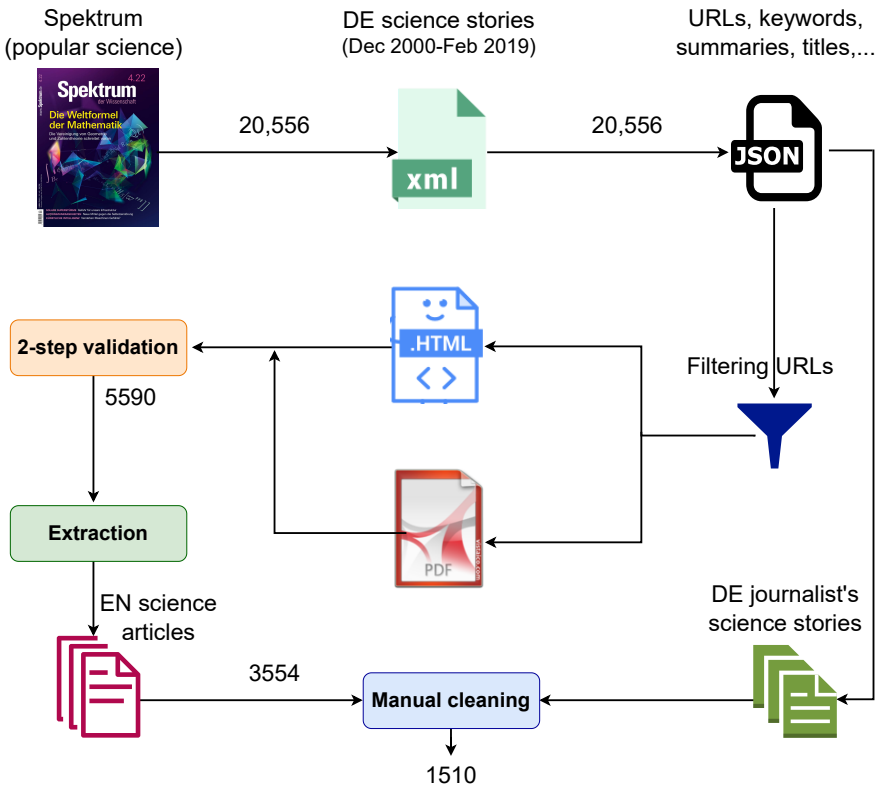

Figure 4.2: SPEKTRUM data collection steps.

The SPEKTRUM management approached us for a research collaboration to automate a part of their process related to SPEKTROGRAMM for facilitating the work of science journalists. After the formal meetings with SPEKTRUM's managing director and head of the digital production and legal processing, SPEKTRUM released their data for the research purpose. Figure 4.2 presents an overview of SPEKTRUM data collection steps. The details of these steps are discussed below.

4.1.1 Raw XML Data

We receive a subset of the SPEKTRUM data in XML format. The released SPEKTRUM raw data contains German summaries and URLs to their source documents. It consists of 20,556 summaries from December 2000 to February 2019. Code 4.1 shows an example from the raw SPEKTRUM data.

```
1 <?xml version="1.0" encoding="UTF-8" ?>
2 <!DOCTYPE sixcms_data SYSTEM "http://www.sixcms.de/dtd/sixcms-1.3.1.dtd">
3 <sixcms_data>
4 <sixcms_info>
5   <system generator="SixCMS 9" version="9.0.12rc3" release="2014-04-01" change="2019-01-29"
      licence="wissenschaftonlinegmbh01" />
```

```
  <document creation_date="2019-05-02 14:24:57" creation_user="dirdjaja" count="11081" />
</sixcms_info>

<sixcms_article>
  <online_date><![CDATA[2019-05-02 11:50:00]]></online_date>
  <field name="r_sdi_art" type="relation" key="title" container="SdW-Verlag: Online-Artikel:
    Artikel-Arten" container_icon="content"><![CDATA[News]]></field>
  <field name="text" type="text">
      <![CDATA[<p>Die Ursachen des atopischen Ekzems - besser bekannt unter der
    veralteten Bezeichnung Neurodermitis - sind komplex und längst nicht geklärt. Die
    Krankheit dürfte häufig genetische Ursachen haben, die das Immunsystem verändern und Kö
    rper von Betroffenen auf äuere Einflüsse extremer reagieren lassen, mit dem
    charakteristisch starken Juckreiz oder Hautrötungen. Eine mitentscheidende Rolle könnten
    aber auch Mikroorganismen auf der Haut haben: Bei Erkrankten finden sich etwa auffällig
    groe Mengen von <em>Staphylococcus aureus</em>, die Toxine ausschütten und Entzü
    ndungsreaktionen verstärken. Forscher um Richard Gallo von der University of California
    in San Diego haben nun untersucht, warum diese Keime bei Betroffenen so schädlich werden&
    nbsp;- und stellen fest, dass daran der Rest der Bakteriengemeinschaft auf der Haut nicht
     unschuldig ist.</p>

      <p>Im <a href="https://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aat8329"
    target="_blank">Fachblatt Science Translational Medicine</a> haben die Forscher nach
    Unterschieden in der Vielfalt der Mikroorganismen auf der Haut von Gesunden und
    Neurodermitikern gefahndet und versucht herauszufinden, ob dies eine Folge oder gar eine
    Ursache des Krankheitsgeschehens sein könnte. Klar war bereits, wie das Immunsystem von
    Patienten das Hautmikrobiom beeinflusst: Eine übergroe Menge bestimmter Zytokin-
    Botenstoffe <a href="https://www.spektrum.de/alias/allergieursachen/hautdefekte-
    beguenstigen-allergien/1173254" target="_blank">schwächt bei ihnen nicht nur die Kö
    rperbarriere</a>, weil sie den <a href="https://www.spektrum.de/lexikon/biochemie/
    filaggrin/2194" target="_blank">Verhornungsprozess von Hautzellen</a> bremst; sie senkt
    zudem die Produktion von bestimmten antimikrobiellen Peptiden, die im Normalfall auf der
    Haut ausgeschüttet werden. Fraglich war allerdings, warum davon <em>S. aureus</em>
    besonders profitiert.</p>

      <p>Die Wissenschaftler haben die Vorgänge nun mit Hautproben von elf Menschen
    mit atopischem Ekzem und von Gesunden sowie auf Hautzellen von Mäusen untersucht, auf
    denen sie <em>S.-aureus</em>-Kolonien angesiedelt haben. Dabei wurde deutlich, dass die
    Keime sich auf der Haut von Patienten effizienter in gröeren Mengen sammeln und
    koordinieren konnten: Sie kommunizierten dabei untereinander über <a href="https://www.
    spektrum.de/lexikon/biologie/quorum-sensing/55384" target="_blank">Quorum-Sensing-Signale
    </a> und gaben schlielich proteinabbauende Enzyme ab, die die Hautzellen stark angreifen.
     Eigentlich könnten dieselben Bakterien dies auch auf der Haut von gesunden Menschen tun&
    nbsp;- hier aber wird ihre Quorum-Sensing-Kommunikation durch den Einfluss anderer
    Bakterien unterbunden, wie die Forscher zeigen. Eine entscheidende Rolle spielt dabei
    etwa die Art <em>Staphylococcus hominis</em>, ein typischer Vertreter des
    durchschnittlichen Hautmikrobioms - diese Keime produzieren ein blockierendes Peptid
     und sorgen so dafür, dass <em>S. aureus</em> harmlos bleibt.</p>
      <!-- Heft 1 XL -->
      <p>Die Interaktion verschiedener Bakterienarten auf der Haut ist demnach für das
    Krankheitsgeschehen wichtig. Womöglich bietet sich hiermit sogar eine bisher nicht
    angedachte Behandlungsoption: So könnten etwas Bakterien von Gesunden, Kulturen von <em>
    Staphylococcus hominis</em> oder ein Gemisch der Blockadesignale auf Kranke
    transplantiert werden, um die zerstörerische Wirkung der schädlichen Keime zu bremsen. Im
     Versuch an Mäusen sind die ersten Experimente in dieser Richtung viel versprechend, so
    die Forscher um Gallo: Sowohl übertragene <em>S.-hominis</em>-Kulturen wie auch der
```

```
    Einsatz der von diesen produzierten blockierenden Peptiden sorgten dafür, dass die
    typischen Hautzellschäden durch <em>S. aureus</em> stark minimiert wurden. Nun
    wollen die Forscher noch untersuchen, welche unterschiedlichen <em>S.-aureus</em>-
    Kulturen so gebremst werden können - und welchen Einfluss möglicherweise weitere
    Mikroorganismen in der Haut haben können.</p>]]>
        </field>
  <field name="r_copyright" type="relation" key="title" container="SdW-Verlag: Online-
    Artikel: Quellen (copyright)" container_icon="content"><![CDATA[Spektrum.de]]></field>
  <field name="dachzeile" type="text"><![CDATA[Atopisches Ekzem]]></field>
  <title><![CDATA[Gesunde Hautflora schützt vor Neurodermitis]]></title>
  <field name="untertitel" type="text"><![CDATA[Hilfreiche Bakterien stoppen zerstörerische
    Keime]]></field>
  <field name="teaser" type="text"><![CDATA[Beim atopischen Ekzem stoen schädliche
    Hautbakterien in die Lücken, die die Krankheit gerissen hat - und sorgen so für Juckreiz
    und Rötung. Andere Keime könnten sie stoppen.]]></field>
  <field name="vorspann" type="text"><![CDATA[Beim atopischen Ekzem stoen schädliche
    Hautbakterien in die Lücken, die die Krankheit gerissen hat - und sorgen so für
    Juckreiz und Rötung. Andere Keime könnten sie stoppen.]]></field>
  <field name="r_mitarbeiter" type="relation" key="title" container="System(übergreifend):
    Autoren- und Mitarbeiterprofile" container_icon="content"><![CDATA[Osterkamp]]></field>
  <field name="summary_title" type="text"></field>
  <field name="summary_content" type="text"></field>
  <field name="r_hauptkategorie" type="relation" key="title" container="System(übergreifend):
     Fachgebiete: Hauptkategorien" container_icon="content"><![CDATA[Medizin]]></field>
  <field name="links_lexika" type="text"></field>
  <field name="l_haupt_fg" type="links" key="title" container="System(übergreifend):
    Fachgebiete: Hauptkategorien" container_icon="content"><![CDATA[Medizin]]></field>
  <field name="quellen" type="text"></field>
  <field name="dachzeile_produkt" type="text"></field>
  <field name="title_produkt" type="text"></field>
  <field name="metatitle" type="text"></field>
  <field name="metadescription" type="text"></field>
  <field name="keywords" type="text"><![CDATA[atopisches ekzem, neurodermitis, hautbakterien,
     quorum sensing, s. aureus, staphylococcus, hautkeime, ]]></field>
  <id><![CDATA[1642530]]></id>
  <gsid><![CDATA[sdw.c.1642530.de]]></gsid>
</sixcms_article>
</sixcms_data>
```

Listing 4.1: An example of SPEKTRUM German story from raw XML data with embedded URLs in it.

4.1.2 Processing of XML Data

We manually analyze the data at first and develop an XML parser to process the following attributes: ids, dates, titles, keywords, summaries, URLs in Quellen (Source), URLs in text.

The raw data only contains German summaries, and we must find the source English articles online and their URLs are either present in the Quellen field or embedded in the text field. However, scraping source articles is a challenging task. We find three cases for URLs:

1. There is no URL present in Quellen and text fields.

2. There is only one URL present.
3. There are multiple URLs present.

The last case makes finding the source hard as multiple links are involved. Upon manual inspection of a subset, it is found that only one URL is a source link while others are further reading links. So, now we have to find which URL best fits for the source article. Code 4.2 shows 6 instances parsed from the raw SPEKTRUM data from the JSON file.). Instance 2 contains no links (case 1), so it will be discarded later.


```
{"Id":{"0": "1642530", "1": "1642462", "2": "1641998", "3": "1642380", "4": "1642290", "5": "
    1642272",}, "Date":{"0": "5-2-2019", "1": "5-2-2019", "2": "5-1-2019", "3": "4-30-2019",
    "4": "4-30-2019", "5": "4-30-2019",}, "Title":{"0": "Gesunde Hautflora schützt vor
    Neurodermitis", "1": "Die unterschätzte Bedeutung von Nilpferdkot", "2": "Blicke ins
    Zentrum unseres Galaxienhaufens", "3": "2000 Jahre altes Himmelsrätsel gelöst", "4": "Das
     Geheimnis der Ibis-Mumien", "5": "Mond könnte aus Lavameer entstanden sein"}, "Keywords"
    :{"0": "atopisches ekzem, neurodermitis, hautbakterien, quorum sensing, s. aureus,
    staphylococcus, hautkeime", "1": "nilpferd, flusspferd, hippo, kot, exkrement, dünger, gü
    lle, ökosystem, silizium, stofftransport, nahrungskette, kieselalgen", "2": "beobachtung,
     galaxien, astronomie, Galaxienhaufen, Markarians Kette, Walfisch-Galaxie", "3": "
    Astronomie, Nova, Chinesisch, Geschichte, Sternexplosion, historische Aufnahme, weier
    Zwerg", "4": "Heiliger Ibis, Altes gypten, Archäologie, gyptologie, Biologie, Mumien,
    Bestattung, Opfer, Tempel", "5": "Mond, Entstehung, junge Erde, Sonnensystem, Astronomie,
     Crash, Magma, Lava",}, "Urls_quellen":{"0": null, "1": null, "2": null, "3": null, "4":
    null, "5": null}, "Urls_text":{"0": ['https://stm.sciencemag.org/lookup/doi/10.1126/
    scitranslmed.aat8329'], "1": ['https://doi.org/10.1126/sciadv.aav0395'], "2": [], "3": ['
    https://arxiv.org/abs/1904.11515', 'https://iopscience.iop.org/article/10.1088/0004-637X
    /756/2/107/meta'], "4": ['https://www.biorxiv.org/content/10.1101/610584v1.full'] "5":['
    https://www.nature.com/articles/s41561-019-0354-2'],},}
```


Listing 4.2: A sample of 6 parsed data instances from the SPEKTRUM raw data.

4.1.3 Translation of Keywords

To find the best-fit source link among the URLs for the source articles, we translate all keywords from German to English by using Google Translate[1]. After that, a manual inspection of translated keywords is conducted to correct any wrong translations manually by a native German speaker. Code 4.3 presents keywords and translation pairs of 6 instances from SPEKTRUM data.


```
{"Keywords":{"0":{"atopisches ekzem": "atopic eczema", "neurodermitis": "neurodermatitis", "
    hautbakterien": "skin bacteria", "quorum sensing": "quorum sensing", "s. aureus": "s.
    aureus", "staphylococcus": "staphylococcus", "hautkeime": "skin germs"},"1":{"nilpferd":
    "hippo", "flusspferd": "hippopotamus", "hippo": "hippo", "kot": "excrement", "exkrement":
     "excrement", "dünger": "fertilizer", "gülle": "slurry", "ökosystem": "ecosystem", "
    silizium": "silicon", "stofftransport": "mass transport", "nahrungskette": "food chain",
    "kieselalgen": "diatoms"},"2":{"beobachtung": "observation", "galaxien": "galaxies", "
```


[1] https://pypi.org/project/googletrans/

```
    astronomie": "astronomy", "galaxienhaufen": "galaxies heap", "markarians kette": "
    markarian's chain", "walfisch-galaxie": "whale-galaxy"},"3":{"astronomie": "astronomy", "
    nova": "nova", "chinesisch": "chinese", "geschichte": "history", "sternexplosion": "
    starburst", "historische aufnahme": "historical recording", "weier zwerg": "white dwarf"
    }, "4":{"heiliger ibis": "holy ibis", "altes ägypten": "old egypt", "archäologie": "
    archeology", "ägyptologie": "egyptology", "biologie": "biology", "mumien": "mummies", "
    bestattung": "funeral", "opfer": "victim", "tempel": "temple"},"5":{"mond": "moon", "
    entstehung": "formation", "junge erde": "young earth", "sonnensystem": "solar system", "
    crash": "crash", "magma": "magma", "lava": "lava"}}}
```


Listing 4.3: A sample of 6 instances translated keywords from SPEKTRUM data via Google Translate.

4.1.4 Filtering of URLs

Before scraping the source URLs, we filter all URLs to social media platforms (Facebook, Twitter, YouTube) [571 links], German websites and non-functional links [2431 links] are filtered out. We also filter the instances with no URL in this step such as instance 2 in Code 4.2. As a result of filtering, only 5,590 instances are left with functional URLs. Code 4.4 presents 5 instances from SPEKTRUM data, and instance 2 has been filtered now.


```
{"1642530": {"Title": "Gesunde Hautflora schützt vor Neurodermitis",
             "Keywords": "atopisches ekzem, neurodermitis, hautbakterien, quorum sensing, s.
    aureus, staphylococcus, hautkeime, ",
             "Urls": {"https://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aat8329":
                     {"En_title": "Quorum sensing between bacterial species on the skin
    protects against epidermal injury in atopic dermatitis | Science Translational Medicine",
                      "Abstract": true,
                      "Structure": 1,
                      "Keyword": "1/7",
                      "Pdfs": ["/content/scitransmed/11/490/eaat8329.full-text.pdf"]}}},
 "1642462": {"Title": "Die unterschätzte Bedeutung von Nilpferdkot ",
             "Keywords": "nilpferd, flusspferd, hippo, kot, exkrement, dünger, gülle, ö
    kosystem, silizium, stofftransport, nahrungskette, kieselalgen",
             "Urls": {"https://doi.org/10.1126/sciadv.aav0395":
                     {"En_title": "Hippos (Hippopotamus amphibius): The animal silicon pump |
     Science Advances",
                      "Abstract": true,
                      "Structure": 5,
                      "Keyword": "4/12",
                      "Pdfs": ["/content/advances/5/5/eaav0395.full-text.pdf"]}}},
 "1642380": {"Title": "2000 Jahre altes Himmelsrätsel gelöst",
             "Keywords": "Astronomie, Nova, Chinesisch, Geschichte, Sternexplosion,
    historische Aufnahme, weier Zwerg", "Urls": {"https://arxiv.org/abs/1904.11515":
                    {"En_title": "[1904.11515] Discovery of an old nova remnant in the
    Galactic globular cluster M 22",
                     "Abstract": false,
                     "Structure": 1,
                     "Keyword": "1/7",
                     "Pdfs": []
                    },
                     "https://iopscience.iop.org/article/10.1088/0004-637X/756/2/107/meta":
```

```
                    {"En_title": "THE INTER-ERUPTION TIMESCALE OF CLASSICAL NOVAE FROM EXPANSION OF THE Z CAMELOPARDALIS SHELL - IOPscience",
                     "Abstract": true,
                     "Structure": 3,
                     "Keyword": "1/7",
                     "Pdfs": []}}},
 "1642290": {"Title": "Das Geheimnis der Ibis-Mumien",
             "Keywords": "Heiliger Ibis, Altes gypten, Archäologie, gyptologie, Biologie, Mumien, Bestattung, Opfer, Tempel",
             "Urls": {"https://www.biorxiv.org/content/10.1101/610584v1.full":
                     {"En_title": "Mitogenomic Diversity in Sacred Ibis Mummies sheds light on early Egyptian practices | bioRxiv",
                     "Abstract": true,
                     "Structure": 6,
                     "Keyword": "1/9",
                     "Pdfs": ["/content/10.1101/610584v1.full.pdf+html", "/content/10.1101/610584v1.full.pdf"]}}},
 "1642272": {"Title": "Mond könnte aus Lavameer entstanden sein",
             "Keywords": "Mond, Entstehung, junge Erde, Sonnensystem, Astronomie, Crash, Magma, Lava",
             "Urls": {"https://www.nature.com/articles/s41561-019-0354-2":
                     {"En_title": "Terrestrial magma ocean origin of the Moon | Nature Geoscience",
                     "Abstract": true,
                     "Structure": 3,
                     "Keyword": "2/8",
                     "Pdfs": ["https://static-content.springer.com/esm/art\%3A10.1038\%2Fs41561-019-0354-2/MediaObjects/41561_2019_354_MOESM1_ESM.pdf"]}}},
}
```


Listing 4.4: A sample of 5 instances after filtering and annotation for each URL.

4.1.5 Annotation of URLs

Each instance has the attributes Title (the German title from the SPEKTRUM data), Keywords (the German keywords from the SPEKTRUM data) and URLs (list of links from the SPEKTRUM data). To decide which link is the best fitting, all links are annotated by inspecting the data present on links as follows:

1. Title of the page (English).
2. Whether the words Abstract or Introduction are part of a heading of the page (such headings are an indicator for scientific texts).
3. Score for the scientific structure of the page (discussed in § 4.1.5.1).
4. Score for keyword matching (discussed in § 4.1.5.2).
5. List of all PDF links of the page.

4.1.5.1 Scientific Structure

This method checks text structure for scientific headings - Abstract, Introduction, Results, Discussion, References and Acknowledgements. The score ranges between 0 to 6 by assigning one point for each heading present in the text. The URLs that score four or higher are selected by assuming that a scientific article has at least four of the six headings.

4.1.5.2 Keyword Matching

This method uses the English translation of parsed German keywords from raw SPEKTRUM data and the English title of the HTML page to calculate a ratio for matched keywords. The ratio of matched keywords is defined as $\frac{m_k}{n_k}$ where m_k denotes the total number of keyword occurrences in the page title and n_k is the total number of German keywords. Code 4.4 shows 5 instances and their annotations from SPEKTRUM data.

4.1.6 Extraction of URLs

The URLs with positive scores were selected for the extraction. There are two types of source links: PDF and HTML. For the instances with PDF links, we use Beautiful Soup[2] for extraction, and Tika[3] for parsing the text. The HTML pages are downloaded via module request and extracted with Beautiful Soup[2]. The final extracted instances from PDF and HTML links are 3,554 in total with their German summaries. Code 4.5 presents the sample of 2 instances after extraction. The first instance will be discarded during cleaning as there is no English article exists in this instance. "[...]" denotes the continuation of text.

```
1 {"1642530": {"De_Summary": "<ARTICLE><TITLE>Gesunde Hautflora schützt vor Neurodermitis</
    TITLE><UNDERTITLE>Hilfreiche Bakterien stoppen zerstörerische Keime</UNDERTITLE><TEASER>
    Beim atopischen Ekzem stoen schädliche Hautbakterien in die Lücken, die die Krankheit
    gerissen hat - und sorgen so für Juckreiz und Rötung. Andere Keime könnten sie stoppen.</
    TEASER><SECTION><S> Die Ursachen des atopischen Ekzems- besser bekannt unter der
    veralteten Bezeichnung Neurodermitis- sind komplex und längst nicht geklärt. </S> <S> Die
     Krankheit dürfte häufig genetische Ursachen haben, die das Immunsystem verändern und Kö
    rper von Betroffenen auf äuere Einflüsse extremer reagieren lassen, mit dem
    charakteristisch starken Juckreiz oder Hautrötungen. </S> <S> Eine mitentscheidende Rolle
     könnten aber auch Mikroorganismen auf der Haut haben: Bei Erkrankten finden sich etwa
    auffällig groe Mengen von Staphylococcus aureus, die Toxine ausschütten und Entzü
    ndungsreaktionen verstärken. </S> <S> Forscher um Richard Gallo von der University of
    California in San Diego haben nun untersucht, warum diese Keime bei Betroffenen so schä
    dlich werden- und stellen fest, dass daran der Rest der Bakteriengemeinschaft auf der
    Haut nicht unschuldigist. </S> </SECTION>[...]</ARTICLE>",
2         "En_Article": "<ARTICLE><TITLE>Quorum sensing between bacterial species on the skin
    protects against epidermal injury in atopic dermatitis | Science Translational Medicine</
    TITLE></ARTICLE>"},
```

[2] https://pypi.org/project/beautifulsoup4/
[3] https://pypi.org/project/tika/

```
"1642462": {"De_Summary": "<ARTICLE><TITLE>Die unterschätzte Bedeutung von Nilpferdkot </
    TITLE><UNDERTITLE>Hippos arbeiten als Siliziumpumpe zwischen Land und Fluss</UNDERTITLE><
    TEASER>Nilpferde fressen, verdauen und scheiden aus - alles in gröerem Umfang. Ohne sie w
    ürde dem kosystem noch viele Kilometer flussabwärts etwas fehlen.</TEASER><SECTION><S>
    Wie jedes kosystem funktioniert auch die Landschaft der afrikanischen Flüsse und Seen ü
    ber den Austausch und die Beziehungen aller Lebewesen: Fällt nur ein Mitspieler aus, dann
     geht es den anderen schlechter. </S> <S> Dies unterstreicht eine internationale
    Forschergruppe am Hin und Her von Flusspferdkot. </S> <S> Der enthält Silizium aus den
    vom Hippos an Land verschlungenen Gräsern und fungiert als kaum verzichtbarer Dünger für
    wachsende Pflanzen. </S> <S> Ohne den typischen Lebensstil der Flusspferde als lebende
    Umwälzpumpen zwischen Land und Fluss würde der Rohstoff nicht in ausreichender Menge vom
    Land in die Flusssysteme gelangen. </S> </SECTION>[...]</ARTICLE>",
        "En_Article": "<ARTICLE><TITLE>Hippos (Hippopotamus amphibius): The animal silicon
    pump </TITLE><HEADING>Abstract</HEADING><SECTION><S> While the importance of grasslands
    in terrestrial silicon (Si) cycling and fluxes to rivers is established, the influence of
     large grazers has not been considered. </S> <S> Here, we show that hippopotamuses are
    key actors in the savannah biogeochemical Si cycle. </S> <S> Through a detailed analysis
    of Si concentrations and stable isotope compositions in multiple ecosystem compartments
    of a savannah-river continuum, we constrain the processes influencing the Si flux. </S>
    [...] </SECTION><HEADING>INTRODUCTION</HEADING><SECTION><S> Animals play an important
    role in the distribution of resources across landscapes, because of their capacity to
    ingest large quantities of food in a different location than that in which they egest,
    excrete, or die . </S> <S> This resource translocation has important effects on carbon
    and nutrient cycling, ecosystem productivity, and food web structure in both the source
    and recipient ecosystems . </S> <S> Studies of animals have predominantly focused on the
    biogeochemical cycles of the nutrients C, N, and P, but animals also move other elements
    that are essential to biological processes, such as silicon (Si). </S> [...] </SECTION
    >[...]</ARTICLE>"}}
```


Listing 4.5: A sample of 2 instances after extraction showing only chunks of complete articles.

4.1.7 Automatic and Manual Cleaning

After the extraction, the English articles are manually inspected to filter the incomplete extractions, garbage text, texts other than English, and shorter than the German summary. We manually cleaned the data by two annotators (native Germans with fluent English skills) over a period of two weeks. Following manual cleaning, the final data consists of 1,510 English articles and German summaries written by experts in science journalism.

4.1.8 Final Processing

The data is processed for lowercase conversion, word and sentence tokenization with NLTK toolkit[4]. The markup tags are used to preserve the structural information on the section and sentence levels. The final version of the dataset is stored in JSON format. Code 4.6 presents the structure of the final JSON file.

[4]https://pypi.org/project/nltk/

```
{"ID": {"De_Summary": "tagged and cleaned German summary", "En_Article": "tagged and cleaned
    English source article"}, ID": {"De_Summary": "tagged and cleaned German summary", "
    En_Article": "tagged and cleaned English source article"}, ...}
```

Listing 4.6: Final structure of SPEKTRUM cleaned dataset.

4.1.9 Spektrum Highlights

The final size after automatic and manual cleaning of the dataset is 1,510 instances only. The SPEKTRUM dataset is a high-quality Cross-lingual Science Journalism dataset, however, the size of the dataset makes it unsuitable for a supervised training paradigm. One possibility is that zero-shot or few-shot training of a model trained on the science journalism dataset. To the best of our knowledge, unfortunately, there are no such models exist that we could adopt for zero-shot or few-shot training. The other possible solution is to collect a similar-nature dataset for training purposes. We decided to opt for this solution and details of the other datasets are presented in the next section. There is another important aspect of the SPEKTRUM dataset that it cannot be published due to the magazine's policies for the research community.

4.2 Wikipedia Data Collection

Wikipedia is viewed as a reliable source for monolingual and multi-lingual data acquisition (Antognini and Faltings, 2020; Gholipour Ghalandari et al., 2020; Hättasch et al., 2020; Frefel, 2020). It maintains a consistent format for the articles[5]. The Wikipedia data is available in several forms for researchers, including data dumps, databases, DBpedia and Wiki-API[6]. These features make Wikipedia an ideal source for the research community.

Figure 4.3: An example of German Wikipedia page.

[5] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style
[6] https://meta.wikimedia.org/wiki/Research:Data

Figure 4.4: An example of English Wikipedia page.

4.2.1 Source

We select Wikipedia Science Portal (WIKIPEDIA) for scientific cross-lingual data collection. It is a popular, crowd-sourced science encyclopedia available in many languages. It is enormous in volume consisting of diversified topics such as biology, agriculture, technology, linguistics, *etc*. In 2019, the size of English Wikipedia was ≈6M, and the size of German Wikipedia was 2.4M in German. Figures 4.3 and 4.4 show an example of a Wikipedia article in two languages - German and English. These two articles are on the same topic; however, the articles are written independently. So these documents share important information about that topic but are not parallel in nature.

4.2.2 Wikipedia API

We use Wiki-API[7] for data collection, which provides an efficient way for extracting articles from a given category. It also facilitates traversing among the existing inter-language links of an article. It also provides user-friendly commands for extracting sections of the article. Figure 4.5 illustrates the data collection process from WIKIPEDIA, where WMS denotes Wikipedia Monolingual Set and WCLS denotes Wikipedia Cross-Linugal Set. The WIKIPEDIA articles are connected with inter-language links, which we used to get the German articles. Figure 4.5 demonstrates the split of a summary (lead) and a body (text) within an article. According to

[7] https://pypi.org/project/wikiapi/

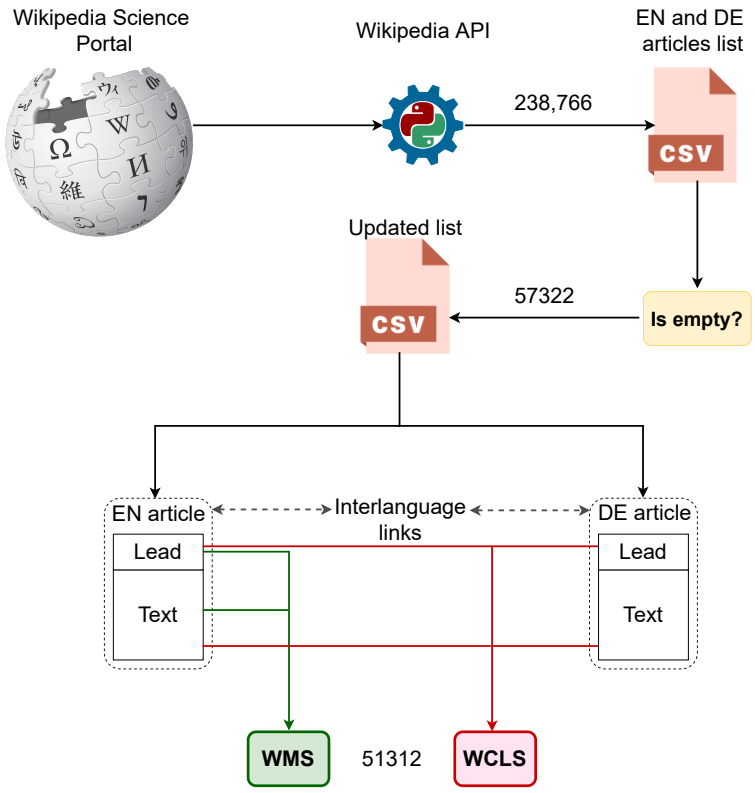

Figure 4.5: WIKIPEDIA data collection steps.

WIKIPEDIA's guidelines[5], the lead is the first paragraph of an article that summarizes the whole article. However, the WIKIPEDIA lead differs from the "lead" or "lede" used in news journalism that only contains important keywords and highlights (Giomelakis and Veglis, 2015).

4.2.3 List of Science Categories

Before extraction, we first create a list of science categories, subcategories and sub-subcategories. Code 4.7 shows the main categories extracted from English Wikipedia Science Portal. "..." denotes the continuation of the list. The main categories list consists of 151 different categories. Due to the limitations of Wiki-API, direct access to articles is not possible. An article can be directly accessed via a subcategory or sub-subcategory.

```
{"Acidbase chemistry", "Acoustics", "Agriculture", "Analytical chemistry", "Anatomy", "
    Anthropology", "Applied and interdisciplinary physics", "Applied mathematics", "Applied
    sciences", "Archaeology", "Architecture", ..., "Computational physics", "Computer science
    ", "Condensed matter physics", "Cosmology", "Cryobiology", "Crystallography", "Dentistry"
    , "Developmental biology", "Earth sciences", "Ecology", "Economics", "Education", "
    Engineering", "Environmental chemistry", "Environmental science", "Epidemiology", "
    Ergonomics", "Ethnobiology", "Ethnology", "Ethology", "Evolutionary biology", "
    Experimental physics",
```

```
... , "Technology", "Theoretical biology", "Theoretical chemistry", "Theoretical physics", "
    Thermodynamics", "Toxicology", "Veterinary medicine", "Volcanology", "Zoology"}
```

Listing 4.7: Main categories list in English Wikipedia Science Portal.

Therefore, the main categories list is further processed to extract subcategories and sub-subcategories (in some cases). The total size of the list is 13,482. Code 4.7 presents the list of subcategories.

```
{"Acidbase chemistry": {"Category:Acid Catalysts", "Category:Acidbase disturbances", "
    Category:Acidic oxides", ..., "Category:Acids in wine"},
"Acoustics": {"Category:Acoustic equations", "Category:Acoustic fingerprinting", "
    Category:Acoustic measurement", "Category:Acoustical engineers", ..., "Category:Acoustics
     journals", "Category:Acoustics software"},
"Agriculture": {"Category:Agnosia", "Category:Agrarianism", "Category:Agrarianists by
    continent", "Category:Agricultural buildings and structures by heritage register", "
    Category:Agricultural cooperatives", ... , "Category:Agricultural journals", "
    Category:Agricultural labor", "Category:Agricultural land development schemes", "
    Category:Agricultural law", "Category:Agricultural machinery", ...},
...,
"Zoology": {"Category:Zoological Societies", "Category:Zoologist stubs", "Category:Zoologists
     by century", "Category:Zoologists by field of research", "Category:Zoologists by
    nationality", "Category:Zoologists of medieval Islam", "Category:Zoologists with author
    abbreviations", ..., "Category:Zoology journal stubs", "Category:Zoology journals", "
    Category:Zoology museums", "Category:Zoology stub templates", "Category:Zoomusicology", "
    Category:Zwitterionic surfactants"}}
```

Listing 4.8: Subcategories list in English Wikipedia Science Portal.

4.2.4 List of Titles

Now we can access the articles, however, the list is prepared using English Wikipedia Science Portal. So we need to find which English articles are connected to German articles. For this purpose, we prepare a list of titles. We extract all English article titles by using the subcategories list. The total number of English titles is 238,766. We get the parallel German titles if they exist. We filter out that title from the list if no German article is present. We also retrieve the URLs of those titles. Code 4.9 shows the snippet of the titles list.

```
{"0": {"EN-Title": "Abdominal aortic aneurysm",
       "EN-Link": "https://en.wikipedia.org/wiki/Abdominal_aortic_aneurysm",
       "DE-Title": "Aortenaneurysma",
       "DE-Link": "https://de.wikipedia.org/wiki/Aortenaneurysma"},
"1": {"EN-Title": "Abdominal aortic plexus",
       "EN-Link": "https://en.wikipedia.org/wiki/Abdominal_aortic_plexus",
       "DE-Title": "Plexus aorticus abdominalis",
       "DE-Link": "https://de.wikipedia.org/wiki/Plexus_aorticus_abdominalis"},
"2":  {"EN-Title": "Abdominal external oblique muscle",
       "EN-Link": "https://en.wikipedia.org/wiki/Abdominal_external_oblique_muscle",
       "DE-Title": "Musculus obliquus externus abdominis",
       "DE-Link": "https://de.wikipedia.org/wiki/Musculus_obliquus_externus_abdominis"},
```

```
    ...,
"199743":  {"EN-Title": "Understanding Comics",
        "EN-Link": "https://en.wikipedia.org/wiki/Understanding_Comics",
        "DE-Title": "Comics richtig lesen",
        "DE-Link": "https://de.wikipedia.org/wiki/Comics_richtig_lesen"},
"199744":  {"EN-Title": "Under the Red Seas",
        "EN-Link": "https://en.wikipedia.org/wiki/Under_the_Red_Sea",
        "DE-Title": "Abenteuer im Roten Meer",
        "DE-Link": "https://de.wikipedia.org/wiki/Abenteuer_im_Roten_Meer"},
"199745":  {"EN-Title": "Underwater acoustic communication",
        "EN-Link": "https://en.wikipedia.org/wiki/Underwater_acoustic_communication",
        "DE-Title": "Akustische Unterwassertelefonie",
        "DE-Link": "https://de.wikipedia.org/wiki/Akustische_Unterwassertelefonie"},
...}
```

Listing 4.9: Titles list of English and German articles from Wikipedia Science Portal.

There is a possibility that an article might fall under two or more subcategories. In that case, that title is added only once.

4.2.5 Validation and Extraction

The titles list is further filtered out for empty criteria for both languages. Figure 4.5 shows how we split an article into lead and text. We take these two parts to define the empty criteria as follows:

- Title is present, but both the lead and text are missing.
- Lead is present and text is missing.
- Lead is absent and text is present.

We filter out all those instances that fulfill empty criteria, either for English or German or both. The final list is used to extract the articles. The total number of extracted pairs is 51,312.

4.2.6 Final Processing

The extracted articles are processed to remove the noise and white spaces. Then it is converted into lowercase and then tokenized for words and sentences with the NTLK toolkit[4]. The markup tags are used to preserve the structural information on the section and sentence levels. The final versions of the dataset are stored in JSON format. We create two versions of the WIKIPEDIA dataset for monolingual and cross-lingual parts as shown in Codes 4.10 and 4.11.

```
{"ID": {"En_Summary": "tagged and cleaned English lead", "En_Article": "tagged and cleaned
    English text"}, "ID": {"En_Summary": "tagged and cleaned English lead", "En_Article": "
    tagged and cleaned English text"},...}
```

Listing 4.10: Final structure of WIKIPEDIA monolingual dataset.

```
{"ID": {"De_Summary": "tagged and cleaned German lead", "De_Article": "tagged and cleaned
    German text", "En_Summary": "tagged and cleaned English lead", "En_Article": "tagged and
    cleaned English text",}, "ID": {"De_Summary": "tagged and cleaned German lead",
    "De_Article": "tagged and cleaned German text", "En_Summary": "tagged and cleaned English
     lead", "En_Article": "tagged and cleaned English text",}, ...}
```

Listing 4.11: Final structure of WIKIPEDIA cross-lingual dataset.

4.2.7 Manual Verification

Online-scraped datasets require manual inspection on a subset to verify their quality. However, due to its complexity, most dataset studies (Antognini and Faltings, 2020; Ladhak et al., 2020; Frefel, 2020) usually omit the manual verification of collected data. Only Hättasch et al. (2020) perform human verification on a subset of 39 summaries from three different parts of their dataset (Harry potter, English and German Star Wars) for one parameter of interest. We also adopt a similar approach to verify the cross-lingual part of the WIKIPEDIA dataset.

4.2.7.1 Guidelines for Verification

To verify the cross-lingual mappings of German summaries and English articles, we randomly select 20 articles from the cross-lingual WIKIPEDIA dataset. We ask two judges who are native German speakers with fluent English skills to annotate those articles for the following parameters.

1. **Relevance:** It determines if the German summary covers the main idea of the English article.
2. **Length:** We define the length criterion on the number of sentences. A summary of ≤ 2 sentences is considered short, otherwise long. The length criterion is crucial because our final objective is to summarize the SPEKTRUM dataset, and we want to have a similar dataset.

We define binary scores for both parameters. For relevance, 0 means the summary is irrelevant, and 1 means the summary is relevant to the given text. For length, 0 denotes a short summary, and 1 denotes a long summary.

4.2.7.2 Scores

For the relevance score, both judges agreed that German summaries are relevant to English articles. For the length, the sample German summaries get an average score of 0.74 with a substantial agreement (Fleiss's $\kappa = 0.76$) between judges. It is worth noting that short summaries

(≈25%) make this dataset quite challenging as it is considered as extreme summarization due to output size (Cachola et al., 2020).

4.2.8 Wikipedia Highlights

The final size of the dataset is 51,312 instances. The WIKIPEDIA dataset consists of two parts - monolingual and cross-lingual. The WIKIPEDIA dataset can be used for the following tasks: monolingual summarization, cross-lingual summarization and Cross-lingual Science Journalism. The WIKIPEDIA dataset is publically available and released under the Creative Commons Attribution-ShareAlike 3.0 Unported License for the research community[8].

4.3 Datasets Statistics

4.3.1 Split and Size

We split the WIKIPEDIA datasets with an 80/10/10 ratio for training, validation and testing sets with random sampling. We use the SPEKTRUM dataset only for zero-shot adaptability as a test set. The size of WIKIPEDIA training set is 41,049 articles. The validation and testing sets of WIKIPEDIA consists of 5,131 and 5,132 articles respectively. The size of the SPEKTRUM dataset is 1,510 articles. Each dataset consists of text and summary pair.

	WIKIPEDIA			**SPEKTRUM**	
	TEXT	EN-SUM	DE-SUM	TEXT	DE-SUM
	TR/VAL/TE	TR/VAL/TE	TR/VAL/TE	TE	TE
Vocabulary	22M/2.7M/2.7M	3.4M/.4M/.4M	2.9M/.3M/.3M	1M	.4M
Total words	64M/8M/7.9M	5.7M/.7M/.7M	4M/.5M/.5M	4.3M	.6M
Avg. words/text	1572/1562/1542	139/140/140	100/101/101	2337	361
Std.	1961/1935/1906	110/114/112	92/109/124	1510	250
Total sentences	2.5M/.3M/.3M	.2M/.03M/.03M	.2M/.02M/.02M	.19M	.03M
Avg. sentences/text	61/61/60	06/04/06	05/05/06	102	69
Std.	76/74/72	06/05/04	05/06/08	17	13
Compression ratio	–	20/20/20	17/18/17	–	30

Table 4.1: SPEKTRUM and WIKIPEDIA datasets statistics.

[8] https://github.com/MehwishFatimah/wsd

Table 4.1 provides statistics for these datasets. Each column of the WIKIPEDIA datasets presents the split of the train (TR), validation (VAL) and test (TE) sets. The first block presents the details of the total number of unique words or the size of the vocabulary. The second block shows the statistics related to words, the total number of words in a set, the average number of words per text and the standard deviation. The third block presents the details of the total number of sentences in a set, the average number of sentences per text and the standard deviation in a set. The last block shows the compression rate of each set.

4.3.2 Compression Ratio

The compression ratio is defined as the word ratio between a text and its summary (Grusky et al., 2018).

$$Compression = \frac{Summary\ word\ count}{Text\ word\ count} \tag{4.1}$$

Table 4.1 shows that the WIKIPEDIA English summaries are typically 20% of the length of the original English texts, while a German summary is 17.5% of the size of the original English texts. Interestingly, both languages belong to the Germanic family, however, the German language is famous for its inflection and compound words. As the compression ratio is calculated on the word counts, it is difficult to determine whether the English summaries are in fact longer than the German summaries or not. From Table 4.1, we find that the WIKIPEDIA splits have similar statistics, suggesting the proper distribution of features among the sets. While the SPEKTRUM set has a few differences from the WCLS test set, especially in compression ratio.

	1-gram	**2-gram**	**3-gram**	**4-gram**	**5-gram**
TR	24.6	69.3	87.6	92.1	93.0
VAL	24.5	69.1	87.4	91.9	92.7
TE	24.7	69.4	87.5	92.1	92.9

Table 4.2: Percentage of novel *n*-grams in WMS summaries.

4.3.3 Abstractiveness

To find out how much these summaries are abstractive in nature, we computed the percentage of *n*-grams that are unique (not present in the original text). The percentage of novel *n*-grams in the summaries serves as a measure of their abstractiveness (Grusky et al., 2018). Table 4.2 presents the percentage of *n*-grams for the WMS dataset, while the value of *n* ranges between $1 - 5$. We find that approximately 25% of uni-grams in summaries are not present in the

original text for each data split (TR, VAL and TE). However, the unique bi-grams percentage is quite high, suggesting almost 70% of the bi-gram combination in summaries is not used in the text. With the higher value of n, the percentage of unique n-grams is increased and reached 93% for 5-grams. From these scores, we infer that the WIKIPEDIA summaries are highly abstractive in nature.

4.4 Datasets Analysis

Analyzing the text language statistically have been deeply rooted in NLP (Aluisio et al., 2010; Hancke et al., 2012; Vajjala and Lučić, 2018; Mosquera, 2022; Weiss and Meurers, 2022). Initially, these statistical features have been developed for English. however, later many researchers adopted these features and some of the variations to analyze non-English texts (DuBay, 2004; Hancke et al., 2012; Weiss and Meurers, 2018, 2019). As a result, linguistic features have been proven to effectively work across languages and across corpora (Mollet et al., 2010; Stills, 2016; Przybyla-Wilkin, 2016; Hsiao et al., 2024). This section discusses the importance of data analysis using statistical features and how these features can be used on cross-lingual data analysis.

4.4.1 Assessment of Text Complexity

The assessment of text complexity is a crucial aspect in both educational and linguistic research. A lot of effort have been made in this regard for cross-language analysis. Anderson (1981) investigates the readability of English and non-English texts in the classroom using Läsbarhetsindex (LIX). LIX considers both sentence length and word length to assess text complexity and readability. The study compares the readability levels of texts in different languages, both English and non-English, by using LIX. This comparative analysis provides valuable insights into the readability of texts across languages. This also helps educators understand the linguistic challenges students may encounter in different language contexts.

Ghivirigâ (2012) investigates how English is used in scientific articles written in Romanian. The purpose of this investigation is to find out whether or not there is a substantial difference between the English used in these articles and that of native English researchers. This study created two corpora, one of Romanian-authored articles and one of native English-authored articles. They compared various linguistic features on these corpora, including first reference pronouns, relative pronouns, passive forms, modal verbs, and the use of tense-aspect forms. The results indicate that although Romanian writers use English quite well overall, there are some noticeable language variations that may affect readability and reception abroad.

Various linguistic features, such as lexical diversity, syntactic complexity, and cohesion, were compared between the two corpora. The results indicate that although Romanian writers use English quite well overall, there are noticeable language variations that may affect readability and reception abroad.

Przybyla-Wilkin (2016) explores readability in English, German, and Polish. The author investigates the difficulties associated with producing easily comprehensible texts in these languages, particularly for accessible communication. This study involves comparing the linguistic features and structures of texts among languages to assess readability levels. Readability formulas, such as the Flesch Reading Ease (FRE) and the Gunning Fog Index, evaluate factors like sentence length, syllable count, and word complexity to determine how easily a text can be comprehended.

Ciobanu and Dinu (2014) investigate the quantitative impact of translation on readability by evaluating translated texts with various readability features. The study employs quantitative metrics to assess changes in readability levels before and after translation across different languages, Romanian, French, Italian, Spanish and Portuguese. The research systematically evaluates readability with various linguistic features such as sentence structure, vocabulary complexity, lexical and morpho-syntactic patterns to quantify the extent of readability alterations induced by translation.

Tillman and Hagberg (2014) explores the compatibility of readability features applied to texts in different languages, specifically Swedish and English. Readability features aim to estimate the ease with which a text can be read and understood. The study tests three readability features: Coleman Liau index (CLI), LIX, and Automated Readability Index (ARI). These features are investigated on collections of Wikipedia articles, *On the Origin of Species* by Charles Darwin, and the Bible, along with their respective translations. The focus lies on Wikipedia articles and *On the Origin of Species* due to their extensive text and similar variables in both languages. The results show that ARI and LIX do well on passages that are less readable, whereas CLI does worse on such texts. However, CLI works well on easily readable texts. To conclude, CLI is appropriate for easy-to-read texts in both Swedish and English, ARI and LIX are effective for hard and moderate texts in both languages.

Friedrich et al. (2020) investigate how entropy operates within legal language across various languages and datasets. Entropy can be used to gauge the complexity and predictability of linguistic patterns within legal texts. They use BGH Zivilsenat, BGH Strafsenat, U.S. Supreme Court, EuroParl German and EuroParl English for legal text analysis.

Considering English and German, despite their linguistic differences, certain features of readability are universally applicable. Weiss et al. (2021) explores readability and and linguistic complexity for German and English texts. They categorized some linguistic features as

Universal Linguistic Features including lexical, syntactic, and morphological patterns. Lexical features, such as word frequency, word length, lexical diversity and readability indices, are essential indicators of vocabulary difficulty and text complexity across languages (Weiss and Meurers, 2018). Syntactic features, including sentence length, clause length, and syntactic tree depth, affect the cognitive load required to parse sentences and are therefore relevant to readability in both English and German (Hancke et al., 2012). These readability features are applicable in the same way to many languages, while some of them (*e.g.*, FRE) needs some calibrations for echoing the similar properties and results as in English readability (Hancke et al., 2012; Vajjala and Meurers, 2014; Weiss and Meurers, 2018; Bengoetxea and Gonzalez-Dios, 2021).

In the next sections, we cover the measures of lexical complexity and readability to analyze the similarities and differences between WIKIPEDIA and SPEKTRUM datasets.

4.4.2 Lexical Richness

Lexical richness refers to the quality, variability, and sophistication of the vocabulary used in a text or speech (Cohen et al., 2019). It is calculated by some form of the Type-Token Ratio (TTR). The number of tokens in a text fragment defines the word count, a simple metric. The number of types describes the number of unique words used in the text, referring to the size of the set of words used in that fragment. When combined, these scores define TTR, an indicator of lexical complexity. The TTR score is sensitive to textual length because when the text length increases, it is less probable that new types of tokens will be added (McCarthy and Jarvis, 2010). Therefore TTR is not considered a reliable indicator. Here we discuss lexical richness scores that are not sensitive to text lengths, such as the article and its summary. In the next sections, we will cover four lexical diversity measures: Hypergeometric Distribution Diversity (HDD), Measure of Textual Lexical Diversity (MTLD), Moving Average Type Token Ratio (MATTR), and Shannon Entropy Estimation (SEE). All these measures can be applied to both English and German, and are generally considered language-independent (McCarthy and Jarvis, 2010; Covington and McFall, 2010; Bentz et al., 2017). These measures assess the richness and variety of vocabulary in a text, providing insights into the complexity and readability of the text. For lexical richness, we use pythonic version of lexical richness formulas from the lexicalrichness library[9].

[9] https://pypi.org/project/lexicalrichness

4.4.2.1 Hypergeometric Distribution Diversity

The HDD feature (McCarthy and Jarvis, 2007) represents the probability of k random successes for a given feature in total n number of draws (trials) without replacement for a finite population of size N that contains total K objects with the observed feature. HDD depends on two characteristics:

- The outcome of each draw is mutually exclusive, either a success or a failure.
- Due to without replacement property, the probability of success changes with each draw because each draw decreases the success population.

To calculate the HDD for each lexical type in a text, the probability of encountering any of its tokens in a random sample of d words drawn from the text. The probabilities for all lexical types in the text are then added together to calculate the index. The HDD score calculates as:

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \tag{4.2}$$

where k is the probability of random successes for a given feature x in total n number of draws from a population of size N.

HDD applies statistical method to model the probability distribution of word frequencies in a text, and is tied to specific linguistic properties of a language, making it suitable for cross-linguistic analysis. It has been effectively used to compare lexical diversity in different languages (Bentz et al., 2017).

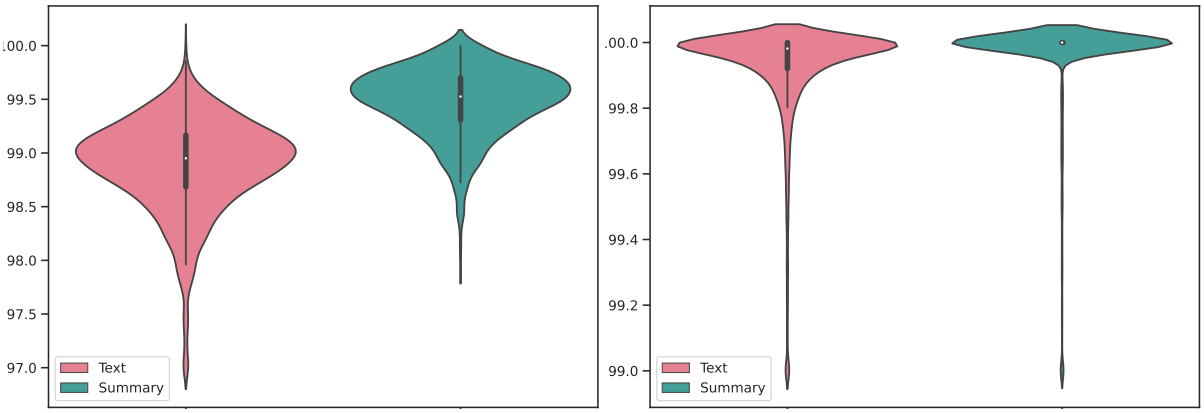

Figure 4.6: Distribution of HDD scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

Figure 4.6 presents the distribution of HDD scores in the SPEKTRUM and WIKIPEDIA sets. From these graphs, we find that SPEKTRUM summaries are the most lexically rich demonstrating the sophistication and high quality of those summaries. Interestingly, the WIKIPEDIA data is also lexically rich but quite skewed. We infer from this graph that the WIKIPEDIA data has high variability of lexicons, however, not well distributed depending on fields and topics.

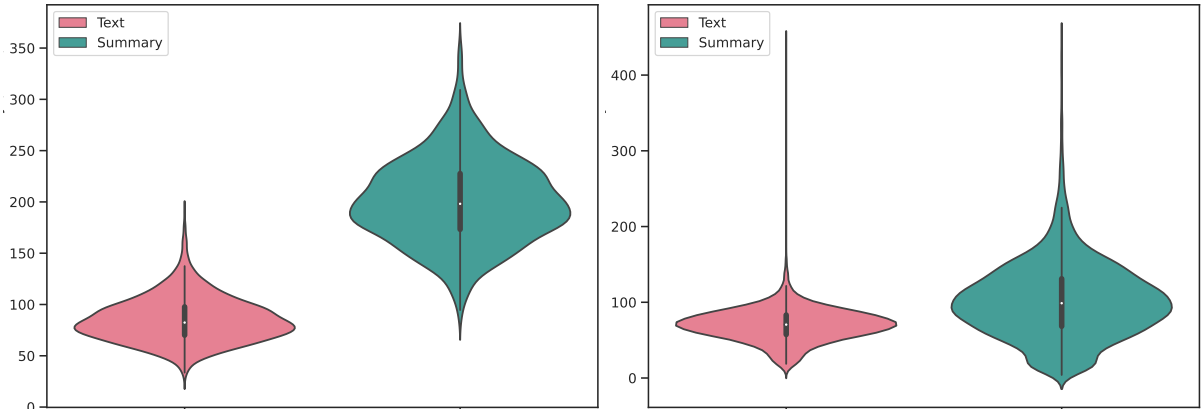

Figure 4.7: Distribution of MTLD scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

4.4.2.2 Measure of Textual Lexical Diversity

The MTLD feature (McCarthy, 2005) calculates lexical diversity with sequential text analysis for a given threshold (default = 0.72). The MTLD score reflects the average number of words in a certain TTR window. Then the TTR window starts again with the next token, and so on, until the last token of the given text is considered. Then, the length of the text is divided by the total number of TTR of 0.72 counted. The same process is repeated in the reverse direction from the last token to the first token. Then the final score is calculated by the average of the forward and backward passes.

MTLD provides a stable measure of lexical diversity that is less sensitive to text length variations. It calculates the average length of sequential word strings that maintain a certain level of TTR. MTLD has been validated across various corpora showing its robustness and applicability as a language-independent measure (McCarthy and Jarvis, 2010).

Figure 4.7 shows the distribution of MTLD in the SPEKTRUM and WIKIPEDIA sets. From these graphs, we interpret that SPEKTRUM and WIKIPEDIA summaries are more lexically rich than their corresponding texts, indicating more lexical variability. Interestingly, the SPEKTRUM summaries have the highest scores, affirming the results of HDD scores.

4.4.2.3 Moving Average Type Token Ratio

The MATTR feature (Covington and McFall, 2008) calculates lexical diversity by calculating TTRs for successive non-overlapping segments of a given text. The MATTR is a variation of TTR with a window (default = 25) for English texts and German summaries which also shows lexical richness in German text.

MATTR mitigates the text length sensitivity of traditional TTR and has been effectively used

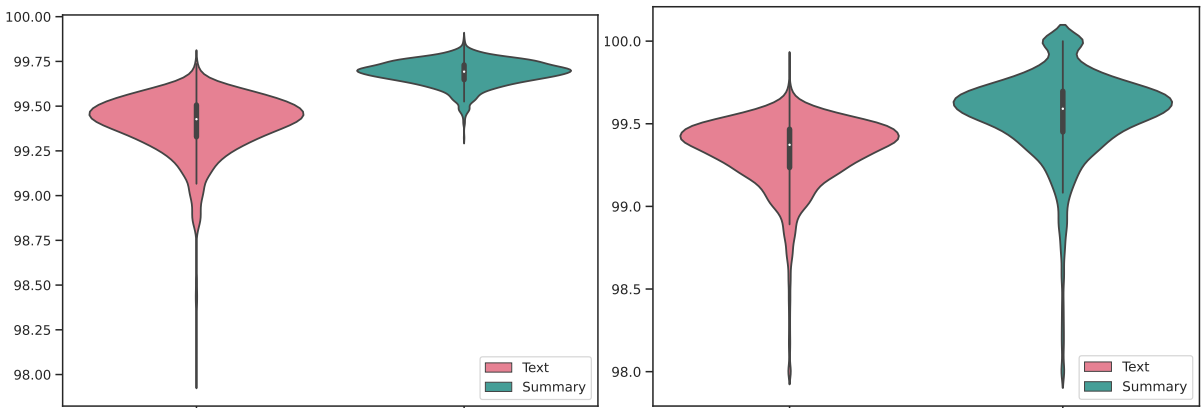

Figure 4.8: Distribution of MATTR scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

in multiple languages (Covington and McFall, 2010). Figure 4.8 presents the distribution of MATTR scores in the SPEKTRUM and WIKIPEDIA sets which also complements to the previous results.

4.4.2.4 Shannon Entropy Estimation

Shannon Entropy Estimation (SEE) refers to the "informational value" present in a text (Shannon, 1948). It is a language-dependent feature, and its value varies for different languages (Shannon, 1951). It is calculated with a frequency table as follows.

$$H(x) = \sum_{i=1}^{n} p(x_i) log_2 \frac{1}{p(x_i)} \tag{4.3}$$

where $H(x)$ is the total amount of information in an entire probability distribution. $P(x_i)$ refers to the frequency of a token appearing in the text, and $1/p(x)$ denotes the information of each case. For SEE, a lower score represents less lexical complexity.

SEE is a measure of unpredictability or information content in a text. It is calculated based on the probabilities of word occurrences and provides insights into the complexity and variability of vocabulary. SEE is inherently language-independent as it relies on statistical properties of word distributions rather than language-specific characteristics. Zanette (2014) investigates the statistical properties of written language, focusing on entropy and other measures that characterize text structure. Building on the foundational work of Shannon, the author explores how statistical methods can reveal underlying patterns in written text. Shannon entropy has been widely used for textual analysis across diverse languages (Largeron et al., 2011; Dye, 2017; Friedrich et al., 2020; Liu et al., 2022).

Figure 4.9 shows higher SEE scores in English texts and WIKIPEDIA German summaries, while SPEKTRUM summaries tend to have lower scores suggesting these have less lexical complexity.

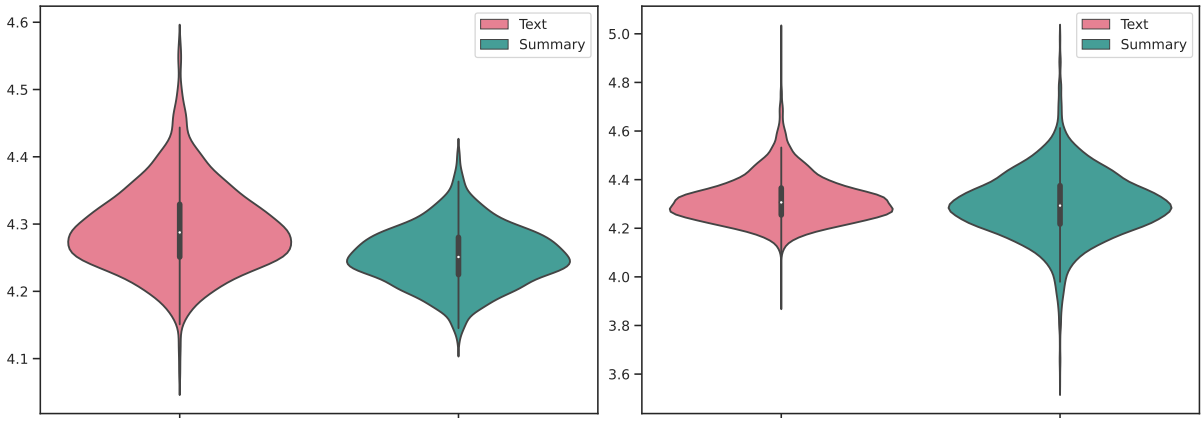

Figure 4.9: Distribution of SEE scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

4.4.3 Readability

Readability indices are computational tools that assess the comprehension levels of a given text. These indices provide quantitative measures of text complexity and ease of understanding, aiding educators, writers, and researchers in evaluating and improving written materials for various audiences. In the next sections, we will cover four readability scores: Flesch Kincaid Reading Ease (FRE), Lensear write formula (LWF), Coleman Liau Index (CLI), and Automated Readability Index (ARI). All these scores except FRE can be applied to both English and German as the same. FRE has constant values in its formula for both English and German because it is designed to provide a standardized measure of readability based on linguistic features common to both languages. The constants in the formula are chosen to balance the influence of sentence length and word length, which are critical factors affecting readability. LWF is also a syllable-based score, however, it is simpler and more adaptable, using counts of easy and hard words and sentence length, without the need for language-specific calibration. CLI and ARI are based on character and sentence counts, which are universally applicable, allowing them to be used across different languages without re-calibration. So, except FRE, these scores are generally considered language-independent scores (Tillman and Hagberg, 2014; Przybyla-Wilkin, 2016; Ciobanu and Dinu, 2014). For readability scores, we use pythonic version of readability formulas from the textstat library[10].

[10]https://pypi.org/project/textstat/

4.4.3.1 Flesch Reading Ease

FRE (Kincaid et al., 1975) is a syllable-based score as it indicates easy and hard readability. Originally developed for English, it uses syllables and sentence length to estimate readability. For cross-language comparison, you might need to adjust the constants based on the target language's syllable structure and average sentence length. The formula for English is given below:

$$FRE_{EN} = 206.835 - (1.015 \times ASL) - (84.6 \times ASW) \tag{4.4}$$

where Average Sentence Length (ASL) is the number of words divided by the number of sentences in the text. The average number of syllables per word (ASW) is the number of syllables divided by the number of words in the text. The numeric values are language-dependent constants. These constants were empirically derived to reflect the ease or difficulty of reading texts at various levels of complexity. The constants ensure that the formula can be applied consistently across different texts within the same language and, with adjustments, across different languages. For German, the formula is calibrated but retains the same structure to ensure comparability. The modified formula for German is given below:

$$FRE_{DE} = 180 - ASL - (58.5 \times ASW) \tag{4.5}$$

For FRE, a higher score represents easy comprehension. FRE calculates readability based on the average number of syllables per word and the average number of words per sentence. While the basic structure of FRE is retained for German, the constants are adjusted to reflect the specific linguistic characteristics of German. For example, German tends to have longer compound words and different syntactic structures, which can affect readability differently compared to English. The adjusted constants ensure that the readability scores accurately reflect these differences.

FRE provides a score that indicates the ease or difficulty of reading a text, with higher scores denoting easier readability. For instance, a text with an FRE score of 70 in English can be compared to a text with an adjusted FRE score in German to gauge relative readability levels. Overall, the use of constants in the FRE formula helps standardize the measurement of readability, making it a versatile tool for assessing text complexity in both English and German (Przybyla-Wilkin, 2016). Adjustments to these constants for different languages account for linguistic differences while maintaining the formula's core function.

Figure 4.10 shows the plots of FRE scores suggesting that the SPEKTRUM summaries are comparatively easier than the English texts. Interestingly, the WIKIPEDIA summaries tend to be comparatively harder than their corresponding text. It seems that the nature of WIKIPEDIA

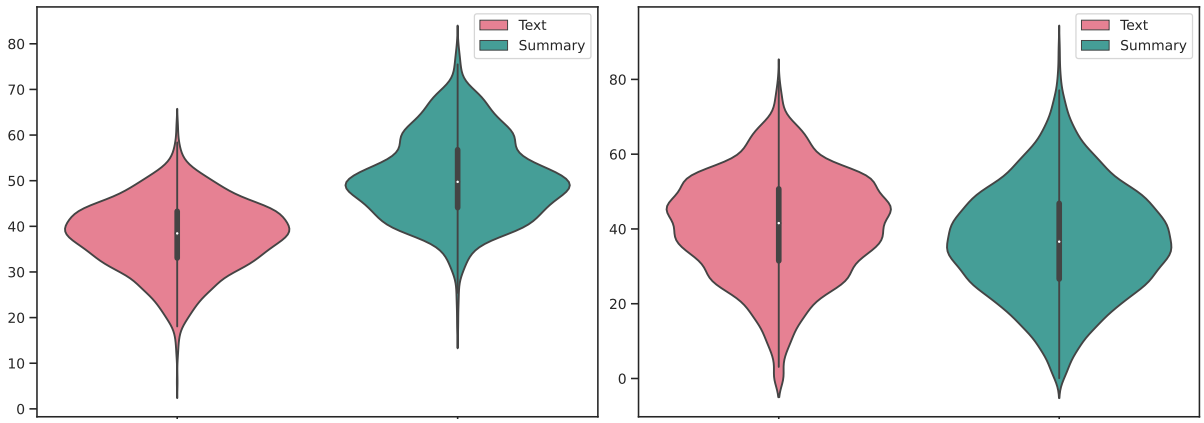

Figure 4.10: Distribution of FRE scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

texts (being more technical or specific) could lead to higher syllable counts in summaries compared to the more generalized SPEKTRUM summaries. Furthermore, FRE is a well-established readability score, however, it calculates on the syllables which can be misleading in some cases (Coleman and Liau, 1975).

4.4.3.2 Lensear Write Formula

LWF (Plavén-Sigray et al., 2017) is comparatively a new measure. It calculates the score by finding the proportion of easy and complex words, sentence length. LWF takes a sample of 100 words and divides them into easy words ($\leq$2 syllables) and hard words ($\geq$3 syllables) scores as follows.

1. For each easy word, add 1 point in r.
2. For each easy word, add 3 points in r.
3. Divide r by the number of sentences in the sample.
4. Adjust the result: if $r > 20$ then LWF $= r/2$, if $r \geq 20$ then LWF $= r/2 - 1$.

LWF is inherently simpler and more adaptable to different languages since it does not rely on specific constants that need adjustment. The classification of words as "easy" or "hard" based on syllable count is a relatively universal concept that does not change significantly between languages. This makes LWF more straightforward and universally applicable for cross-language comparisons without requiring detailed linguistic calibration.

For LWF, a higher score is considered better. Usually, the recommended easy score for an adult reader is between $70 - 80$. Figure 4.11 shows the distribution of LWF scores in the SPEKTRUM and WIKIPEDIA sets. Interestingly, the LWF results mirror the FRE findings for the

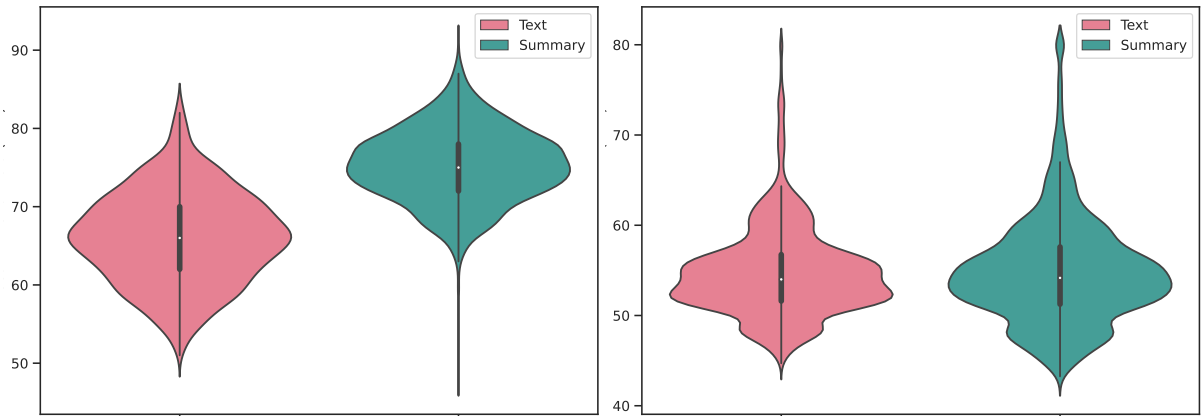

Figure 4.11: Distribution of LWF scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

SPEKTRUM summaries, clearly indicating that the SPEKTRUM summaries are easier to read than the input texts. However, these results also suggest that WIKIPEDIA summaries are comparatively easier than their texts. Moreover, these results suggest that the WIKIPEDIA data has hard readability than the SPEKTRUM data. These results suggest that LWF can identify simpler word usage in WIKIPEDIA summaries, aligning with practical readability improvements not captured by FRE.

4.4.3.3 Coleman Liau Index

CLI (Coleman and Liau, 1975) is a readability score that computes the reading level of a text by using sentences and letters and does not rely on syllables. It is calculated as follows:

$$Coleman\ Liau\ Index = 5.88 \times \frac{L}{W} - 29.6 \times \frac{S}{W} - 15.8 \tag{4.6}$$

where L is the total number of characters (including numbers and punctuation), W is the total number of words, and S is the total number of sentences in a given text.

CLI computes the score by taking average number of letters per 100 words and average number of sentences per 100 words. The CLI formula relies on character counts (letters and sentences) rather than syllables or word complexity, which are less language-dependent. This makes the CLI applicable across different languages without needing specific calibrations.

Figure 4.12 shows the distribution of CLI scores in the SPEKTRUM and WIKIPEDIA sets. The CLI scores are interpreted as a lower score means better readability. These results suggest that the German summaries are comparatively harder than their English texts. CLI relies on the number of letters and sentences, which might differ significantly between languages due to word length and sentence structure variations. German typically has longer compound words,

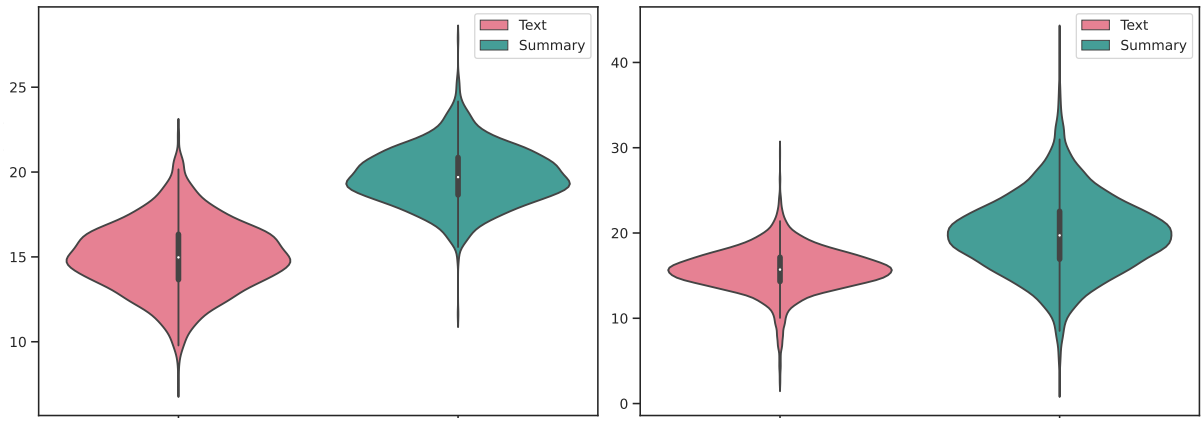

Figure 4.12: Distribution of CLI scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

which can result in higher letter counts per word, affecting CLI scores and making summaries appear to be harder to read.

4.4.3.4 Automated Readability Index

ARI (Senter and Smith, 1967) is calculated from ratios of word difficulty (number of letters per word) and sentence difficulty (number of words per sentence). The ARI score is computed as follows:

$$Automated\ Readability\ Index = 4.71 \times \frac{L}{W} + 0.5 \times \frac{W}{S} - 21.43 \tag{4.7}$$

where L is the total number of characters (including numbers and punctuation), W is the total number of words, and S is the total number of sentences in a given text.

ARI computes average number of characters per word and average number of words per sentence to determine readability. Similar to CLI, ARI uses character and word counts, making it less sensitive to specific linguistic structures of different languages. The use of characters and words per sentence as metrics allows ARI to be applied across various languages without re-calibration, since these elements are common across languages.

Figure 4.13 shows the distribution of ARI scores in the SPEKTRUM and WIKIPEDIA sets. The ARI scores are also interpreted as a lower score means better readability. Interestingly these scores suggest that the German summaries are easier than their English texts are easy to read. These results also indicate the similar mean value of the SPEKTRUM and WIKIPEDIA summaries (≈ 15). ARI considers both characters per word and words per sentence. If German summaries use shorter sentences, this might lower ARI scores despite longer words. Differences in the nature of summaries and texts might impact ARI differently than CLI, possibly capturing simpler sentence structures in German summaries.

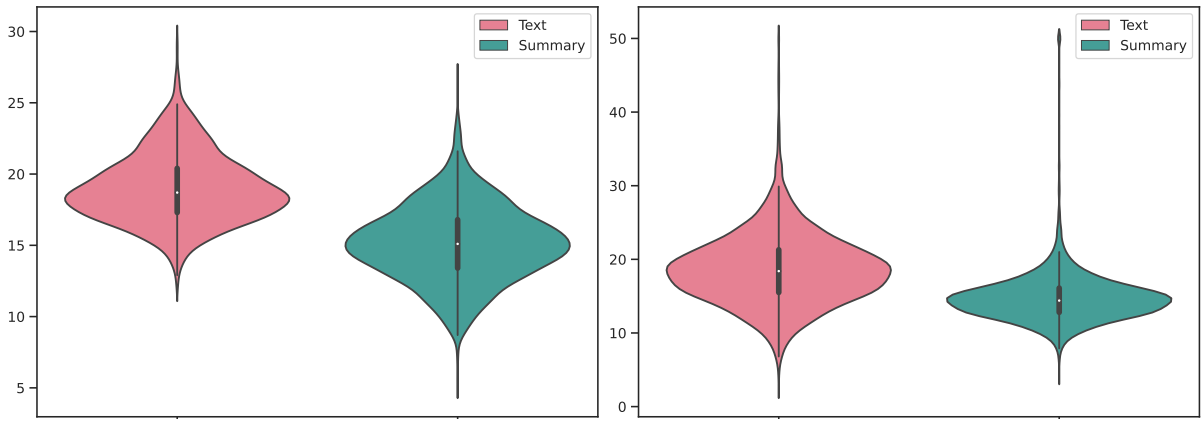

Figure 4.13: Distribution of ARI scores in the SPEKTRUM (left) and WIKIPEDIA (right) sets.

4.4.4 Discussion

Integrating various lexical richness and readability scores enriches the analysis by providing a comprehensive assessment of text complexity from multiple dimensions. Here's how the inclusion of different scores enhances the comprehensiveness of the analysis:

1. Lexical richness scores: Assess the variability and sophistication of vocabulary in texts. They indicate how diverse and rich the language use is, offering insights into the linguistic complexity and depth of expression.
2. Readability scores: Measure the ease of comprehension based on factors like sentence structure, word complexity, and syntactic clarity. They provide quantitative metrics to gauge how accessible and understandable texts are to readers.

Examining both lexical richness and readability helps gaining a holistic understanding of text complexity. Lexical richness scores reveal the breadth and depth of vocabulary usage, while readability scores highlight the overall readability and ease of understanding. Furthermore, each readability index has distinct metrics (syllables, word lengths, sentence lengths) leading to different sensitivity levels to language-specific features and summarization changes. FRE is highly sensitive to syllables, making it less reliable across languages without calibration. LWF's focus on easy and hard words offers a more consistent measure across languages. CLI and ARI's reliance on character counts can vary with language structure but offer a different perspective from syllable-based indices. Different types of content (encyclopedic vs. scientific articles vs. general non-expert summaries) inherently have different readability profiles, influencing the scores differently.

Based on these results, we find that LWF and ARI are less sensitive to language-specific nuances like word length differences, making them more adaptable for cross-linguistic comparisons without calibration. While FRE and CLI require adjustments for languages with distinct linguistic features (*e.g.*, compound words in German), ensuring accurate readability assessments. In conclusion, integrating different lexical richness and readability scores provides a comprehensive framework for analyzing text complexity.

4.5 Datasets Evaluation

We conduct an empirical evaluation for two tasks - monolingual summarization and cross-lingual summarization.

4.5.1 Experiments

4.5.1.1 Datasets

For monolingual summarization, we use the WMS dataset with an 80-10-10 split ratio. For cross-lingual summarization, we use the WCLS dataset with the same split ratio and SPEKTRUM dataset as our case study.

4.5.1.2 Models

For monolingual summarization, we use extractive methods as baselines: (i) SUM-BASIC, (ii) LUHN, (iii) KLSUM, (iv) LSA, (v) LRANK, (vi) TRANK, and (vii) BERT[11,12]. We train three existing abstractive summarization models: (i) Attention-based S2S model (S2S) (Bahdanau et al., 2015), (ii) Pointer Generator Network (PGN) (See et al., 2017), and (iii) Transformer-based S2S model (TRF) (Vaswani et al., 2017). We select these models because these models show good results in previous studies See et al. (2017); Ouyang et al. (2019); Duan et al. (2019). For cross-lingual summarization baselines, we create two existing pipelines - Trans-Sum and Sum-Trans to create synthetic cross-lingual data from WMS with FairSeq[13].

4.5.1.3 Training and Inference

4.5.1.3.1 Libraries and Hardware We run all abstractive models with Pytorch[14] on a single Tesla P40 GPU with 24GB RAM.

[11] (i-vi) https://pypi.org/project/sumy/
[12] (vii) https://pypi.org/project/bert-extractive-summarizer/
[13] https://github.com/pytorch/fairseq
[14] https://pytorch.org/

4.5.1.3.2 Hyperparameters We train S2S and PGN models with almost the same settings as in See et al. (2017). The size of word embeddings is 128 dimensions with hidden layers of 256 dimensions. The vocabulary size is 100K and 50K at the encoder and decoder sides, without the OOV words handling as used in the PGN model. However, we choose BPE instead of the *n*-gram vocabulary to solve the OOV words[15]. We use the Adam optimizer with a learning rate of 0.15 and a mini-batch of size 16. We train all models for 40 epochs and the validation loss is our metric to determine the best-trained model.

We train TRF with the same hyper-parameters settings as in Vaswani et al. (2017). The word embeddings have dimensions of 512 and hidden layers have dimensions of 786. The model consists of six stacks of encoders and decoders, each having 6 layers and 8 multi-attention heads at the decoder side. To make the results comparable among all models, we opt for the same vocabulary size of 100K and 50K for this model. We use the Adam optimizer with a learning rate of 0.0001 and a residual dropout of 0.1. For all abstractive models, the encoder and decoder length is fixed to 400 and 100 words as in See et al. (2017) and a beam search of size 4 is applied in the inference phase.

4.5.1.3.3 Training Time The S2S and TRF models take around 6 days, and the PGN model takes 3 days for training and inference.

4.5.1.4 Evaluation

For automatic evaluation, we report precision, recall and F-scores of ROUGE-1, ROUGE-2 and ROUGE-L. We also conduct a human evaluation to investigate two linguistic properties - **relevance** and **fluency**. The previous monolingual scientific summarization studies (Cohan et al., 2018; Dangovski et al., 2019) have not considered human evaluations due to its demanding nature. For human evaluation of scientific articles, human judges must read and comprehend long domain-specific articles with summaries to evaluate the linguistic qualities of system summaries. It is more challenging to conduct cross-lingual evaluations as it requires bilingual comprehension for articles tailored to various science topics.

4.5.2 Monolingual Results

Table 4.3 presents the monolingual summarization results of the WMS dataset. All the results are the average of three runs for each model. The BERT model achieves the highest scores for ROUGE-1 and ROUGE-2 , whereas Sum-Basic performs well for ROUGE-L. Overall, all

[15]These experiments were conducted in 2019 and at that time n-gram fixed size vocabulary was in practice with neural models.

MODELS	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
EXTRACTIVE									
SUM-BASIC	28.67	22.82	38.68	07.15	05.53	10.12	**25.24**	20.06	34.02
TRANK	26.29	18.82	43.57	07.11	04.86	13.27	22.98	16.43	38.19
KLSUM	24.96	17.73	42.13	06.40	04.42	11.58	21.64	15.35	36.65
LUHN	25.67	19.25	38.50	06.75	04.86	11.03	22.54	16.88	33.89
LRANK	26.53	20.11	38.95	06.69	04.92	10.47	23.22	17.58	34.18
RD	28.05	21.20	41.45	**07.51**	05.44	12.13	24.55	18.52	36.39
LSA	26.51	18.97	44.01	07.40	05.03	13.96	23.09	16.50	38.45
BERT	**28.74**	23.56	36.83	**07.51**	06.02	09.98	25.03	20.51	32.10
ABSTRACTIVE									
PGN	22.25	47.88	14.49	05.34	11.90	03.44	20.58	44.52	13.38
S2S	20.94	54.98	12.93	04.75	11.67	02.98	19.31	51.40	11.89
TRF	**25.53**	40.95	18.55	**06.29**	07.03	05.69	**22.76**	36.83	16.47

Table 4.3: Monolingual summarization results on the WMS dataset. The best results are marked as **Bold**.

extractive models yield similar results and abstractive models perform fairly well, however slightly lower than the extractive models. In general, all the summarization methods perform worse for ROUGE-2 than ROUGE-1 and ROUGE-L.

We consider two factors while comparing extractive and abstractive results: (i) the impact of novel *n*-grams in the reference summaries, and (ii) the length of output summaries. Regarding the impact of novel *n*-grams, extractive methods are not impacted by the presence/absence of novel *n*-grams. For example, if we consider novel unigrams, as mentioned in Table 4.2, approximately 25% of the summary unigrams are not present in the corresponding text, but the remaining 75% unigrams can overlap. As the extractive methods extract the sentences from the actual text and maintain a good percentage of overlapped words. However, as abstractive models do not rely on extraction, their results can be influenced by the presence/absence of novel *n*-grams.

Moreover, we note that the extractive models have higher recall than precision, indicating that the system summaries are longer than the reference summaries. However, precision is higher in the abstractive models indicating that the system summaries are shorter than reference summaries. Ideally, the system summaries should be similar to reference summaries. However, since the models have been only designed and evaluated on news datasets, they tend

Source: united nations conference on sustainable development background in 1992, the first conference of its kind, the united nations conference on environment and development (unced), commonly referred to as the rio conference or earth summit, succeeded in raising public awareness of the need to integrate environment and development. the conference drew 109 heads of state to rio de janeiro, brazil, to address what were dubbed urgent problems of environmental protection and socio-economic development. [...]
Target: the united nations conference on sustainable development (uncsd), also known as rio 2012, or earth summit 2012 was the third international conference on sustainable development aimed at reconciling the economic and environmental goals of the global community. hosted by brazil in rio de janeiro from 13 to 22 june 2012, [...]
BERT: united nations conference on sustainable development background in 1992, the first conference of its kind, the united nations conference on environment and development (unced), commonly referred to as the rio conference or earth summit, succeeded in raising public awareness of the need to integrate environment and development. [...]
LRANK: the world conference on human rights, for example, focused on the right of people to a healthy environment and the right to development; controversial demands that had met with resistance from some member states until the earth summit. it also created new international institutions, among them the commission on sustainable development, tasked with the follow-up to the rio conference and led to the reform of the global environment facility. [...]
PGN: the united nations environment programme (unep–1898) is an international non-governmental organization devoted to the united nations convention on climate change in the united states. it is the world's fourth-largest programme in the united states and **the united nations environment programme (unep) and the united nations environment programme (unep)**. the international commission on climate change is the **world's fourth-largest programme.**
TRF: the first conference of its kind, the united nations conference on environment and development (unced) was held in 1992. the conference drew heads of state to rio de janeiro, brazil to address what were dubbed urgent problems of environmental protection and socio-economic development. the earth summit influenced subsequent un conferences and set the global green agenda.

Figure 4.14: An example of monolingual system summaries from the WMS outputs. New words are marked as Blue, incorrect information is marked as Red, extractive parts are marked as Orange and repetition are marked as **Bold**.

to produce short summaries.

4.5.2.1 Example

Figure 4.14 shows an example of monolingual generated and reference summaries. In Figure 4.14, we find that both PGN and TRF produce good summaries, however, PGN tend to produce more repetitions while TRF tend to have more extractive chunks.

4.5.3 Cross-lingual Results

Table 4.4 presents the results of cross-lingual summarization. All the results are the average of three runs for each model. We cannot compare our results with those of recent cross-lingual

summarization studies since they used synthetic cross-lingual data from the news domain. We overcome this problem by using two baselines, Trans-Sum and Sum-Trans, which have been used in recent studies.

MODELS	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
Trans-Sum									
S2S	14.18	30.49	09.24	01.51	02.55	01.07	12.77	27.67	08.30
PGN	15.81	31.35	10.57	02.86	05.58	01.92	14.69	29.14	09.82
TRF	16.15	32.56	11.41	03.66	05.84	02.72	15.25	32.06	09.29
Sum-Trans									
S2S	15.04	32.85	09.75	01.48	02.56	01.04	13.64	28.03	08.82
PGN	18.24	28.62	13.38	04.14	10.98	02.55	16.04	30.45	11.11
TRF	19.31	26.77	14.18	04.23	11.67	02.84	17.37	31.74	12.18
WCLS									
S2S	18.37	37.93	12.12	04.04	09.91	02.54	16.55	34.57	10.88
PGN	20.72	30.34	15.73	03.79	05.93	02.79	18.68	27.48	14.15
TRF	**21.61**†	26.81	18.10	**04.37**†	05.16	03.79	**18.10**†	22.42	15.18
SPEKTRUM									
S2S	16.47	26.42	11.97	03.42	03.43	03.41	11.87	25.47	07.74
PGN	18.64	29.74	13.54	03.83	04.05	03.63	15.65	26.42	11.12
TRF	**20.81**†	31.39	14.47	**04.19**	05.43	03.41	**17.54**†	21.73	15.29

Table 4.4: Cross-lingual summarization results on the WIKIPEDIA and SPEKTRUM datasets. The best results are marked as **Bold**. † denotes a significant improvement ($p < 0.05$).

Among the baselines, the Sum-Trans models perform better than the Trans-Sum models. However, these baseline models do not perform well than the direct cross-lingual models. The direct models show significant improvements on the WCLS dataset than the Sum-Trans and Trans-Sum models ($p < 0.05$). These results suggest that machine translation introduced noise to synthetic cross-lingual data, which acts as a bias and affects the performance of these models. The cross-lingual summarization models learn the mappings between encoder and decoder sides language distributions along with compression. Therefore, distortion in language distributions (*e.g.*, wrong translated tokens, UNK tokens) can affect the generalization of those mappings. Overall, the abstractive models perform well by learning the structural mappings between English and German languages and tend to maintain a logical structure of sentences in summaries.

Source: d' arrest discovered ngc525 using his 11-inch refractor telescope at copenhagen. he located the galaxy 's position with a total of two observations. as he also noted the mag 11-12 star just 2' northwest, his position is fairly accurate. the galaxy was later catalogued by john louis emil dreyer in the new general catalogue, where it was described as very faint, very small, 11th or 12th magnitude star 5 seconds of time to west. the galaxy appears very dim in the sky as it only has an apparent visual magnitude of 13.3 and thus can only be observed with telescopes. [...]
Target: ngc525 ist eine linsenförmige galaxie vom hubble - typ s0 im sternbild fische auf der ekliptik. sie ist schätzungsweise 99 millionen lichtjahre von der milchstraße entfernt und hat einen durchmesser von etwa 40. 000 lichtjahren. im selben himmelsareal befinden sich u.a. die galaxien ngc516, ngc524, ic101, ic102. das objekt wurde am 25. september 1862 von dem deutsch-dänischen astronomen heinrich ludwig d' arrest entdeckt. **Translation:** ngc525 is a lenticular galaxy of the hubble type s0 in the constellation pisces on the ecliptic. it is estimated to be 99 million light years from the milky way and about 40,000 light years across. the galaxies ngc516, ngc524, ic101, ic102 are located in the same area of the sky. the object was discovered on september 25,1862 by the german danish astronomer heinrich ludwig d' arrest.
Trans-Sum-PGN: ngc142 ist eine unregelmäßige galaxie im sternbild eridanus. sein d <unk> l wurde mit hilfe der kugelsternhaufenluminosität auf 31,01 ± 0,21 geschätzt, was etwa 52 ms entspricht. er ist das hellste <unk> <unk>. **Translation:** ngc142 is an irregular galaxy in the constellation of eridanus. its d<unk>l was estimated with the help of the globular cluster luminosity to 31. 01 ± 0. 21, which corresponds to about 52 ms. it is the brightest <unk> <unk>.
WCLS-PGN: ngc499 ist eine elliptische galaxie mit aktivem galaxienkern vom hubble - typ e0 im sternbild fische am nordsternhimmel. sie ist schätzungsweise 22 millionen lichtjahre von der milchstraße entfernt und hat einen durchmesser von etwa 70.000 lichtjahren. **Translation:** ngc499 is an elliptical galaxy with an active galaxy core of the hubble type e0 in the constellation pisces in the north star sky. it is an estimated 22 million light years from the milky way and has a diameter of around 70,000 light years.
WCLS-TRF: d' arrest entdeckte ngc990 mit seinem 11 - zoll - refraktorteleskop. die galaxie wurde von john ratter im neuen katalog katalogisiert, wo sie als sehr kleiner stern beschrieben wurde. **Translation:** d' arrest discovered ngc990 with its 11 - inch refractor telescope. the galaxy was cataloged by john ratter in the new catalog, where it was described as a very small star.

Figure 4.15: An example of cross-lingual system summaries. New words are marked as Blue, incorrect information is marked as Red, extractive parts are marked as Orange and repetition are marked as **Bold**.

4.5.3.1 Example

Figure 4.15 shows an example of cross-lingual generated and reference summaries. We find that both PGN and TRF produce good summaries, however, PGN tends to produce more errors for numeric embeddings, while TRF tends to have translated extractive chunks.

4.5.3.2 Case Study: SPEKTRUM

We extend our experiments with the SPEKTRUM dataset. In these experiments, we examine how on-the-ground cross-lingual summarization models perform on a real-world dataset. The cross-lingual summarization models under-perform on the SPEKTRUM set ($p < 0.05$). The

slight drop in performance is because the decoder is a conditional model that learns contextual representations from training data. Moreover, it seems that BPE vocabulary caters to the unseen words of SPEKTRUM set, as both test sets (WIKIPEDIA and SPEKTRUM) have not been used for vocabulary construction. All models perform poorly with ROUGE-2 than ROUGE-1 and ROUGE-L. Due to the short summaries produced by the models, precision is higher than recall, affecting the F-score. Earlier, we noted that these models are designed for news datasets, which do not require long summaries. We selected these neural models because they have performed well in machine translation and summarization. However, their implementation has not been tested for cross-lingual scientific texts that are comparatively much longer than the previously used datasets.

4.5.4 Human Evaluation

To evaluate the models, we randomly select 20 output summaries, their reference summaries, and the input articles (10 from PGN_MS and 10 from PGN_CLS). We evaluate these summaries for **relevance** and **fluency** on a scale of $1-3$ as used by Ouyang et al. (2019). The average score for **relevance** is 2.10, and **fluency** is 2.65, with a moderate agreement (Fleiss's κ = 0.60 and 0.58) between judges for monolingual summaries. Whereas the average score for **relevance** is 1.65 and **fluency** is 1.96, with a substantial agreement (Fleiss's κ = 0.70) for both scores between judges for cross-lingual summaries. We infer from these results that these abstractive models retain an appropriate structure of the output summaries, however, tend to miss some relevant information. The cross-lingual models tend to produce irrelevant content in some summaries.

4.5.5 Limitations

We conclude this section with the following findings: (1) the existing abstractive models struggle with longer inputs as these models are mostly evaluated with news datasets that are comparatively shorter from scientific text. (2) these models also tend to repeat tokens and extract phrases from the input.

4.6 Summary

This chapter presents the details of our data curation and verification process. We perform a detailed analysis of our datasets for lexical and readability complexity, from which we infer that the German summaries of both datasets are lexically rich and comparatively readable than

their corresponding English text. We further empirically evaluate our datasets for summarization which suggests the viability and amenability of our datasets. These results also highlight the challenging nature of these datasets for the existing summarization models. These results show that existing summarization models are insufficient to generate long cross-lingual scientific summaries. Moreover, the uniqueness of the SPEKTRUM dataset directs us towards text simplification.

Part III

Models for Cross-lingual Science Journalism

Chapter 5

SSR: Select, Simplify and Rewrite

> "Divide each difficulty into as many parts as is feasible and necessary to resolve it."
>
> Rene Descartes

In the previous chapter, we collected two datasets from SPEKTRUM and WIKIPEDIA, performed readability analysis on the datasets and evaluated them for summarization task. From readability analysis, it is identified that German summaries of both datasets are lexically rich and comparatively readable than their corresponding English text. From the datasets evaluation, we found that the existing abstractive models did not perform well with longer inputs of scientific text. In this chapter, we investigate our next research question:

3. **Can a combination of summarization and simplification models perform better than the existing summarization models for Cross-lingual Science Journalism?**

To address the properties of our datasets and the limitation of existing models, we develop a pipeline model - **S**ELECT, **S**IMPLIFY and **R**EWRITE (SSR) for Cross-lingual Science Journalism. Sections 5.1 and 5.2 provide an overview of our approach and detailed descriptions of components, respectively. Sections 5.3, 5.4 and 5.6 include the experimental setup and the results. Section **??** presents the readability analysis of generated outputs.

5.1 Overview

Figure 5.1 illustrates SSR's information flow between its components. In the middle, we have the input and target pair. The input is processed by **SELECT** that accepts English texts and generates English outputs consisting of salient sentences of given inputs. Then **SIMPLIFY**

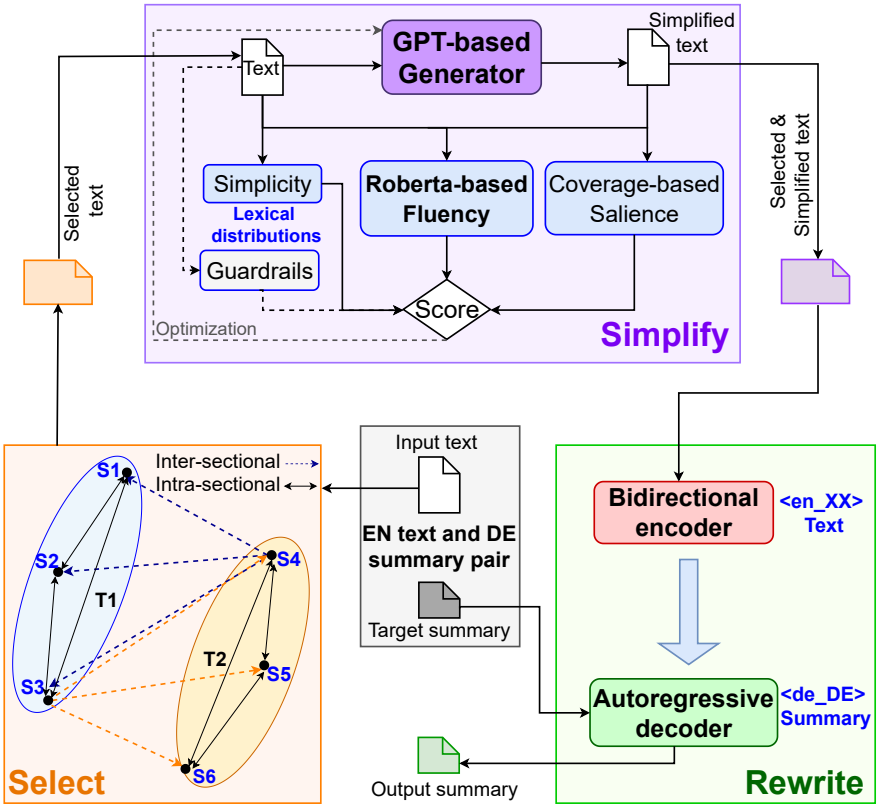

Figure 5.1: The SSR architecture - in the middle, we have the input and target pair. The input is processed by **SELECT** and **SIMPLIFY** for extraction and simplification. Then **REWRITE** processes this transformed input with the target summary to generate the cross-lingual summary.

receives the **SELECT** outputs and generates a linguistically simplified version of given inputs (English). In the last, **REWRITE** accepts the output of **SIMPLIFY** at the encoder as the input and a reference German summary at the decoder to generate a German summary as output. We apply a divide-and-conquer approach to break down the task into manageable components. We divide cross-lingual scientific summarization into two further components: monolingual scientific summarization and cross-lingual abstractive summarization. Here we discuss the rationale behind it before discussing SSR's components.

1. **Scientific Discourse.** For the scientific text, summarization models should include the salient information in summary from all sections because the pivotal content is spread over the entire text, following an "hourglass" structure (see Figure 2.1). The existing models accept only "lead" tokens from the source while discarding the rest. Mostly summarization models have been built on news datasets, which follow an "inverted pyramid" structure, so this conventional method is reliable for news but ineffective for

scientific texts. Some studies have developed models to exploit scientific texts' structural properties for better summaries (Cohan et al., 2018; Dong et al., 2021).

2. **Text length.** The average length of scientific documents is 1560 to 4900 words (Fatima and Strube, 2021). Even recently, there has been a significant gap between the average and accepted input lengths (max. 2048 tokens) by traditional and pre-trained models, *i.e.*, BART, GPT, LONGFORMER, *etc*. These recent models are still struggling to handle sizable documents (Jin et al., 2020). Some researchers have developed hybrid summarization models to deal with this problem (Liu and Lapata, 2019; Pilault et al., 2020), but the recent models are still struggling to handle sizable documents (Jin et al., 2020). Long texts often lead to model degradation resulting in hallucinations and factual inconsistencies (Maynez et al., 2020).

Considering all these challenges, we aim to deal with them with the SSR model. The SSR model is a proficient, adaptable and convenient plug-and-play application that has the strength to modify or change components without disturbing the information's flow.

5.2 SSR

5.2.1 Select

We apply the HIPORANK (HRANK) (Dong et al., 2021) model as **SELECT**, which is a hierarchical discourse model for scientific summarization.

5.2.1.1 Graph-based Ranking

It takes a document as a graph $G = (V, E)$, where V is the set of sentences and E is the set of relations between sentences. A directed edge e_{ij} from sentence v_j to sentence v_i is weighted by a (cosine) similarity score:

$$w_{ij} = f(sim(v_i, v_j)) \tag{5.1}$$

where f is an additional weight function.

5.2.1.2 Hierarchical Connections

A hierarchical graph is created upon sections and sentences for intra-sectional (local) and inter-sectional (global) hierarchies. The asymmetric edge weights are calculated on the hierarchical graph. The asymmetric edge weighting works on boundary functions at sentence and section levels to find important sentences.

5.2.1.3 Similarity of Pairs

Before calculating asymmetric edge weights over boundaries, a sentence-sentence pair similarity $sim(v_j^T, v_i^T)$ and a section-sentence pair similarity $sim(v^S, v_i^T)$ are computed with cosine similarity with various vector representations. However, these similarity scores cannot capture the salience well, so asymmetric edge weights are calculated and injected over intra-section and inter-section connections.

5.2.1.4 Asymmetric edge weighting over sentences

To find important sentences near the boundaries, a sentence boundary function (s_b) computes scores over sentences (v_i^T) in a section T:

$$s_b(v_i^T) = min(x_i^T, \alpha(n^T - x_i^T)) \tag{5.2}$$

where n^T is the number of sentences in section T and x_i^T represent sentence at i^{th} position in the section T. α is a hyper-parameter that controls the relative importance of the start or end of a section or document. The sentence boundary function allows integration of directionality in edges and weighing edges differently based upon their occurrence with a more or less important sentence in the same section. The weight w_{ji}^T for intra-section edges (incoming edges for i) is defined as:

$$w_{ji}^T = \begin{Bmatrix} \lambda_1 * sim(v_j^T, v_i^T),\ if\ s_b(v_i^T) \geq s_b(v_j^T) \\ \lambda_2 * sim(v_j^T, v_i^T),\ if\ s_b(v_i^T) < s_b(v_j^T) \end{Bmatrix} \tag{5.3}$$

where $\lambda_1 < \lambda_2$ for an edge e_{ji} occurs with i is weighted more if i is closer to the text boundary than j.

5.2.1.5 Asymmetric edge weighting over sections

Similarly, a section boundary function (d_b) computes the importance of a section (v^T) to reflect that sections near a document's boundaries are more important:

$$d_b(v^T) = min(x^T, \alpha(N - x^T)) \tag{5.4}$$

where N is the number of sections in the document and x^T represents section at T^{th} position in the document. The section boundary function enables injecting asymmetric edge weighting w_i^{ST} section edges.

$$w_i^{ST} = \begin{Bmatrix} \lambda_1 * sim(v^S, v_i^T),\ if\ d_b(v^T) \geq d_b(v^S) \\ \lambda_2 * sim(v^S, v_i^T),\ if\ d_b(v^T) < d_b(v^S) \end{Bmatrix} \tag{5.5}$$

where $\lambda_1 < \lambda_2$ for an edge e_i^{ST} occurs to $i \ \epsilon \ T$ is weighted more if section T is closer to the text boundary than section S. The boundary functions in Equation 5.2 and 5.4 naturally prevent *redundancy* because similar sentences have different boundary positional scores.

5.2.1.6 Overall Importance

It is computed as the weighted sum of local and global centrality scores:

$$c(v_i^T) = \mu \cdot c_{inter}(v_i^T) + c_{intra}(v_i^T), \tag{5.6}$$

$$c_{intra}(v_i^T) = \sum_{v_j^T \in T} \frac{w_{ji}^T}{|T|}, \tag{5.7}$$

$$c_{inter}(v_i^T) = \sum_{v_j \in D} \frac{w_i^{JT}}{|D|} \tag{5.8}$$

where T is the neighboring sentences set of v_i^T, D is the neighboring sections set, and μ is an inter-section centrality weighting factor.

5.2.1.7 Generation

A summary is generated by greedy extraction with the highest importance scores until a predefined word limit is reached.

5.2.2 Simplify

We adopt the KEEP-IT-SIMPLE (KIS) (Laban et al., 2021) model as **SIMPLIFY**, a reinforcement learning syntactic and lexical simplification model. The KIS model trains on maximizing the reward of its four components: simplicity, fluency, salience and guardrails. Here we discuss the details of its components and training objective.

5.2.2.1 Simplicity

Simplicity is computed at syntactic and lexical levels: S_{score} is calculated by Flesch Kincaid Grade Level (FKGL) with linear approximation, and L_{score} is computed with the input paragraph (W_1) and the output paragraph (W_2) as follows:

$$L_{score}(W_1, W_2) = \left[\frac{1 \ - \ \Delta Z(W_1, W_2) \ - \ c}{c} \right]^+ \tag{5.9}$$

where $\Delta Z(W_1, W_2) = Z(W_2 - W_1) - Z(W_1 - W_2)$ is the average Zipf frequency of inserted words and deleted words, clipped between 0 and 1 (denoted as $[\cdot]^+$), and c is a median value to target Zipf shift in the L_{score}.

5.2.2.2 Fluency

The fluency component consists of a GPT-based Language Model (LM) generator and a ROBERTA-based discriminator. The fluency score is computed with a likelihood of the original paragraph ($LM(p)$) and the generated output ($LM(q)$):

$$LM_{score}(p,q) = \left[\frac{1 \; - \; LM(p) \; - \; LM(q)}{\lambda}\right]^{+} \tag{5.10}$$

where λ is a trainable hyperparameter.

If the $LM(q) < LM(p)$ by λ or more, $LM_{score}(p,q) = 0$. If $LM(q) \geq LM(p)$, then $LM_{score}(p,q) = 1$, otherwise it is a linear interpolation. As LM_{score} is static and deterministic, a dynamic discriminator is trained jointly with the generator for the dynamic adaption of the fluency score is an optimized solution.

The ROBERTA-based discriminator is a classifier with two labels: 1 = authentic paragraphs and 0 = generator outputs. The discriminator is trained on the training buffer. The discriminator score is computed on the probability that a paragraph (q) is authentic:

$$D_{Score}(q) = p_{disc}(Y = 1|X = q) \tag{5.11}$$

where X denotes the input and Y is the output probability.

5.2.2.3 Salience

The salience component is based on a transformer-based coverage model trained to look at the generated text and answer fill-in-the-blank questions about the original text. Its score is based on the model's accuracy: the more filled results in relevant content and the higher score. All non-stop words are masked, as the task expects most of the original text should be recoverable.

5.2.2.4 Guardrails

Brevity and inaccuracy are simple pattern-based binary scores to improve the generation. The brevity ensures the similar lengths of the original paragraph (L_1) and generated paragraph (L_2). The brevity is defined as compression: $C = L_2/L_1$ where the passing range of C is $C_{min} \leq C \leq C_{max}$. The inaccuracy is a Named Entity Recognition (NER) model for extracting entities from the original paragraph (E_1) and the output paragraph (E_2). It triggers if an entity is present in E_2 is not in E_1.

5.2.2.5 Training

The KIS model trains on a variation of Self-Critical Sequence Training (SCST) named k-SCST, so the loss is redefined for conditional generation probability:

$$\mathcal{L} = \sum_{j=1}^{k} \bar{R}^{S} - R^{Sj} \sum_{i=0}^{N} \log p(w_i^{Sj} | w_{<i}^{Sj}, P) \tag{5.12}$$

where k is the number of sampled candidates, and R^{Sj} and $\bar{R}^{S}$ denote the candidate and sampled mean rewards, P is the input paragraph and N is the number of generated words. All these components are jointly optimized by using the product of all components as the total reward.

5.2.3 Rewrite

We adopt multilingual BART (mBART) (Liu et al., 2020) model as **REWRITE**, which is a transformer-based architecture with a noise function. The mBART model consists of 12 layers on each side in the encoder and decoder, closely related to the architecture used in BERT.

5.2.3.1 Self-attention

Every layer of the encoder and decoder has its own self-attention, consisting of keys, values, and queries from the same sequence.

$$A(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{5.13}$$

where Q is a query, K^T is transposed K (key) and V is the value. All parallel attentions are concatenated to generate multi-head attention scaled with a weight matrix W.

$$MH(Q, K, V) = Concat(A_1, \cdots, A_h) \cdot W^O \tag{5.14}$$

5.2.3.2 Cross-attention

Each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder (as in the vanilla transformer S2S model). The cross-attention is the attention between the encoder and decoder, which gives the decoder a weight distribution at each step, indicating the importance of each input token in the current context.

5.2.3.3 Conditional Generation

The MBART model is a S2S model parameterized by θ, accepts an input text as $x = (x_1, \cdots, x_n)$ and generates a summary $y = (y_1, \cdots, y_m)$ as output. The generation probability of y is conditioned on x and trainable parameters θ as follows:

$$p(y|x, \theta) = \prod_{t=1}^{m} p(y_t|y_{<t}, x, \theta) \tag{5.15}$$

5.3 Experiments

5.3.1 Datasets

We use the WIKIPEDIA dataset with a split of 80-10-10 for experiments, while the SPEKTRUM dataset is used for zero-shot adaptability as a case study.

5.3.2 Models

We define various extractive and abstractive baselines with diverse experimental settings. Our first four models are based on Summarize-then-translate. As we are using extractive summarization models here, so we define them as EXT-TRANS: (1) LEAD: it takes the initial n tokens of the input document, (2) TEXTRANK (TRANK) (Mihalcea and Tarau, 2004): it is a graph-based algorithm, (3) ORACLE (Nallapati et al., 2017): it is based on greedy extraction, (4) HRANK with SENTENCE-BERT (SB)[1] (Dong et al., 2021).

Our second group is non-pre-trained models which we train from scratch. We define them as direct CLS models: (5) LSTM & attention-based S2S (S2S), (6) pointer generator network (PGN), (7) transformer-based encoder-decoder (TRF) (Fatima and Strube, 2021).

Our last group consists of pre-trained models, which we fine-tune with our dataset: (8) MT5 (Xue et al., 2021), (9) MBART (Liu et al., 2020) and (10) LongFormer encoder-decoder (LONG-ED) (Beltagy et al., 2020).

The input to the encoder for all these models is the first (lead) 1024 tokens of each document. We create the EXT-TRANS pipeline with T5 for translation wherever required.

5.3.3 Training and Inference

The HRANK model is a readily available unsupervised model, so it does not require training. While the KIS and MBART models are trained independently.

[1]We apply four embeddings with HRANK: RANDOM (RD), BIOMED (BM), SENTENCE-BERT (SB) and PACSUM (PS) to find the best one.

5.3.3.1 Libraries and Hardware

We run all of our experiments with Pytorch[2], Hugging Face[3] and Apex[4] libraries. For all models, we complete training and inference on a single Tesla P40 GPU with 24GB memory.

5.3.3.2 Hyperparmeters

For all the abstractive baselines, we train or fine-tune all models for a maximum of 30 epochs with early stopping. We use a batch size of 4 with a learning rate (LR) of $5e^{-5}$, and 100 warm-up steps to avoid over-fitting of fine-tuned model. We use the Adam optimizer ($beta_1 = 0.9$, $beta_2 = 0.99$, $\epsilon = 1e^{-08}$) with LR linearly decayed LR scheduler. During decoding, we use the maximum length of 200 tokens with a beam size of 4. The encoder language is set to English, and the decoder language is German.

We adopt similar settings for the REWRITE component as used for baselines. The REWRITE component is based on mBART-large-50. For the KIS model, we initialize GPT-2-medium model with the Adam optimizer at a learning rate of 10^{-6}, a batch size of 4 and $k = 4$. We initialize ROBERTA-base with the Adam optimizer at a learning rate of 10^{-5} and a batch size of 4.

5.3.3.3 Training Time

The S2S and TRF models take 6 days, the PGN model takes 3 days, the mBART model takes 6 days and the KIS and LONG-ED models take 14 days for training.

5.3.4 Automatic Evaluation

We evaluate all models with three automatic metrics - ROUGE, BERT-SCORE and FRE. For human evaluation, we consider **relevance**, **fluency**, **readability** and **overall ranking**.

5.3.4.1 Human Evaluation

We conduct a human evaluation to compare the outputs of SSR with mBART (baseline) for the same linguistic properties. It is noteworthy to mention here that human evaluation of long cross-lingual scientific text is quite challenging because it requires bi-lingual annotators with some scientific background.

[2] https://pytorch.org/
[3] https://huggingface.co/
[4] https://github.com/NVIDIA/apex

We published a job on University Job Portal with the task description, requirements, implications, working hours, wage per hour and location. We hired five annotators from a German University who are native Germans, fluent in English and master's or bachelor's science students. The selected students for the evaluation task submitted their consent while agreeing to the job. We compensated them at €15 per hour, while the minimum student wage ranges between €$9.5 - 12$ in 2022 according to German law[5]. Further details about the guidelines for human evaluation are presented in Appendix 5.A.

5.4 Wikipedia Results

We report FRE and the F-SCORE for ROUGE and BERT-SCORE of all models for the WIKIPEDIA test set in Table 5.1. All the results are the average of five runs for each model. The first block includes the EXT-TRANS baselines, the second and third blocks present direct CLS and fine-tuned models, and the last block includes the different variations of SSR models. We compute the statistical significance of the results with the Mann-Whitney two-tailed test for a p-value ($p < .001$) against the fine-tuned models.

From Table 5.1, we find that all EXT-TRANS models perform quite similarly considering the ROUGE, BERT-SCORE and FRE scores. For all these extractive models, the FRE scores fall in hard readability. For the direct CLS models in Table 5.1, the TRF model performs better than the PGN and S2S models for the ROUGE, BERT-SCORE and FRE scores. Interestingly, the FRE scores of these models are similar to EXT-TRANS models. One reason behind the low scores for the PGN and S2S models is that these models use word-based fixed-size vocabulary, due to which UNK tokens are present in the outputs. Moreover, the PGN model heavily relied on coverage of original text, due to which the FRE score is low.

Considering fine-tuned models in Table 5.1, the MBART model performs the best in this group, the MT5 model's performance is also good, however, the LONG-ED model performs quite low. We also run some experiments for the LONG-ED model with 2048 tokens at the encoder side, resulting in much worse performance. From this comparison, we infer that longer inputs of lead tokens are not helpful for scientific summarization. These models produce easier readability outputs except the LONG-ED model. As these models are pre-trained with large-size datasets, we infer that these models have latent simplification properties.

Comparing the performance of the best baseline with SSR from Table 5.1, the SSR model outperforms the MBART model by wide margins for the ROUGE, BERT-SCORE and FRE scores. We infer from these results that transforming input texts by **SELECT** and **SIMPLIFY** components helps the SSR model learn better contextual representations.

[5] Minimum wage in Germany

MODELS	ROUGE-1	ROUGE-2	ROUGE-L	BERT-SCORE	FRE
EXT-TRANS					
LEAD	18.90	2.68	12.40	64.28	22.11
TRANK	17.83	2.25	11.59	63.81	24.45
ORACLE	19.63	2.78	12.49	64.30	25.19
HRANK	18.09	2.25	11.52	63.75	25.18
CLS					
S2S	18.37	4.04	16.55	52.76	25.14
PGN	20.72	3.79	18.68	55.67	26.56
TRF	21.61	4.37	18.10	60.95	29.75
FINE-TUNED					
mT5	24.57	7.66	18.34	68.40	40.18
LONG-ED	15.35	4.57	14.39	63.89	23.66
mBART	27.02	8.93	20.46	70.16	42.23
OURS					
SIM+RE	27.65	6.65	18.35	70.34	46.05
SEL+RE	28.50	9.71	21.85	**70.47**†	44.52
SSR	**30.07**†	**12.60**†	**24.14**†	70.45	**50.45**†

Table 5.1: The WIKIPEDIA results for all baselines and SSR. Underline refers to the best baseline results and **bold**† denotes significant improvements ($p < .001$).

5.4.1 Component Analysis

Table 5.2 presents the performance of ablated models. All the results are the average of five runs for each model. **SIM+RE** denotes the model without **SELECT**, resulting in a significant decrease in performance for ROUGE and and FRE as compared to SSR but maintaining the performance for BERT-SCORE. **SEL+RE** refers to the model without **SIMPLIFY**, also resulting in a notable drop in performance ROUGE and FRE as compared to SSR, while showing similar performance for BERT-SCORE. Overall, the full SSR model (the last row) demonstrates that all three components are necessary to generate good-quality simplified cross-lingual stories.

MODELS	ROUGE-1	ROUGE-2	ROUGE-L	BERT-SCORE	FRE
SIM+RE					
MBART	27.65	6.65	18.35	70.34	46.05
SEL+RE					
TRANK	26.70	8.60	20.06	70.07	38.15
ORACLE	29.27	10.11	21.89	**70.99**†	40.11
HRANK	28.50	9.71	21.85	70.47	44.52
SEL+SIM+RE					
MT5	26.74	10.25	21.63	69.52	45.57
LONG-ED	17.25	6.58	14.99	65.32	27.23
MBART	**30.07**†	**12.60**†	**24.14**†	70.45	**50.45**†

Table 5.2: The WIKIPEDIA results for SSR component analysis, where **bold**† denotes significant improvements ($p < .001$).

5.4.1.1 Component Replacement.

We also explore the behavior of SSR by component replacement with their counterparts. For **SELECT**, we replace HRANK with TRANK and ORACLE to compare their performances. Interestingly, ORACLE shows slightly higher performance as compared to HRANK. We manually analyzed the outputs of HRANK and ORACLE. During the manual analysis of a subset, we find that the HRANK model changes the order of sentences according to the importance score calculation of the section in some examples. We infer that it is the reason for the slightly low performance of HRANK. Overall, these results indicate the importance of **SELECT**.

For **SIMPLIFY**, we could not find any comparable paragraph-based simplification model to replace KIS. For **REWRITE**, we replace MBART with MT5 and LONG-ED to compare their performances. The MBART model demonstrates superior performance over its replacements. However, the performance of replacement models is improved compared to fine-tuned models. Overall, these results suggest that MBART is data-efficient and can be easily adapted for small-sized or low-resource data compared to MT5 (Lee et al., 2022). To conclude the component analysis, we infer from these results that the component replacements demonstrate the resilience and robustness of SSR with intact information flow.

MODELS	ROUGE-1	ROUGE-2	ROUGE-L	BERT-SCORE	FRE
CLS					
S2S	16.47	3.42	11.87	44.01	24.55
PGN	18.64	3.83	15.65	46.89	25.86
TRF	<u>20.81</u>	<u>4.19</u>	<u>17.54</u>	46.87	28.88
FINE-TUNED					
mT5	11.13	0.88	8.03	59.57	38.92
LONG-ED	1.98	0.10	1.29	50.65	29.31
mBART	18.16	1.48	9.54	<u>62.61</u>	<u>39.38</u>
OURS					
SSR	**23.24**†	**5.28**†	15.56	**64.90**†	**43.14**†

Table 5.3: The SPEKTRUM results for all baselines and SSR. <u>Underline</u> refers to the best baseline results and **bold**† denotes significant improvements ($p < .001$).

5.5 Case Study: Spektrum

Table 5.3 presents FRE and the F-SCORE for ROUGE and BERT-SCORE of baselines and the SSR model on the SPEKTRUM dataset. All the results are the average of five runs for each model. The first block includes the direct CLS models, the second block presents the fine-tuned models, and the last block includes the SSR model.

From the results presented in Table 5.3, we infer that the SSR model performs quite well on the SPEKTRUM set. We find a similar performance pattern among the models on the SPEKTRUM dataset. However, these results are lower than those on the WIKIPEDIA test set because these models are trained on the WIKIPEDIA training and validation sets. Comparing the baseline models, the MBART model outperforms other baselines for BERT-SCORE, while TRF takes the lead for ROUGE. However, SSR shows higher performance over all the baselines. Overall, the results on the WIKIPEDIA and SPEKTRUM datasets suggest the superior performance of the SSR model.

5.5.1 Human Evaluation

We further investigate the linguistic properties of generated summaries by comparing the outputs of the SSR and MBART models by human evaluation. We provide 25×2 (for each model)

random summaries with their original texts and gold reference summaries. The annotators were given a Likert scale from $1-5$ (1:worst, 2:bad, 3:neutral, 4:good, 5:best) for **relevance**, **fluency**, **simplicity** (readability) and **overall ranking**.

MODELS	FLUENCY (α)	RELEVANCE (α)	SIMPLICITY (α)	OVERALL (α)
mBART	3.08 (0.52)	1.74 (0.61)	3.65 (0.60)	2.31 (0.53)
SSR	3.95 (0.62)	3.27 (0.74)	3.83 (0.78)	3.49 (0.57)

Table 5.4: The SPEKTRUM human evaluation for mBART and SSR- the average scores for each linguistic property (Krippendorff's α).

We compute the average scores and inter-rater reliability using Krippendorff's α[6] over 20 samples, excluding the first five examples. Table 5.4 presents the results of human evaluation. From these results, we find that the SSR outputs are significantly better than the mBART model for **fluency**, **relevance**, **simplicity** and in the overall ranking. We summarize the overall performance of the SSR model on the SPEKTRUM dataset based on automatic scores and human evaluation that the SSR model outperforms all the baselines for Cross-lingual Science Journalism.

5.6 Analysis: Spektrum

We further extend our investigation for the readability analysis to compare the gold references and outputs. For all graphs, Text represents English documents, Gold denotes German references, FT represents the outputs of mBART and SSR denotes the SSR outputs.

5.6.1 Lexical Diversity

We compute Hypergeometric Distribution Diversity (HDD) and Measure of Textual Lexical Diversity (MTLD) to find lexical richness. Figure 5.2 presents the graphs for HDD and MTLD. We find from these graphs that gold summaries have higher lexical diversity, while both system summaries are slightly lower. These results indicate that the system summaries are not as lexically diverse as the gold references and are similar to the text.

5.6.2 Readability Index

We calculate two scores for readability - Coleman Liau Index (CLI) and Linsear Write Formula (LWF). For LWF, the recommended easy score for an adult reader is between $70-80$.

[6] https:/pypi.org/project/krippendorff/.

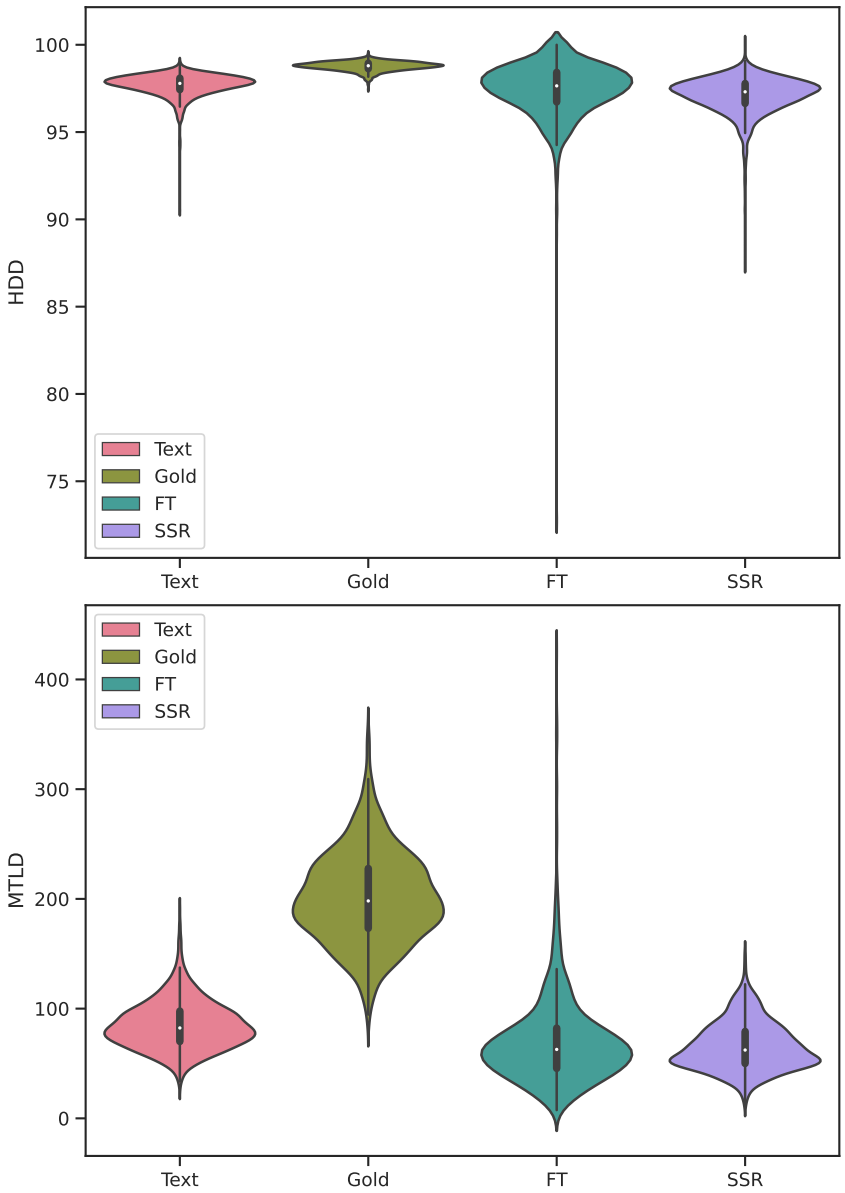

Figure 5.2: Distribution of lexical diversity. For HDD and MTLD ↑ score is better.

Figure 5.3 shows the violin plots of CLI and LWF. The CLI plot indicates that English texts have the highest readability among all. Then gold summaries and the SSR outputs stand on slightly less readability than English texts. While the mBART outputs are the most difficult among all. The LWF graphs show that the gold references and the SSR outputs are the easiest among all, while the mBART outputs are the most difficult. So we can conclude from the results of CLI and LWF that the SSR summaries are comparatively easy compared to the mBARToutputs. However, the difference in results with LWF and CLI is due to the difference in features used for calculation.

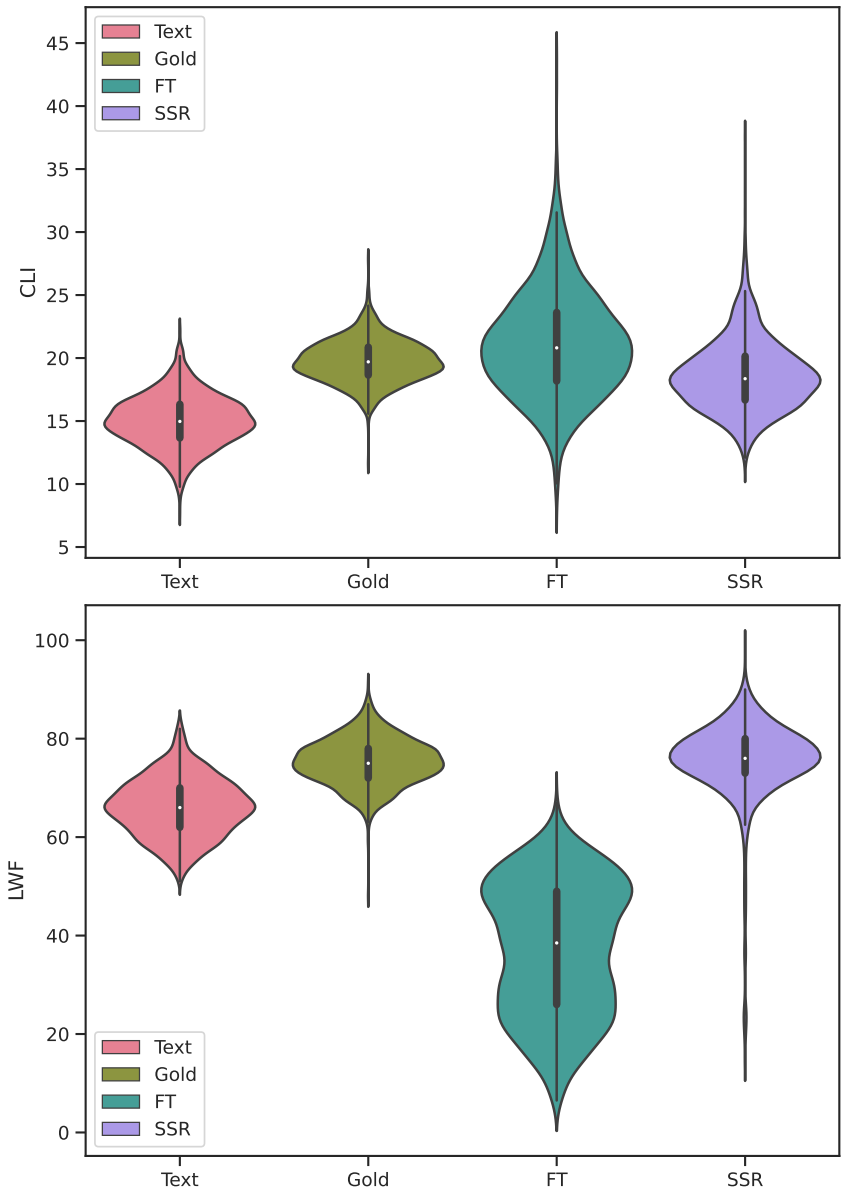

Figure 5.3: Distribution of readability scores. For CLI ↓ score is better. For LWF ↑ score is better.

5.6.3 Density Distribution

Word density (WD) and sentence density (SD) measure how much information is carried in a word and a sentence. However, word and sentence densities are correlated and can be a language function. The WD score is calculated by the average number of characters present in a word and the SD score refers to the average number of lexical words present in a sentence.

Figure 5.4 presents the violin plots for density scores. From these graphs, we interpret that the mBART model has the highest dense sentences among all, while the word densities of

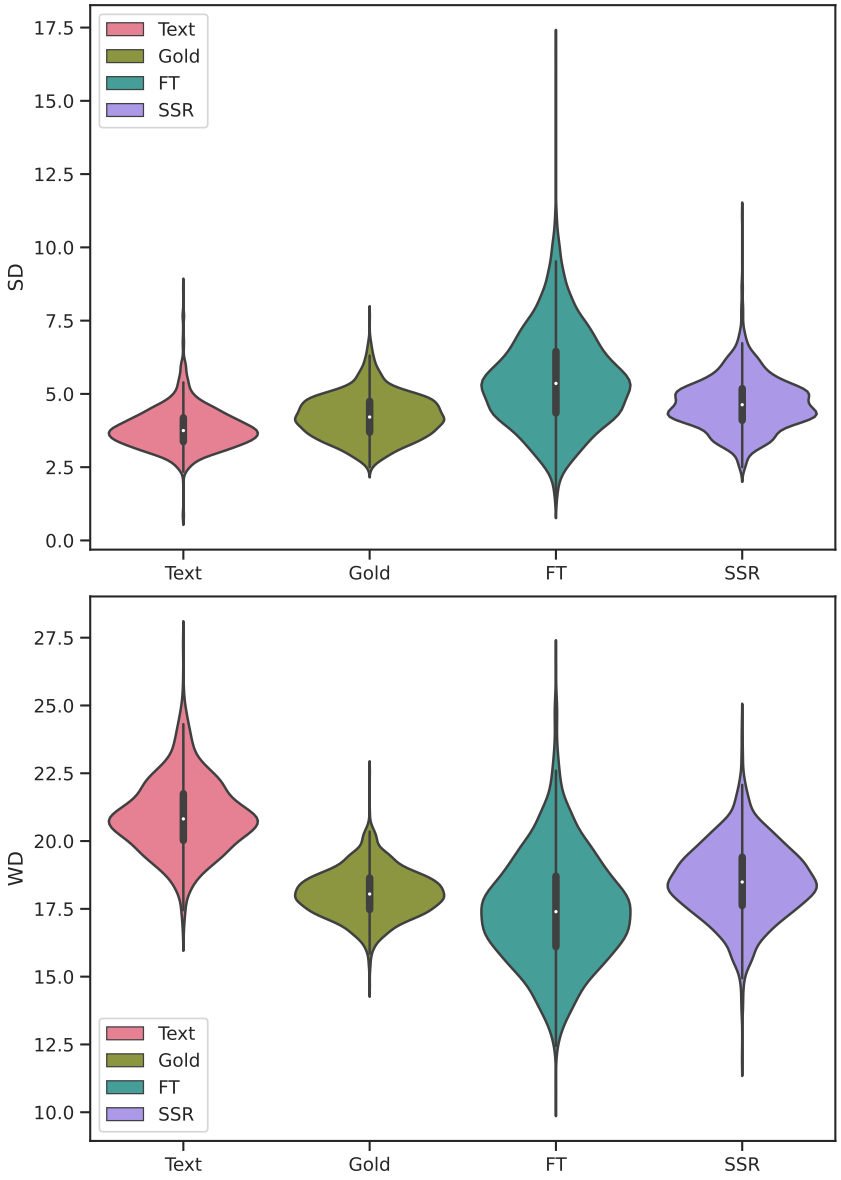

Figure 5.4: Distribution of density scores. For WD and SD ↓ score is better.

the SSR model are slightly higher than the gold references. Surprisingly, English texts have higher word density, even though German is famous for its inflections and compound words, suggesting that these English texts are harder to read for an average non-expert adult person.

To conclude the readability analysis, the scores of lexical diversity, readability indices and density scores suggest improved and better readability of the SSR outputs compared to the mBART model.

5.6.4 SSR Examples

We present some random outputs generated by the SSR and mBART models (in the last pages of this chapter). We mark wrong words or sentences with red and new words and unfaithful information with blue.

5.7 Limitations

We investigated Cross-lingual Science Journalism with **SELECT**, **SIMPLIFY** and **REWRITE**. We adopted HIPORANK as **SELECT** because it is a lightweight, unsupervised model that extracts a summary in a discourse-aware manner. However, when we replaced it with other extractive models during the component analysis, we found no significant difference in overall performance. We adopted KEEP-IT-SIMPLE for **SIMPLIFY** because it facilitates paragraph simplification. We found the model is quite heavy, making it slow during training. To the best of our knowledge, there is no paragraph-based simplification model we could explore in component replacement. The choice among various pre-trained models for **REWRITE** was quite challenging, as all these models are variations of transformer-based architectures. So we adopted the latest three SOTA models, which are efficient and effective summarization models. We also trained the vanilla S2S model, pointer-generator model, and transformer as our baselines to provide sufficient variations of SOTA models. We found the mBART model is more promising performance-wise in our experiments. However, its training time is also slow for our datasets due to longer inputs. In the future, we plan to improve the training time by model parallelization.

5.8 Summary

In this chapter, we investigate Cross-lingual Science Journalism as a fusion of text simplification and cross-lingual scientific summarization. We discuss the details of SSR and evaluate it against several strong baselines. Our empirical evaluation shows that SSR outperforms all baselines by wide margins on WIKIPEDIA and performs well on SPEKTRUM. We further explore the ablated models with component replacements, demonstrating the resilience and robustness of the SSR application. We conduct a human evaluation of the SPEKTRUM outputs, indicating the good linguistic properties, further affirmed by readability analysis. The SSR model directs us towards joint training of **SIMPLIFY** and **REWRITE** models as these models are similar in architectures and nature of tasks.

Target: ein mädchen aus südafrika lebt seit neun jahren mit einer hiv-infektion, ohne den erreger mit medikamenten kontrollieren zu müssen. das berichteten fachleute um avy violari von der university of the witwatersrand in johannesburg auf einer konferenz in paris. bei dem kind einer hiv-positiven mutter war die infektion nach der geburt im jahr 2007 festgestellt und im rahmen einer klinischen studie direkt nach der diagnose 40 wochen lang mit medikamenten behandelt worden. [...]
Translation: a girl from south africa has been living with an hiv infection for nine years without having to control the pathogen with medication. this is what experts around avy violari from the university of the witwatersrand in johannesburg reported at a conference in paris. the child of an hiv-positive mother was diagnosed with the infection after birth in 2007 and was treated with medication for 40 weeks immediately after the diagnosis as part of a clinical study. [...]

MBART: das kind, das **fiziert** wird mit dem hi-virus bei der geburt, ist die **geburtsache** für das infizieren des virus bei der geburt. die infektion des ungeborenen kindes mit dem hiv wird als latente hiv-infektion bezeichnet. die ursache für die erkrankung ist noch ungeklärt.
Translation: the child who is <UNK> with the hi virus at birth is the <UNK> for infecting the virus at birth. the infection of the unborn child with hiv is referred to as latent hiv infection. the cause of the disease is still unclear.

SSR: ein neunjähriges kind, das ohne behandlung mit einer antiretroviralen therapie die welt retten kann, ist eine person, die eine hiv-infektion überlebt hat. das berichtet eine arbeitsgruppe um avy violari in der fachzeitschrift proceedings of the national allergy and infectious diseases institute. das team hatte das kind im alter von neun jahren ohne die behandlung mit einer antiretroviralen therapie im rahmen einer klinischen studie in den usa geimpft. mit der antiretroviralen therapie wurde das kind auf eine hiv-infektion geschützt, ohne dass die infektion nach der geburt aufhörte. die entwicklung einer antiretroviralen therapie hat eine große wahrscheinlichkeit, dass die hiv-infektion nach der geburt wieder zurückkehrt.
Translation: a nine-year-old child who can save the world without treatment with antiretroviral therapy is a person who survived hiv infection. this is reported by a working group led by avy violari in the specialist journal proceedings of the national allergy and infectious diseases institute. the team had vaccinated the child at the age of nine without treatment with antiretroviral therapy as part of a clinical trial in the us. with the antiretroviral therapy, the child was protected against an hiv infection, without the infection stopping after birth. the development of an antiretroviral therapy has a high probability that the hiv infection will return after the birth.

Figure 5.5: An example where the MBART model generates a very short summary with wrong words in it. However, in the SSR model's output, we find quite a few wrong words and unfaithful information, while there is more topic-relevant information present in the summary.

5.A Guidelines for Human Evaluation

5.A.1 Task Description

We present annotators with 25 examples of documents paired with a reference summary and two system-generated summaries. The models' identities are hidden. The annotators were asked to evaluate each model summary for the following linguistic features after reading the original English text. The annotators were given a Likert scale from $1-5$ (1=worst, 2=bad, 3=neutral/ok, 4=good, 5=best). They were asked to use the first 5 examples to resolve the

Target: bakterien leben überall-auch auf 400jahre alten gemälden. welche das sind und was für konsequenzen das für die kunstwerke haben könnte, berichtet eine arbeitsgruppe um elisabetta caselli von der università degli studi di ferrara anhand des barockgemäldes incoronazione della vergine. wie das team in plos one berichtet, weist das werk des malers carlo bononi, das seit etwa1620 unter der decke der basilika santa maria in vado in ferrara montiert war, neben rissen auch spuren mikrobiellen abbaus durch die besiedelnden pilze und bakterien auf. [...]
Translation: bacteria live everywhere, even on paintings that are 400 years old. a working group led by elisabetta caselli from the università degli studi di ferrara reports on the baroque painting incoronazione della vergine what these are and what consequences they could have for the works of art. as the team in plos one reports, the work by the painter carlo bononi, which had been mounted under the ceiling of the basilica santa maria in vado in ferrara since around 1620, shows not only cracks but also traces of microbial degradation caused by colonizing fungi and bacteria. [...]

MBART: biodeterioration bezeichnet den abbau von kunstwerken durch mikroorganismen. man unterscheidet zwischen organischer und anorganischer deterioration. im gegensatz dazu werden biologische abbauprodukte durch anorganische abbauprodukte, die durch abbau von fäkalien oder desinfektionen bewirkt werden.
Translation: biodeterioration refers to the degradation of works of art by microorganisms. a distinction is made between organic and inorganic degradation. in contrast, biological degradation products are replaced by inorganic degradation products caused by the degradation of faeces or disinfectants.

SSR: mikroorganismen wie bakterien, pilze oder parasiten begünstigen den kunstwerkbau, indem sie das kunstwerk durch ihre lebensnotwendigen mikroorganismen schädigen. das berichtet eine arbeitsgruppe um den italienischen maler carlo bononi. wie das team in der zeitschrift frontiers in microbiology berichtet, handelt es sich um eine mischung aus mikroben, die sich auf der oberfläche des kunstwerks bilden, und mikroorganismen, deren lebensbedingungen das kunstwerk beeinträchtigen. wie das team in frontiers in microbiology berichtet, sind die mikroorganismen bestandteil der abdeckung und bildung des kunstwerks, die sich auf der oberfläche des kunstwerks bilden. [...]
Translation: microorganisms such as bacteria, fungi or parasites favor the artwork construction of art by damaging the work of art with their vital microorganisms. this is reported by a working group led by the italian painter carlo bononi. as the team reports in the journal frontiers in microbiology, it is a mixture of microbes that form on the surface of the artwork and microorganisms whose living conditions affect the artwork. as the team reports in frontiers in microbiology, the microorganisms that form on the surface of the artwork are part of the artwork's covering and formation. [...]

Figure 5.6: An example where the MBART model has wrong topic attention, generating more frequent wrong words than the SSR model and some unfaithful information. While the SSR model's output is better with more relevant information and has fewer wrong words or unfaithful information.

annotator's conflict, while the rest examples were to be evaluated independently.

5.A.2 Linguistic Features

We asked annotators to evaluate each summary for the following features.

Relevance. A summary delivers adequate information about the original text. Relevance determines the content relevancy of the summary.

Fluency. The words and phrases fit together within a sentence, and so do the sentences.

Target: alle wassermoleküle bestehen aus einem sauerstoff-und zwei wasserstoffatomen-und doch gibt es zwei arten von ihnen. die kernspins der beiden wasserstoffatome können in die gleiche richtung zeigen, oder aber in entgegengesetzte. der unterschied zwischen ortho-wasser und para-wasser ist klein, aber durchaus bedeutsam, zeigte jetzt eine arbeitsgruppe um stefan willitsch von der universität basel. [...]
Translation: all water molecules consist of one oxygen and two hydrogen atoms-yet there are two types of them. the nuclear spins of the two hydrogen atoms can point in the same direction or in opposite directions. the difference between ortho water and para water is small but significant, as a working group led by stefan willitsch from the university of basel has shown. [...]

MBART: para-wasser (auch para-wasser oder ortho-wasser) ist ein molekül aus der gruppe der wasserstoff-isomere. es besteht para-wasser in zwei formen, para-sonde und ortho-wasser. para-wasser ist strukturell mit wasser gleich brüchig und hat die form eines sphärischen moleküls. wasser ist ein molekül aus der gruppe der ortho-isomere. mit hilfe der quantenmechanischen methode kann der grundzustand des moleküls abgeschätzt werden.
Translation: para-water (also para-water or ortho-water) is a molecule from the group of hydrogen isomers. There are para water in two forms, para probe and ortho water. Para-water is structurally as brittle as water and has the shape of a spherical molecule. water is a molecule from the group of ortho isomers. the ground state of the molecule can be estimated using the quantum mechanical method.

SSR: wasser ist nicht nur ein chemisches element, sondern auch ein physikalisches element. es besitzt zwei unterschiedliche isomere (para-wasser und ortho-wasser), die sich in zwei verschiedenen formen unterscheiden : para-wasser und ortho-wasser. die beiden wasserstoffatome unterscheiden sich allerdings in den eigenschaften des wasserstoffs, weil die beiden formen von wasserstoffatomen dasselbe atom-oder molekülzentrum besitzen. das berichtet eine arbeitsgruppe um paul virilio von der universität genf. die forscher haben nun herausgefunden, dass wassermoleküle in zwei unterschiedlichen, unterschiedlichen zuständen unterschieden werden können : para-wasser und ortho-wasser. die beiden wasserstoffatome unterscheiden sich dabei nicht, weil sie nur einen wasserstoffatomen-oder molekül.
Translation: water is not only a chemical element but also a physical element. it has two different isomers (para-water and ortho-water) which differ in two different forms : para-water and ortho-water. However, the two hydrogen atoms differ in the properties of hydrogen because both forms of hydrogen atoms have the same atomic or molecular center. this is reported by a working group led by paul virilio from the university of geneva. The researchers have now discovered that water molecules can be distinguished in two distinct, distinct states: para-water and ortho-water. the two hydrogen atoms do not differ because they are only one hydrogen atom or molecule.

Figure 5.7: An example where both the MBART and SSR models produce wrong phrases and repetitions of similar words. Also, there is some unfaithful information present in both outputs.

Fluency determines the structural and grammatical properties of a summary.

Simplicity. Lexical (word) and syntactic (syntax) simplicity of sentences. A simple summary should have minimal use of complex words/phrases and sentence structure.

Overall Ranking. Compared with reference summaries, how is the overall coherence of each model's summary?

Chapter 6

SimCSum: Joint Learning of Simplification and Cross-lingual Summarization

> "You can never understand one language until you understand at least two."
>
> Geoffrey Willans

In the previous chapter, we evaluate the performance of a pipeline model - SSR for Cross-lingual Science Journalism. The SSR model outperforms all baselines by a wide margin. From the SSR experiments, we find that the simplification model positively impacts the SSR model. However, the model used in SSR for simplification is quite slow for training. The simplification model is similar to the abstractive summarization model. Therefore, we can train summarization and simplification models together. In this chapter, we investigate our last research questions:

4. **Can joint training of cross-lingual summarization and simplification help to improve the quality of generated summaries for Cross-lingual Science Journalism?**

To answer this question, we present a Multitask Learning-based model for joint training of **Sim**plification and **C**ross-lingual **Sum**marization - SIMCSUM. Moreover, it is quite lightweight for training and inference. To the best of our knowledge, no prior work combines simplification and cross-lingual summarization for joint learning. Sections 6.1 and 6.2 present the background and architecture of SIMCSUM, Sections 6.3, 6.4 and 6.5 discuss the experimental setup and results. Section 6.6 presents a detailed analysis of generated summaries.

6.1 Multitask Learning

SIMCSUM is a multitask model that considers cross-lingual summarization as the main task and simplification as the auxiliary task. Multitask Learning is an approach in deep learning which improves generalization by learning different noise patterns from data related to different tasks. We define our Multitask Learning-based model trained on two tasks: simplification and cross-lingual summarization. We adopt hard parameter sharing as it improves the positive transfer and reduces the risk of overfitting (Ruder, 2017). We define our tasks here and then discuss the architecture of our model.

6.1.1 Summarization

6.1.1.1 Monolingual

Monolingual single-document abstractive summarization takes a single piece of text in a source language and aims to generate a human-like summary in the source language. For producing a human-like summary, the abstractive summarization model relies on the paraphrasing of salient information. We define single-document abstractive summarization mathematically as follows.

Given a text $X = \{x_1, \cdots, x_m\}$ with m number of sentences comprising of a set of words (vocabulary) $W_X = \{w_1, \cdots, w_X\}$, an encoder-decoder-based abstractive summarizer generates a summary $Y = \{y_1, \cdots, y_n\}$ with n number of sentences that contain salient information of X, where m and Y consisting of a set of words $W_Y = \{w_1, \cdots, w_Y | \exists w_i \notin W_X\}$ as the summarizer might generate some new words that are not present in X. The decoder learns the conditional probability distribution over the given input and all previously generated words, where t denotes the time step.

$$P_\theta(Y|X) = \log P(y_t|\ y_{<t}, X) \tag{6.1}$$

6.1.1.2 Cross-lingual

Cross-lingual summarization adds another dimension of language for simultaneous translation and summarization. Cross-lingual single-document abstractive summarization takes a single piece of text in a source language and aims to generate a human-like summary in the given target language. We define single-document cross-lingual summarization mathematically as follows.

Given a text $X^l = \{x_1^l, \cdots, x_m^l\}$ in a language l with m number of sentences comprising of a vocabulary $W_X^l = \{w_1^l, \cdots, w_X^l\}$, a cross-lingual summarizer generates a summary $Y^k =$

$\{y_1^k, \cdots, y_n^k\}$ in a language k that contains salient information in X, where m and Y consisting of a vocabulary $W_Y^k = \{w_1^k, \cdots, w_Y^k | \exists w_i \notin W_X^l\}$ as there is a possibility of common words in the source and target languages. The conditional probability is the same as in Equation 6.1, the only difference being that the language on the decoder is different from the encoder.

6.1.2 Simplification

Document-level simplification takes a single piece of text and aims to produce a simplified version of that text by reducing linguistic complexity at lexical and syntactic levels. We define the document-level simplification task as follows.

Given a text $X = \{x_1, \cdots, x_m\}$ with m number of sentences comprising of a vocabulary $W_X = \{w_1, \cdots, w_X\}$, a simplification model generates the output text $Y = \{y_1, \cdots, y_n\}$ that retains the primary meaning of X, yet more comprehensible as compared to X, where m and Y consisting of a vocabulary $W_Y = \{w_1, \cdots, w_Y | \exists w_i \notin W_X\}$. The conditional probability is also the same as in Equation 6.1.

6.2 SimCSum

We illustrate the framework of SIMCSUM in Figure 6.1. The SIMCSUM model jointly trains on simplification and cross-lingual summarization. The SIMCSUM model adopts hard parameter sharing where the encoder is shared between the tasks while having two task-specific decoders. The decoders share the cross-attention layer, and the loss is combined to update the parameters (θ). We opt for two decoders because each task's output language and length differ, and we do not want to pad the summaries for optimal training.

The training method is described in Algorithm 1. Here we discuss the further details of the SIMCSUM model. For all mathematical definitions, $\mathcal{T} \in \{sim, sum\}$ denotes a task.

6.2.1 Architecture

Considering the excellent text generation performance of multilingual BART (mBART) (Liu et al., 2020), we implement the SIMCSUM model based on it and modify it for two decoding sides for each task. Each encoder and decoder stack consists of 12 layers with dimensions of 1024 on 16 heads (∼680M parameters). The mBART model has an additional layer which is the normalization layer on top of both the encoder and decoder.

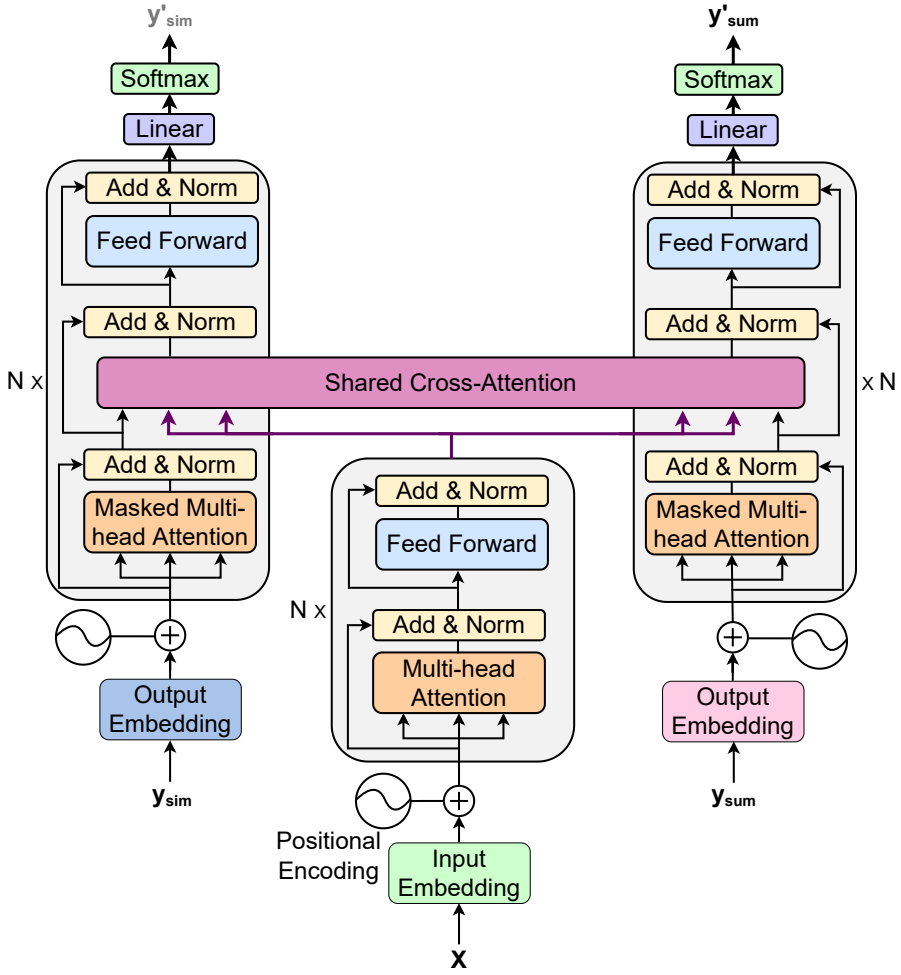

Figure 6.1: SIMCSUM architecture consists of one shared encoder with two decoding sides for Simplification and cross-lingual Summarization.

6.2.1.1 Self-Attention

Each layer of the encoder and decoder has its self-attention, consisting of keys, values, and queries generated from the same sequence, where the key is a label of a word used to distinguish between different words. The query represents an active request for specific information, checks all available keys, and selects the best-matched one. A value is information that a word contains.

$$A(Q, K, V) = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{6.2}$$

where Q is a query, K^T is transposed K (key) and V is the value, "$\cdot$" represents the dot products and $Softmax$ is the scaling function over the result vector.

Algorithm 1 SIMCSUM training

Input:

for each $d \in trainset$ **do**

▷ *Process each instance* d *of dataset* D *for tuples* I *of input* x *and targets for each task* $\mathcal{T}$

Create $I\langle x, y_{\mathcal{T}}\rangle$

end for

Initialize model parameters θ

Set maximum Epoch Ep

for epoch 1 to Ep **do**

for $b \in trainset$ **do**

▷ b *is a mini-batch containing I from* $trainset$

▷ SIMCSUM *consists of Encoder* E*, two Decoders* $D_{\mathcal{T}}$

Feed x to E and get the cross-attention

Feed $y_{\mathcal{T}}$ to $D_{\mathcal{T}}$

Feed the cross-attention to $D_{\mathcal{T}}$ [eq. (6.4)]

$t \leftarrow 0$

while θ_t is not converged **do**

$t \leftarrow t + 1$

Compute $\mathcal{L}(\theta)$ [eq. (6.6)]

Compute gradient $\nabla(\theta_t)$

Update $\theta_t \leftarrow \theta_{t-1} - \eta\nabla(\theta)$

end while

end for

end for

6.2.1.2 Multi-head Attention

As the MBART model contains multiple heads, all parallel attentions are concatenated to generate multi-head attention scaled with a weight matrix W. It helps the model to attend to multiple pieces of information simultaneously.

$$MH(Q, K, V) = Concat(A_1, \cdots, A_h) \cdot W^O \tag{6.3}$$

6.2.1.3 Cross-attention

Cross-attention connects the encoder and decoder and provides the decoder with a weight distribution at each step, indicating the importance of each input token in the current context. Cross-attention combines asymmetrically two separate embedding sequences of the same dimension, unlike self-attention, where input is a single embedding sequence. In cross-attention, one of the sequences serves as a query, while the other as a key and value.

As the SIMCSUM model has dual decoders, we concatenate the cross-attention of both decoders to generate the shared cross-attention.

$$A(E, D_{\mathcal{T}}) = Concat(Softmax(\frac{D_{\mathcal{T}} \cdot E^T}{\sqrt{d_k}}) \cdot E) \tag{6.4}$$

where E is the encoder representation, $D_{\mathcal{T}}$ is the task-specific decoder contextual representation, and d_k is the model size.

6.2.2 Training Objective

We train our model end-to-end to maximize the conditional probability of the target sequence given a source sequence. We define the task-specific loss as follows.

$$\mathcal{L}_{\mathcal{T}}(\theta) = \sum_{n=1}^{N} \log P(y_{\mathcal{T}_t} |\, y_{\mathcal{T}_{<t}}, x; \theta) \tag{6.5}$$

where x represents the input, y is the target, N is the mini-batch size, t is the time step and θ denotes to learnable parameters. We define the total loss of our model by task-specific losses where $\lambda_{\mathcal{T}}$ is an assigned weight to each task.

$$\mathcal{L}(\theta) = \sum \lambda_{\mathcal{T}} \cdot \mathcal{L}_{\mathcal{T}}(\theta) \tag{6.6}$$

6.3 Experiments

6.3.1 Datasets

We use two non-synthetic cross-lingual scientific summarization datasets: WIKIPEDIA and SPEKTRUM. We construct a synthetic WIKIPEDIA dataset for the simplification task by applying Keep-It-Simple (KIS) (Laban et al., 2021). To create the simplified WIKIPEDIA, we fine-tune KIS on WIKIPEDIA English articles as KIS is an unsupervised model and does not require parallel data. The simplified WIKIPEDIA consists of the original English articles paired with simplified English articles. We perform English text simplification because most of the

simplification work has been done in the English language (Al-Thanyyan and Azmi, 2021), and very few studies cover the German language (Aumiller and Gertz, 2022; Weiss and Meurers, 2018; Hancke et al., 2012) for children and dyslexic persons (not suitable for scientific simplification). Moreover, most of the work focuses on lexical or sentence level (Sun et al., 2021). To the best of our knowledge, KIS is the only SOTA paragraph-level unsupervised simplification model.

6.3.1.1 Split and Usage

We use WIKIPEDIA for training, validation and testing (80/10/10), while we use SPEKTRUM for zero-shot adaptability as a case study. All pre-trained language models (PLMs) are fine-tuned on WIKIPEDIA where each instance I in the training set consists of $< x, y >$ where x is the input English text, and y is the target German summary. The pipeline-based baseline is also trained on the WIKIPEDIA dataset. The first module - KIS accepts English articles as input and generates the simplified articles as output. Then the second module - MBART accepts the generated simplified English articles as input and German summaries as a target. The SIMCSUM model is trained on WIKIPEDIA where each instance I in the training set contains $< x, y_{sim}, y_{sum} >$ where x denotes the input English article and y_{sim} refers to the simplified English article and y_{sum} is the target German summary.

6.3.2 Models

Almost all cross-lingual multitask learning models in §3.2.1 are based on translation and summarization, and none of them applies simplification. So we select several SOTA PLMs that accept long input texts as baselines. We fine-tune the following baselines: (1) MT5 (Xue et al., 2021), (2) MBART (Liu et al., 2020), (3) PEGASUS (Zhang et al., 2020a), (4) LongFormer Encoder-Decorder (LONG-ED) (Beltagy et al., 2020), and (5) XLSUM (Hasan et al., 2021) and (6) BIGBIRD (Zaheer et al., 2020). In addition, we define a baseline, Simplify-Then-Summarize, based on KIS and MBART models as a pipeline. We report it as KIS-MBART in our experiments. We set $\lambda_{Sum} = 0.75$ for the SIMCSUM model based on the best results on the WIKIPEDIA validation set.

6.3.3 Training and Inference

6.3.3.1 Libraries and Hardware

We train all models with Pytorch[1], Hugging Face[2] integrated with DeepSpeed[3] for parallel model training with ZeRO-2. We apply ZeRO-2[4] to enable model parallelism. ZeRO-2 reduces the memory footprints for gradients and optimizer because it shards the optimizer states and gradients across GPUs. For all models, we complete training and inference on 4 Tesla P40 GPUs each with 24GB memory.

6.3.3.2 Hyperparameters

We fine-tune all models for a maximum of 25 epochs and average the results of 5 runs for each model. We use a batch size of 4-16, depending on the model size. We use a learning rate (LR) of $5e^{-5}$ and 100 warm-up steps to avoid over-fitting of the fine-tuned models. We use the Adam optimizer with a LR linearly decayed LR scheduler. The encoder language is set to English, and the decoder language is German.

We adopt similar settings for SIMCSUM as used for baselines, except for the batch size fixed to 4. We only generate tokens from the Summarization decoder side in the inference period. We use beam search of size 5 and a tri-gram block during the decoding stage to avoid repetition.

6.3.3.3 Training Time

With DeepSpeed, the MBART model takes 1 day and 17 hours, the MT5 model takes 10 hours, the PEGASUS model takes 1 day and 8 hours, the XLSUM model takes 1 day and 3 hours, the LONG-ED model takes almost 4 days, and the BIGBIRD model takes 2 days to complete 25 epochs. The SIMCSUM model takes 2 days to complete 25 epochs.

6.3.4 Automatic Evaluation

We evaluate all models with three automatic metrics - ROUGE, BERT-SCORE and FRE. For human evaluation, we consider **relevance**, **fluency** and **readability**.

[1] https://pytorch.org/
[2] https://huggingface.co/
[3] https://www.microsoft.com/en-us/research/project/deepspeed/
[4] Initially, we used ZeRO-3 offload with FP16 evaluation, and the training became quite slow as it consumes a lot of time for offloading during evaluation.

6.3.4.1 Human Evaluation

We conduct a human evaluation to compare the outputs of SIMCSUM with MBART (baseline) for the same linguistic properties. It is worth mentioning that human evaluation of long cross-lingual scientific text is challenging and costly because it requires bi-lingual annotators with a scientific background.

Due to budget and time constraints, we opted for two annotators. One of the annotators was working as a Hiwi in the NLP Group at HITS, Germany and studying at Heidelberg University. One of his class fellow volunteered for the annotations. Our annotators were two university students from the Computational Linguistics department with fluent German and English skills. Further details about the guidelines for human evaluation are presented in Appendix 6.A.

MODELS	ROUGE-1	ROUGE-2	ROUGE-L	BERT-SCORE	FRE
GOLD	-	-	-	-	36.93
MT5	26.79	12.65	23.40	69.12	45.42
MBART	31.43	13.20	25.12	70.52	44.67
PEGASUS	29.30	13.93	24.62	69.83	43.39
XLSUM	31.91	13.30	24.14	70.04	37.83
BIGBIRD	29.23	13.72	24.60	69.19	41.42
LONG-ED	15.11	06.82	13.67	63.94	24.48
KIS-MBART	32.02	12.39	24.72	70.52	45.76
SIMCSUM	**34.50**†	**14.36**†	**25.85**†	**71.60**†	**46.86**†

Table 6.1: The WIKIPEDIA results for all baselines and SIMCSUM. GOLD denotes the reference summaries. Underline refers to the best baseline results, and **bold**† denotes the best overall results with significant improvements ($p < .001$).

6.4 Wikipedia Results

We report F-SCORE of ROUGE and BERT-score and FRE of all models in Table 6.1. The first block includes the fine-tuned PLMs models, the second block presents the pipeline baseline, and the last block includes the SIMCSUM model. From Table 6.1, we find that the SIMCSUM model outperforms all baselines for every metric. We compute the statistical significance of the results with the Mann-Whitney two-tailed test for a p-value ($p < .001$).

Interestingly, the WIKIPEDIA reference summaries are not as much simplified as the SPEKTRUM summaries; still, the SIMCSUM model performs better on the WIKIPEDIA dataset than the baselines. We interpret that the simplification auxiliary task helps the SIMCSUM model to learn better contextual representation and produce more relevant German words. We infer from the results that joint learning of simplification and cross-lingual summarization improves the quality of summaries.

Among the baselines, almost all models demonstrate comparable performance except the LONG-ED model. For ROUGE-1, the KIS-MBART pipeline performs better than other models, however, the MBART and XLSUM models' performance are also similar. The PEGASUS model takes the lead for ROUGE-2, and the MBART model shows higher performance for ROUGE-L. The KIS-MBART pipeline and the MBART model take the lead for BERT-SCORE among the baselines. For FRE, a score between $30 - 50$ presents the readability level best understood by college graduates. The WIKIPEDIA summaries fall in this range. For FRE, the KIS-MBART pipeline performs better than the other baselines. Interestingly, almost all baselines except the BIGBIRD and XLSUM models demonstrate good performance. Comparing the KIS-MBART pipeline and the MBARTmodel, the KIS-MBART pipeline performs slightly better than the MBART model for ROUGE-1 and FRE, equal for BERT-SCORE and slightly lower for ROUGE-2 and ROUGE-L. We infer that it is due to the impact of the simplification module in the KIS-MBART pipeline.

MODELS	ROUGE-1	ROUGE-2	ROUGE-L	BERT-SCORE	FRE
GOLD	-	-	-	-	40.76
MT5	09.21	00.75	06.50	58.52	38.18
MBART	16.16	01.47	<u>13.89</u>	62.11	39.17
PEGASUS	11.49	00.95	08.01	60.56	37.93
XLSUM	17.10	<u>01.63</u>	09.79	<u>62.25</u>	33.83
BIGBIRD	12.28	01.04	08.65	59.97	36.24
LONG-ED	01.32	00.11	01.18	51.85	30.16
KIS-MBART	<u>17.33</u>	01.61	12.97	61.83	<u>39.21</u>
SIMCSUM	**18.88**†	**01.82**†	**14.16**†	**63.47**†	**40.03**†

Table 6.2: The SPEKTRUM results for all baselines and SIMCSUM. GOLD denotes the reference summaries. <u>Underline</u> refers to the best baseline results, and **bold** with † denotes the best overall results with significant improvements ($p < .001$).

6.5 Case Study: Spektrum

Table 6.2 presents the results of all models on the SPEKTRUM dataset. The first block includes the fine-tuned PLMs models, the second block presents the pipeline baseline, and the last block includes the SIMCSUM model. We find a similar pattern that the SIMCSUM model outperforms all baselines. We also compute the statistical significance of these results with the same procedure. The SPEKTRUM results are on the lower side compared to the WIKIPEDIA results due to the zero-shot adaptability, especially for ROUGE-2. We infer that it is due to the impact of the computation method of ROUGE score as it is an n-gram-based metric (Ng and Abrecht, 2015). The SPEKTRUM summaries have higher FRE scores than the WIKIPEDIA summaries. Interestingly, we find that all baselines perform lower than the GOLD summaries. However, the SIMCSUM score is similar to the GOLD summaries. Comparing the performance of the MBART model and the KIS-MBART pipeline, the KIS-MBART pipeline performs slightly lower than the MBART model for all scores except ROUGE-1 because only the WIKIPEDIA dataset is used for fine-tuning of both models in the KIS-MBART pipeline.

6.5.1 Human Evaluation

We provide 30×2 (for each model) random summaries of MBART and SIMCSUM outputs with their original texts. We ask two annotators to evaluate each document for three linguistic properties - **relevance**, **fluency** and **simplicity** (readability) on a Likert scale from $1 - 5$ (1:worst | 2:bad | 3:ok | 4:good | 5:best). The first five samples are used to calibrate the annotations of annotators, and then each annotator provides independent judgments on the rest of the samples. We compare the SIMCSUM and MBART outputs for analyzing linguistic qualities because the SIMCSUM's architecture is based on the MBART model.

MODELS	FLUENCY (α)	RELEVANCE (α)	SIMPLICITY (α)
MBART	2.28 (0.64)	1.64 (0.70)	1.86 (0.56)
SIMCSUM	2.62 (0.87)	2.76 (0.78)	2.88 (0.81)

Table 6.3: The SPEKTRUM human evaluation for MBART and SIMCSUM. The average scores (Krippendorff's α) for each linguistic feature are presented here.

Table 6.3 shows the human evaluation results. The samples used for calibration are not used for computing the scores. We compute the inter-rater reliability by using Krippendorff's α[5]. We find that SIMCSUM improves the **fluency**, **relevance** and **readability** of outputs. To

[5] https://github.com/LightTag/simpledorff

conclude, both automatic and human evaluation show the dominance of the SIMCSUM model over the baselines.

6.5.1.1 Reproducibility

We find that the results of Table 6.3 and 5.4 are different under the same evaluation criteria. We conducted human evaluations independently for each experiment with different annotators and their count. It helps to mitigate biases in evaluation process and ensures diversity in opinion. However, differences in human evaluation scores can stem from varying annotator backgrounds, understanding of evaluation criteria, and subjective perceptions of summary quality. Such variations are expected and do not necessarily indicate lack of reproducibility or unreliability in the experimental setup. Reproducibility in deep learning experiments is primarily concerned with the ability to replicate results using the same data, models, and evaluation metrics. Automatic metrics like ROUGE and BERT-SCORE demonstrate that the MBART model perform similarly under controlled conditions (see Tables 5.3 and 6.2), which ensures reproducibility. However, differences in human evaluation scores, while notable for qualitative analysis, do not invalidate the reproducibility of objective findings (ROUGE and BERT-SCORE). They highlight the complementary nature of objective and subjective evaluations in understanding model performance.

6.5.2 SimCSum Examples

We present some examples showing the difference between the SIMCSUM and MBART models from the SPEKTRUM outputs (in the last pages of this chapter). All the summaries are translated via Google translate.

6.6 Analysis: Spektrum

We explore three further dimensions along with extended readability for in-depth analysis: lexical diversity, syntactic properties and error types to determine the quality of generated summaries. The lexical diversity and readability scores are computed over all SPEKTRUM's reference summaries (Gold) and outputs of MBART and SIMCSUM. The gold summaries' score is a guideline for how similar the models' outputs are to gold summaries.

FEATURES	GOLD	mBART	SIMCSUM
Shannon Entropy Estimation ↓	4.25 (0.04)	4.26 (0.1)	4.25 (0.1)
Measure of Textual Lexical Diversity ↑	201 (41.4)	65.13 (33.3)	91.75 (33.1)

Table 6.4: Lexical diversity features' average scores (standard deviation).

6.6.1 Lexical Diversity

Lexical diversity estimates how language is distributed overall and how much cohesion is present in the text as synonyms. It is a good indicator of the readability of a text. We calculate Shannon Entropy Estimation (SEE) and Measure of Textual Lexical Diversity (MTLD) to find lexical diversity. SEE presents a text's "informational value" and language diversity. It is a language-dependent feature, and its value varies for different languages. Higher SEE scores suggest higher lexical diversity. We aim to get a similar SEE as Gold summaries. Table 6.4 shows the SEE scores of mBART and SIMCSUM that are similar to Gold summaries suggesting the similar informational value of all summaries.

Table 6.4 presents MTLD scores of mBART and SIMCSUM. The gold summaries have the highest scores, while SIMCSUM is the second highest and mBART has the lowest score. These scores suggest that the lexical richness of all three groups is not similar, in contrast to SEE results. However, the SIMCSUM outputs are more lexically diverse than the mBART outputs. We infer from the improved SIMCSUM scores that joint learning of simplification and cross-lingual summarization impacts word generation. These results also suggest that the MTLD scores provide a better estimation of lexical diversity for our summaries.

FEATURES	GOLD	mBART	SIMCSUM
Coleman Liau Index ↓	18.45 (1.7)	21.64 (4.7)	20.96 (4.8)
Automated Readability Index ↓	18.99 (2.4)	21.07 (5.5)	20.26 (5.2)

Table 6.5: Readability features' average scores (standard deviation).

6.6.2 Readability Scores

Readability scores measure comprehension levels of the text. We calculate Coleman Liau Index (CLI) and Automated Readability Index (ARI) as these do not rely on syllables. CLI computes scores on word lengths. The ARI computes scores on characters, words and sen-

tences. For both CLI and ARI, the lower score is better as it shows the ease of reading and understanding. We interpret from Table 6.5 that Gold summaries have the lowest score, SIMCSUM has the second-lowest score, and mBART has the highest score. We infer from the improved SIMCSUM scores that joint learning of simplification and cross-lingual summarization impacts both word and sentence level because the CLI only focuses on words, while the ARI includes sentences also.

6.6.3 Syntactic Properties

Syntactic analysis elaborates on how words and phrases are related in a sentence structure. We perform it with constituency trees on 25×2 (for each model) random summaries from mBART, SIMCSUM and the gold summaries. The total number of sentences for mBART is 70, for SIMCSUM is 80 and for gold is 131.

We use Stanza[6] to extract dependency relations and Stanford Parser[7] to extract constituency trees for each summary. Before tree generation, we replace all German umlauts (ä, ö, ü and ß) in the summaries with their replacements (ae, oe, ue and ss) due to encoding issues of the Stanford Parser. Table 6.6 presents four syntactic features defined as follows.

1. **Average Sentence Length** It is the number of tokens in the sentences averaged over the number of sentences in a summary.
2. **Average Dependency Distance** It is the averaged dependency distance over the sentences, which means the distance between the dependency heads and their dependents.
3. **Average Dependents per Word** It computes the average number of dependents for each word.
4. **Average Tree Height** For computing the average tree height of a summary, we calculate the height of every tree and average it over the sentences.

We infer from the average sentence length (ASL) that the SIMCSUM model produces shorter sentences than the mBART model and gold summaries, which exhibits syntactic simplicity. A small average dependency distance (ADD) shows that words with a dependency relation are close together, making the text easier to understand. Table 6.6 shows that the SIMCSUM summaries have a smaller average dependency than the mBART summaries, much closer to gold summaries. Fewer dependents per word (ADW) make a text less ambiguous and thus easier to follow. Table 6.6 shows the SIMCSUM outputs have fewer dependents than the mBART

[6] https://stanfordnlp.github.io/stanza/constituency.html
[7] https://nlp.stanford.edu/software/lex-parser.shtml

FEATURES↓	GOLD	MBART	SIMCSUM
Average Sentence Length	24.09 (4.2)	24.15 (7.2)	20.97 (6.5)
Average Dependency Distance	3.60 (0.3)	4.16 (1.1)	3.91 (1.1)
Average Dependents per Word	0.93 (0.04)	0.95 (0.02)	0.94 (0.04)
Average Tree Height	8.32 (0.7)	8.72 (1.5)	8.57 (1.5)

Table 6.6: Syntactic features' average scores (standard deviation).

outputs and are similar to gold summaries. The average tree height (ATH) explains the syntactic structural complexity of a sentence. Table 6.6 shows that SIMCSUM outputs are less structurally complex than the MBART outputs, however, gold summaries have the least average tree height. We infer from the syntactic analysis that joint learning of simplification and cross-lingual summarization positively impacts the syntactic properties of summaries.

6.6.4 Error Analysis

To further explore the challenges of improving cross-lingual science summaries, we randomly select 25×2 (for each model) summaries from the SIMCSUM and MBART outputs. We find three main categories of errors in the manual inspection. We first present the informal guidelines for the error analysis and then present the analysis in Table 6.7 and some examples.

6.6.4.1 Guidelines for Error Analysis

To find the errors in the MBART and SIMCSUM outputs, we compare them to each other, to the SPEKTRUM German gold summary and the original English text.

1. **Non-German Words.** To find them, it is sufficient to read through our model outputs and look up any unknown words. If one of the unknown words turns out to be a non-German word, we mark them in red.

2. **Wrong name entities.** We find wrong-name entities by comparing the names in both system outputs to the reference summary. If the names differ, we verify with the original text that they refer to the same person and thus represent a mistake by the model, and we mark them in blue.

3. **Unfaithful information.** We find new/unfaithful information by looking up every piece of information in the model outputs in the reference summary. We search for this in-

formation in the original text, and if it is not present there, it is clear that the model produced new information that is not faithful to the source text. We mark this information in orange.

ERROR TYPES	MBART	SIMCSUM
Non-German words	83	35
Wrong name entities	1	2
Unfaithful information	3	3

Table 6.7: Error occurrences for MBART and SIMCSUM summaries. The summaries may contain multiple errors.

6.6.4.2 Non-German Words

It is the error type where the models either produce non-existent German words or partially English-German or another language words. We find that MBART is more prone to produce such errors. We infer that it is due to the imbalance between the pre-trained and fine-tuned dataset sizes. These models are pre-trained on many languages and usually fine-tuned on comparatively smaller data. SIMCSUM tends to produce fewer errors due to data augmentation (simplification data) during the training.

6.6.4.3 Wrong Name Entities

It is the error type where the models produce wrong name entities, such as cities or country names and persons' first and last names. We find that both models tend to produce such errors, however, the percentage of such errors is quite low. We infer that the models overestimate or underestimate the probability of word sequences present in data.

6.6.4.4 Unfaithful Information

It is the error type where we find some (new) information in generated summaries that is not faithful to the source documents. We infer that this error is caused by long inputs where the model tends to hallucinate and generates some content that cannot be verified from the source. We find that SIMCSUM makes similar errors as MBART for this error type.

6.6.4.5 Error Examples

In this section, we present some output examples to show the performance difference between the SIMCSUM and MBART in terms of errors. All the summaries are translated via Google translate. Further details about the error analysis are presented in Appendix 6.B.

Target: im freigehege aufzuwachsen und sich dort im schlamm zu suhlen, stärkt offenbar das immunsystem von schweinen : verglichen mit artgenossen, die in einem stall gehalten wurden, hatten freilandschweine eine höhere anzahl von gesundheitsfördernden darmbakterien. zu diesem ergebnis kommen jetzt wissenschaftler um denise kelly von der university of aberdeen. auch gene, die auf entzündungsreaktionen hindeuten, waren bei tieren, die im außenbereich lebten, wesentlich weniger aktiv. [...]
Translation: Growing up in an outdoor enclosure and wallowing in the mud there apparently strengthens the immune system of pigs : compared to conspecifics that were kept in a pen, free-range pigs had a higher number of health-promoting intestinal bacteria. scientists led by denise kelly from the university of aberdeen are now coming to this conclusion. genes that indicate inflammatory responses were also significantly less active in animals that lived outdoors. [...]

MBART: als fruehmikrobielle kolonisation wird dieabdeckung der brustdruese von schweinen bei der fruehen entwicklung des erwachsenen menschen bezeichnet. sie wird in der medizin, der mikrobiologie und der pathologie erforscht. early microbial colonization of the gut reduced the incidence of infectious, entzuendlichen und autoimmunischen diseasen reduced the incidence of infectious, inflammatory and autoimmune diseases.
Translation: the covering of the mammary gland of pigs during the early development of the adult human is referred to as early microbial colonization. it is researched in medicine, microbiology and pathology. early microbial colonization of the gut reduced the incidence of infectious, inflammatory and autoimmune diseases reduced the incidence of infectious, inflammatory and autoimmune diseases.

SIMCSUM: die fruehzeitshygiene ist ein begriff aus der entwicklungsbiologie und bezeichnet das phaenomen, dass die fruehzeitliche besiedlung des darmes durch krankheitserreger verhindert wird. die fruehzeitshygiene unterscheidet sich von anderen entwicklungsbiologischen forschungsgebieten wie der entwicklungsphysiologie, der haematologie und der palaeontologie dadurch, dass in ihrer gesamtheit zur fruehen entwicklungsphase die mikrobielle vielfalt des darmes zaehlt.
Translation: early hygiene is a term from developmental biology and describes the phenomenon that prevents early colonization of the intestines by pathogens. early hygiene differs from other developmental biological research areas such as developmental physiology, haematology and palaeontology in that the microbial diversity of the intestine counts in its entirety for the early development phase.

Figure 6.2: An example where the MBART model produces non-German words (marked as red), while the SIMCSUM model generates a summary focusing its attention on hygiene, which is not much relevant to the topic.

Target: nicht nur stress und genetische faktoren tragen zu essstörungen bei, sondern auch vorübergehende hungerphasen. nun stießen forscher vom europäischen zentrum für hirnforschung in rom auf einen möglichen grund dafür - und zwar in den gehirnen von labormäuse, die sich auch durch schmerzen nicht vom naschen abbringen ließen. die psychologin rossella ventura und ihre kollegen trainierten mäuse darauf, in einer von zwei kammern nach einer belohnung in form von schokolade zu suchen. dann versetzten sie den tieren ausgerechnet in dem raum, wo die leckerei lockte, schmerzhafte elektroschocks. die mäuse mieden fortan die schokolade und hielten sich vorwiegend in der anderen kammer auf - außer jene, die zuvor einige zeit lang kaum futter bekommen hatten. [...]
Translation: Not only stress and genetic factors contribute to eating disorders, but also periods of starvation. now researchers from the european center for brain research in rome have discovered a possible reason for this - namely in the brains of laboratory mice, which were not deterred from snacking even by pain. psychologist rossella ventura and her colleagues trained mice to look for a chocolate reward in one of two chambers. then they gave the animals painful electric shocks in the room where the treat lured them. From then on, the mice avoided the chocolate and mainly stayed in the other chamber - except for those that had hardly received any food for a while. [...]

MBART: eating disorders (engl. “ess-stoerungen”) nennt man eine vielzahl von ursachen und mechanismen, die zur auspraegung von essstoerungen fuehren koennen.
Translation: eating disorders are a number of causes and mechanisms that can lead to the development of eating disorders.

SIMCSUM: als essverhalten oder magersucht werden die verhaltensweisen von menschen und tieren bezeichnet, die excessive (vergoren) oder chronisch (z.b. nach unfaellen oder erkrankungen) zu es verlangen oder dieses verhalten unterdruecken. essverhalten kann zu verhaltensstoerungen fuehren, die ursache fuer die entstehung von essstoerungen sind innere oder aeussere ursachen des essverhaltens.
Translation: eating habits or anorexia are the behaviors of people and animals that demand excessive (fermented) or chronic (e.g. after accidents or illnesses) or suppress this behavior. eating behavior can lead to behavioral disorders, the cause of eating disorders are internal or external causes of eating behavior.

Figure 6.3: An example where both models generate non-German words (marked as red). However, the summary produced by the MBART model is quite short.

Target: der tod kam auf dem höhepunkt der macht : cangrandei. della scala, der berühmteste spross der scaliger aus verona, hatte in zahlreichen kriegerischen auseinandersetzungen die herrschaft über ganz venetien an sich gerissen. doch kurz nach seinem triumphalen einzug in die norditalienische stadt treviso am 18. juli 1329 begannen mit durchfall und erbrechen die ersten symptome einer tödlichen erkrankung. [...]
Translation: death came at the height of power : cangrandei. della scala, the most famous offspring of the scaliger family from verona, seized control of all of veneto in numerous military conflicts. but shortly after his triumphant entry into the northern italian city of treviso on july 18, 1329, the first symptoms of a fatal illness began with diarrhea and vomiting. [...]

MBART: alberto cangrande (* 1291 oder 1329 in verona ; † 29. oktober 1329 ebenda) war ein italienischer dichter, schriftsteller und politiker der flämischen republik. in der zeit der italienischen republik bestieg er von seinen brudern die dynastie der scaligeri. bei der eroberung der elfenbeinküste 1311 wurde er zum maßgebenden gründer der faction in rom. ab 1314 war er herrscher im königreich venedig. cangrande gilt als vater der italienischen
Translation: alberto cangrande (b. 1291 or 1329 in verona ; † october 29, 1329 ibid) was an italian poet, writer and politician of the flemish republic. in the time of the italian republic he ascended from his brothers the dynasty of the scaligeri. when he conquered the ivory coast in 1311, he became the main founder of the faction in rome. from 1314 he was ruler of the kingdom of venice. cangrande is considered the father of the italian

SIMCSUM: cangrande della scala oder cangrande cangrande (* 9. märz 1291 in verona ; † 22. juli 1329 ebenda) war ein begründer der von den scaligeri beeinflussten dynastie, der unter den herrschern von verona von 1308 bis 1387 regierte. der sohn des veronaer stadtbegründers und stadtmeisters alberto della scala (genannt „ albert “ oder „ adelsmann “), der nur als unterstellt war, wurde nach seinem tod 1301 zum bruder von alberto verbannte.
Translation: cangrande della scala or cangrande cangrande (born march 9, 1291 in verona ; † july 22, 1329 ibid) was a founder of the dynasty influenced by the scaligeri, who reigned under the rulers of verona from 1308 to 1387. the son of verona's founder and mayor alberto della scala (called " albert " or " nobleman "), who was only subordinate, was banished to alberto's brother after his death in 1301.

Figure 6.4: An example where the MBART model produces a wrong named entity, while the SIMCSUM model generates the correct named entity but with a wrong alias for this person.

Target: inoffizieller weltrekordhalter im nicht - frieren trotz eisiger kälte ist der niederländer wimhof : er lief bereits ziemlich nackt halbmarathons nördlich des polarkreises und nahm das längste eisbad seit beginn der aufzeichnungen - ohne sich ernsthaft zu beklagen. wissenschaftliche untersuchungen des eismannes hatten bisher keine körperlichen auffälligkeiten gefunden, die erklären könnten, warum hof mit kälte offensichtlich besser umgeht als der durchschnittsmensch. [...]
Translation: the unofficial world record holder for not freezing despite the icy cold is the dutchman wimhof : he ran half marathons north of the arctic circle quite naked and took the longest ice bath since records began - without seriously complaining. scientific examinations of the ice cream man had not found any physical abnormalities that could explain why hof obviously deals with the cold better than the average person. [...]

MBART: wim hof (* 20. april 1959) ist ein niederländischer sportler, der sich unter anderem mit eisschwimmern beschäftigt.
Translation: wim hof (born april 20, 1959) is a dutch athlete who is involved with ice swimmers, among other things.

SIMCSUM: wim hof (* 20. april 1959), auch als eismann bekannt, ist ein niederländischer swimmer und ein ehemaliger olympiateilnehmer. er hält seit dem 28. februar 2015 den bislang längsten direkten kontakt menschlichen körpers mit eis.
Translation: Wim Hof (born April 20, 1959), also known as Eismann, is a Dutch swimmer and a former Olympic competitor. since february 28, 2015, he has been in the longest direct contact between the human body and ice.

Figure 6.5: An example where both MBART and SIMCSUM produce unfaithful information (marked in orange) to the original text.

6.7 Limitations

The SIMCSUM model jointly trains on two high-level NLP tasks, which takes slightly more time than its base model - MBART, as it has some additional parameters to learn during the training. Moreover, the SIMCSUM model is trained on synthetic simplification data, which may create a dependency on the simplification model - KIS. We also find during error analysis that both MBART and SIMCSUM models have problems (repetition or unfaithful information) with long inputs, which need further investigation in the future.

6.8 Summary

In this chapter, we explore the task of Cross-lingual Science Journalism with a novel Multitask model - SIMCSUM. The SIMCSUM combines two high-level NLP tasks, simplification and cross-lingual summarization. The SIMCSUM model jointly trains for reducing linguistic complexity and cross-lingual abstractive summarization. Our empirical investigation shows the significantly superior performance of the SIMCSUM model over the SOTA baselines on two non-synthetic cross-lingual scientific datasets, also indicated by human evaluation. Furthermore, our in-depth linguistics analysis shows how multitask learning in the SIMCSUM model

has lexical and syntactic impacts on the generated summaries.

Target: für menschen ist der einbeinige stand immer eine wackelpartie , vor allem bei geschlossenen augen . um nicht umzukippen , müssen die muskeln permanent die leichten schwankungen ausgleichen . nicht so bei flamingos : sie kostet es weniger kraft , auf einem bein zu stehen als auf zweien . deswegen können sie auch beruhigt auf einem bein ein schläfchen machen , ohne dabei umzukippen . das zeigt jetzt eine untersuchung von young - hui chang und lena ting . die wissenschaftler haben dazu messungen mit flamingoküken angestellt und zwei tote exemplare aus einem zoo seziert . [...]
Translation: standing on one leg is always a shaky game for humans, especially with closed eyes. in order not to tip over , the muscles have to constantly compensate for the slight fluctuations . Not so with flamingos: it takes less strength for them to stand on one leg than on two. that's why you can take a nap on one leg without tipping over. this is now shown by a study by young - hui chang and lena ting . the scientists took measurements with flamingo chicks and dissected two dead specimens from a zoo . [...]

MBART: die biogerontologie (von griech. bíos „ leben " und lat. protes „ befestigt ") beschäftigt sich mit dem verhältnis von körpergewicht und körperhaltung. so untersucht die biogerontologie das verhältnis von körpergewicht und körperhaltung : welche gewichtszustände und welche muskeln notwendig sind, um einen pilz, der einer krankheit ausgesetzt ist, zu tragen? wie sehr wichtig es ist, den jeweiligen körpergewichtszustand zu messen. dies wird in der
Translation: biogerontology (from greek bíos "life" and lat. protes "fixed") deals with the relationship between body weight and posture. this is how biogerontology examines the relationship between body weight and posture: which weight conditions and which muscles are necessary to carry a fungus that is exposed to a disease? how very important it is to measure the respective body weight condition. this will be in the

SIMCSUM: flamingos (phoenicopteridae) oft sitzen auf einem bein, um eine muskelkontraktion zu erzeugen. sie haben die fähigkeit, das körpergewicht aufzunehmen und zu tragen, ohne dabei eine muskelaktive wirkung auszuüben.
Translation: Flamingos (phoenicopteridae) often perch on one leg to produce muscle contraction. they have the ability to absorb and carry body weight without exerting a muscle-active effect.

Figure 6.6: An example where the MBART model skips the salient point in the original text and focuses on biogerontology in the summary which has no relevancy in this case. So this information is not faithful to the original text (marked in orange). While the SIMCSUM model captures the salient idea in the original text and generates a summary that better captures that idea.

Target: eine besondere strategie, seinen fressfeinden zu entkommen, hat der bombardierkäfer pheropsophus jessoensis entwickelt. er wehrt sich sogar noch, nachdem er eigentlich schon verloren hat, und das sogar recht erfolgreich, wie jetzt shinji sugiura und takuya sato von der universität kobe in den biology letters berichten. gerade verschluckt, schießt der käfer im magen der kröte einen strahl heißes, ätzendes sekret aus der afterdrüse - eine strategie, die den bombardierkäfern ihren namen eingebracht hat. normalerweise allerdings dient die waffe zur abschreckung. [...]
Translation: the bombardier beetle pheropsophus jessoensis has developed a special strategy for escaping its predators. he even defends himself after he has actually already lost, and even quite successfully, as shinji sugiura and takuya sato from the university of kobe now report in the biology letters. just swallowed, the beetle in the toad's stomach shoots a jet of hot, caustic secretion out of the anal gland - a strategy that has given the bombardier beetle its name. normally, however, the weapon serves as a deterrent. [...]

MBART: unter zoologischer kompatibilitaet versteht man in der biologie die faehigkeit von beutetieremischlingen, nach einem laubabschweif von artgenossen in den verdauungstrakt von artgenossen zu entweichen. die oekologische kompatibilitaet ist ein wesentlicher bestandteil der erfolgreichen entstehung von beutegreifern.beim laubabschweifexperiment mit hilfe des insekten- pheropsophus jessoensis aus zwei kaeferarten, dem bufo japonicus und dem b. torrenticola, konnte nachgewiesen werden, dass der kaefer pheropsophus jessoensis unter laboratorischen bedingungen eine in den magensaftgastropfen der jagdsekten einschleudern kann.
Translation: in biology, zoological compatibility means the ability of hybrid prey animals to escape into the digestive tract of conspecifics after the leaves have drifted away from conspecifics. ecological compatibility is an essential part of the successful emergence of predators.in the leaf-dropping experiment with the help of the insect pheropsophus jessoensis from two beetle species, the bufo japonicus and the b. torrenticola, it could be proven that the beetle pheropsophus jessoensis can inject one into the gastric juice droplets of the hunting sects under laboratory conditions.

SIMCSUM: als wehrverhalten oderverteidigungsverhalten bezeichnet man in der verhaltensbiologie verschiedene verhaltensweisen von tieren, die wehren gegen die fressfeinde oder andere fresstiere bilden. nicht zu verwechseln ist das wehrverhalten mit der phylogenetik, die sich aus der evolutionsgeschichte entwickelt hat.
Translation: in behavioral biology, defensive behavior is the term used to describe various behaviors of animals that defend themselves against predators or other predators. the defensive behavior should not be confused with the phylogenetics that have developed from the history of evolution.

Figure 6.7: An example where the performance of the MBART model and the SIMCSUM model are comparable and reasonable. However, in this case, we prefer the summary generated by the SIMCSUM model because it explains the basic idea of the beetle defending itself against predators quite well. While the summary produced by the MBART model is filled with some unnecessary and complex words and some out-of-context information (marked in orange).

6.A Guidelines for Human Evaluation

6.A.1 Task Description

We provided annotators with 30 examples of documents paired with a reference summary and two system-generated summaries. The models' identities had not been revealed. The annotators had to rate each model summary for the following linguistic properties after reading

Target: manche farben bleiben besser im gedächtnis haften als andere, doch auf die erinnerung an den entsprechend gefärbten gegenstand hat das keinen einfluss. das zeigt nun eine studie von psychologen um christof kuhbandner von der universität regensburg. die ergebnisse machen deutlich, dass der gegenstand selbst und dessen eigenschaften vom gehirn nicht immer als einheit betrachtet werden. die wissenschaftler testeten für ihre studie das gedächtnis ihrer versuchspersonen. dabei variierten sie sowohl, was zu merken war, als auch, wie es dargeboten wurde. insbesondere präsentierten sie wörter und gegenstände in unterschiedlichen farben. [...]
Translation: some colors stick in the memory better than others, but this has no influence on the memory of the correspondingly colored object. this is now shown by a study by psychologists led by christof kuhbandner from the university of regensburg. the results make it clear that the object itself and its properties are not always viewed as a unit by the brain. for their study, the scientists tested the memory of their subjects. they varied both what was to be remembered and how it was presented. in particular, they presented words and objects in different colors. [...]

MBART: die farbpsychologie ist ein teilgebiet der psychologie und beschäftigt sich mit der wirkung von farben auf die wahrnehmung und das gedächtnis.
Translation: color psychology is a branch of psychology and deals with the effect of colors on perception and memory.

SIMCSUM: die farbe ist ein farbmuster, das uns wahrnehmungen und die sinneswahrnehmung durch das sehen vermittelt. die farbwahrnehmung gehört zu den kognitionswissenschaften. wir sehen sie nicht nur als eindimensionales spektrum möglicher farben, sondern auch als eindimensionales, oftmals auch als „ visuelles system “, das die sinneswahrnehmung mitanpasst. [...]

Translation: color is a color pattern that conveys perceptions and sensory perception to us through seeing. Color perception is one of the cognitive sciences. we see them not only as a one-dimensional spectrum of possible colors, but also as a one-dimensional, often also as a " visual system " that also adapts the sensory perception. [...]

Figure 6.8: An example where the performance of the MBART model is better than the SIMCSUM model. Here, the MBART model generates a summary that is too short, however, it recapitulates the main idea quite well. While the SIMCSUM model produces an incoherent and unfaithful summary (marked in orange).

the English document and the German summaries. The annotators were given a Likert scale from $1-5$ (1=worst, 2=bad, 3=neutral/ok, 4=good, 5=best). We asked annotators to use the first 5 examples to resolve the annotator's conflict and to find a common consensus for rating the linguistic aspects. However, the rest of the examples were annotated independently.

6.A.2 Linguistic Properties

We asked annotators to annotate each summary for the following linguistic properties.

Relevance. A summary delivers adequate information about the original text. Relevance determines the content relevancy of the summary.

Fluency. The words and phrases fit together within a sentence, and so do the sentences.

Fluency determines the structural and grammatical properties of a summary.

Simplicity. Lexical (word) and syntactic (syntax) simplicity of sentences. A simple summary should have minimal use of complex words/phrases and sentence structure.

6.B Guidelines for Error Analysis

To find the errors in the MBART and SIMCSUM outputs, we compare them to each other, to the SPEKTRUM German gold summary and the original English text.

1. **Non-German Words.** To find them, it is sufficient to read through our model outputs and look up any unknown words. If one of the unknown words turns out to be a non-German word, we mark them in red.

2. **Wrong name entities.** We find wrong-name entities by comparing the names in both system outputs to the reference summary. If the names differ, we verify with the original text that they refer to the same person and thus represent a mistake by the model, and we mark them in blue.

3. **Unfaithful information.** We find new/unfaithful information by looking up every piece of information in the model outputs in the reference summary. We search for this information in the original text, and if it is not present there, it is clear that the model produced new information that is not faithful to the source text. We mark this information in orange.

Part IV

Conclusions

Chapter 7

Conclusions

> "Every ending is a beginning. We just don't know it at the time."
>
> Mitch Album

In this thesis, we introduce a novel task - Cross-lingual Science Journalism as a use case of cross-lingual abstractive summarization, and we also incorporate text simplification in it. In this regard, we focus on collecting two new datasets and developing and evaluating models which show promising results, further assessed by different linguistic features. This chapter summarizes our contributions and discusses three possible future directions.

7.1 Contributions

Our contributions to this thesis are as follows:

1. **Collecting and Analyzing Datasets**

   We develop a systematic methodology to collect and verify non-synthetic cross-lingual datasets from online sources for Cross-lingual Summarization and Science Journalism - SPEKTRUM and WIKIPEDIA. The WIKIPEDIA dataset also contains a monolingual part. Moreover, the WIKIPEDIA dataset is publicly available for the summarization research community. The data collection steps for SPEKTRUM can be applied with minimal modification to any similar raw data to collect source articles. The data collection methodology for WIKIPEDIA can be applied to any pair of languages and a set of languages. We also conduct a thorough analysis based on various statistical and linguistic features to investigate the properties of our datasets, which provide a basic set of traditional features for future summarization and science journalism studies.

2. **Evaluating Datasets for Summarization**

   We evaluate our datasets for monolingual and cross-lingual summarization. For monolingual summarization, we consider extractive summarization methods as the baselines and compare the performance of three neural-based abstractive models against them. We create two pipelines as the baselines for cross-lingual summarization and train the same abstractive models with cross-lingual data. The cross-lingual abstractive models achieve better results than the baselines with 95% confidence for ROUGE scores. However, these models have two major limitations - fixed n-gram-based vocabulary and limited input-output size. We solve the vocabulary problem during our experiments by replacing it with BPE based vocabulary. However, these models could not accept longer input lengths; moreover, these models produce repetitive phrases in output, suggesting these models could not retain long-term dependencies.

3. **Developing and Evaluating SSR for Cross-lingual Science Journalism**

   To address the limited input-output size of abstractive models and to incorporate simplification as indicated by data analysis, we develop a component-based model - **S**ELECT, **S**IMPLIFY and **R**EWRITE (SSR). We train individual components in a pipeline fashion and compare SSR against SOTA models. The SSR model achieves higher results than the baselines for three evaluation metrics with 99% confidence, further indicated by human judgment and readability analysis. The ablated models and readability analysis indicate the positive impact of the SIMPLIFY component. However, the SIMPLIFY component is computationally expensive and takes a long time for training and inference. Moreover, as SIMPLIFY and REWRITE are based on similar PLMs, we can train them for a positive inductive transfer.

4. **Developing and Evaluating SimCSum for Cross-lingual Science Journalism**

   We further develop a novel multitask learning model for joint training of **Sim**plification and **C**ross-lingual abstractive **Sum**marization (SIMCSUM). It is an end-to-end model based on MBART architecture. To evaluate this model, we construct a synthetic simplification dataset from our WIKIPEDIA dataset and then train SIMCSUM with simplification and summarization data for the joint training. The SIMCSUM model achieves higher results than the baseline models with 99% confidence for ROUGE, BERT-SCORE and FRE, further indicated by human evaluation and readability analysis. We also conduct an error analysis to investigate error types produced by SIMCSUM.

The SSR and SIMCSUM provide a basis for developing and evaluating cross-lingual scientific summarization and science journalism models. These models are based on generalized

solutions, so our models and their modifications can be implemented for other domains.

7.2 Future Directions

In this section, we discuss three possible follow-ups for the work that is presented in this thesis:

1. **Combining extractive summarization with SimCSum**

   One possible future direction for investigating Cross-lingual Science Journalism is combining the extractive summarization model with SIMCSUM in a pipeline fashion. With the SSR model, we observe better retention of long-term dependencies, and the model produces longer summaries without repetitions. It would be interesting to explore how this combination performs for Cross-lingual Science Journalism.

2. **Infusing linguistic features in cross-lingual abstractive summarization**

   Another possible direction is to induce readability features in cross-lingual abstractive summarization. However, a major obstacle to infusing such traditional features with neural networks is that these features are non-differentiable and non-continuous and cannot be adopted as training objectives. One possible solution is adding linguistic features with inputs, and another possibility is using Gumbel Softmax with them to get continuous representations.

3. **ChatGPT for Cross-lingual Science Journalism**

   Recently, a reinforcement-based learning model - CHATGPT has been introduced by OpenAI[1]. Due to its diverse performance and smart solutions, it immediately has gained popularity and attention from the AI-NLP community and the general public. So here a question arises how can CHATGPT reshape Cross-lingual Science Journalism? To answer this, we could not conduct a study in this regard. However, we find a similar work by Wang et al. (2023). They investigate the performance of CHATGPT for cross-lingual summarization against the SOTA models, including MBART. The authors evaluate the performance of these models for English-Chinese and English-German language pairs. Interestingly CHATGPT could not beat MBART for cross-lingual summarization. Based on this study, we infer that the performance of CHATGPT for Cross-lingual Science Journalism needs a thorough investigation and can be considered another future direction.

[1] https://openai.com/blog/chatgpt

List of Figures

List of Tables

List of Algorithms

Bibliography

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 1–9.

Jonathan Anderson. 1981. Analysing the readability of english and non-english texts in the classroom with lix.

Diego Antognini and Boi Faltings. 2020. GameWikiSum: a novel large multi-document summarization dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6645–6650, Marseille, France. European Language Resources Association.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German Dataset for Joint Summarization and Simplification. *ArXiv preprint*, abs/2201.07198.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yu Bai, Yang Gao, and He-Yan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. 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)*, pages 6910–6924.

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Yiming Zhu, Jiaao Zhan, Zewen Chi, and Boxing Chen. 2022. Unifying cross-lingual summarization and machine translation with compression rate. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1087–1097.

Yael Barel-Ben David, Erez S Garty, and Ayelet Baram-Tsabari. 2020. Can Scientists Fill the Science Journalism Void? Online Public Engagement with Science Stories Authored by Scientists. *Plos One*, 15(1):e0222250.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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)*, pages 3615–3620.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Christian Bentz, Dimitrios Alikaniotis, Tanja Samardžić, and Paula Buttery. 2017. Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, 24(2-3):128–162.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2019. Unsupervised opinion summarization as copycat-review generation. *arXiv preprint arXiv:1911.02247*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. 1993. The mathematics of statistical machine translation: Parameter estimation.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020a. Expertise style transfer: A new task towards better communication between experts and laymen. *arXiv preprint arXiv:2005.00701*.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020b. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6220–6231.

Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou. 2016. Tgsum: Build tweet guided multi-document summarization dataset. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Rich Caruana. 1998. *Multitask learning*. Springer.

Silvia Casola and Alberto Lavelli. 2022. Summarization, simplification, and generation: The case of patents. *Expert Systems with Applications*, page 117627.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. Lexical substitution dataset for german.

Alina Maria Ciobanu and Liviu P Dinu. 2014. A quantitative insight into the impact of translation on readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 104–113.

Arman Cohan, 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)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

K Bretonnel Cohen, Lawrence E Hunter, and Peter S Pressman. 2019. P-hacking lexical richness through definitions of "type" and "token". *Studies in health technology and informatics*, 264:1433.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine Scoring. *Journal of Applied Psychology*, 60(2):283.

Miguel Collantes, Maureen Hipe, Juan Lorenzo Sorilla, Laurenz Tolentino, and Briane Samson. 2015. Simpatico: A text simplification system for senate and house bills. In *Proceedings of the 11th National Natural Language Processing Research Symposium*, pages 26–32.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

M Covington and Joe D McFall. 2008. The moving-average type-token ratio. *Linguistics Society of America, Chicago, IL, United States*.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Rumen Dangovski, Li Jing, Preslav Nakov, Mićo Tatalović, and Marin Soljačić. 2019. Rotational unit of memory: A novel representation unit for RNNs with scalable applications. *Transactions of the Association for Computational Linguistics*, 7:121–138.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakova, and Marin Soljacic. 2021. We Can Explain Your Research in Layman's Terms: Towards Automating Science Journalism at Scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12728–12737, Online.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

William H DuBay. 2004. The principles of readability. *Online Submission*.

Melody Dye. 2017. Bridging levels of analysis: Learning, information theory, and the lexicon. *ProQuest LLC*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Liana Ermakova, Patrice Bellot, Pavel Braslavski, Jaap Kamps, Josiane Mothe, Diana Nurbakova, Irina Ovchinnikova, and Eric San-Juan. 2021. Text simplification for scientific information access: Clef 2021 simpletext workshop. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 583–592. Springer.

Liana Ermakova, Patrice Bellot, Jaap Kamps, Diana Nurbakova, Irina Ovchinnikova, Eric SanJuan, Elise Mathurin, Sílvia Araújo, Radia Hannachi, Stéphane Huet, et al. 2022. Automatic Simplification of Scientific Texts: SimpleText Lab at CLEF-2022. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Proceedings, Part II*, pages 364–373, Stavanger, Norway. Springer.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Mehwish Fatima, Tim Kolber, Katja Markert, and Michael Strube. 2023. Simcsum: Joint learning of simplification and cross-lingual summarization for cross-lingual science journalism. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 24–40.

Mehwish Fatima and Michael Strube. 2021. A novel Wikipedia based dataset for monolingual and cross-lingual summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 39–50, Online and in Dominican Republic. Association for Computational Linguistics.

Mehwish Fatima 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)*, pages 1843–1861.

Dominik Frefel. 2020. Summarization corpora of Wikipedia articles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6651–6655, Marseille, France. European Language Resources Association.

Roland Friedrich, Mauro Luzzatto, and Elliott Ash. 2020. Entropy in legal language. In *NLLP 2020 Natural Legal Language Processing Workshop 2020. Proceedings of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, volume 2645, pages 25–30. CEUR-WS.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain:{I} nsights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Teodora Ghivirigâ. 2012. Scientific articles in english in economics journals. a case study for romania. *Procedia-Social and Behavioral Sciences*, 46:4230–4235.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria. Association for Computational Linguistics.

George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011. TAC2011 MultiLing Pilot Overview. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011*, Gaithersburg, Maryland, USA. NIST.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual

summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Gillick, Benoit Favre, Dilek Hakkani-Tür, Bernd Bohnet, Yang Liu, and Shasha Xie. 2009. The icsi/utd summarization system at tac 2009. In *Tac*. Citeseer.

Dimitrios Giomelakis and Andreas Veglis. 2015. Employing search engine optimization techniques in online news. *Studies in media and communication*, 3(1):22–33.

Annette Rios Gonzales, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in german. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161.

Ian J Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. 2013. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*.

Max Grusky, 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)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Benjamin Hättasch, Nadja Geisler, Christian M. Meyer, and Carsten Binnig. 2020. Summarization beyond news: The automatically acquired fandom corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6700–6708, Marseille, France. European Language Resources Association.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Yaling Hsiao, Nicola J Dawson, Nilanjana Banerji, and Kate Nation. 2024. A corpus-based developmental investigation of linguistic complexity in children's writing. *Applied Corpus Linguistics*, 4(1):100084.

Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2007. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 901–904.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254, Online. Association for Computational Linguistics.

Akanksha Joshi, Eduardo Fidalgo, Enrique Alegre, and Laura Fernández-Robles. 2019. Summcoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*, 129:200–215.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158.

Kazuya Kawakami. 2008. *Supervised sequence labelling with recurrent neural networks*. Ph.D. thesis, Technical University of Munich.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186.

Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *arXiv preprint arXiv:2006.01997*.

Minsoo Kim, Dennis Singh Moirangthem, and Minho Lee. 2016a. Towards abstraction from extraction: Multiple timescale gated recurrent unit for summarization. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 70–77, Berlin, Germany. Association for Computational Linguistics.

Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016b. SimpleScience: Lexical simplification of scientific terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1071, Austin, Texas. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a german/simple german parallel corpus for automatic text simplification.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph 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)*, pages 6365–6378, Online. Association for Computational Linguistics.

Faisal Ladhak, 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*, pages 4034–4048, Online. Association for Computational Linguistics.

Christine Largeron, Christophe Moulin, and Mathias Géry. 2011. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM symposium on applied computing*, pages 924–928.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2091–2100.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086.

Kanglong Liu, Rongguang Ye, Liu Zhongzhu, and Rongye Ye. 2022. Entropy-based discrimination between translated chinese and original chinese using data mining techniques. *Plos one*, 17(3):e0265633.

Yang Liu 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Annie Louis and Ani Nenkova. 2013a. A Corpus of Science Journalism for Analyzing Writing Quality. *Dialogue & Discourse*, 4(2):87–117.

Annie Louis and Ani Nenkova. 2013b. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Joshua Maynez, 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*, pages 1906–1919, Online. Association for Computational Linguistics.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Antonio Menta and Ana Garcia-Serrano. 2022. Controllable sentence simplification using transfer learning. *Proceedings of the Working Notes of CLEF*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of biomedical informatics*, 52:457–467.

Eugène Mollet, Alison Wray, Tess Fitzpatrick, Naomi R Wray, and Margaret J Wright. 2010. Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics*, 15(4):429–473.

José Monteiro, Micaela Aguiar, and Sílvia Araújo. 2022. Using a pre-trained simplet5 model for text simplification in a limited corpus. *Proceedings of the Working Notes of CLEF*.

Alejandro Mosquera. 2022. Tackling data drift with adversarial validation: An application for german text complexity estimation. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 39–44.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Courtney Napoles, Matthew R Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX)*, pages 95–100.

Shashi Narayan, 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*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. 2018. Data-driven Summarization of Scientific Articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. 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)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40.

Chris D Paice. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 172–191.

Chris D Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple ppdb: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, and Vijay Mago. 2022. Medical text simplification using reinforcement learning (teslea): Deep learning–based text simplification approach. *JMIR Medical Informatics*, 10(11):e38095.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The Readability of Scientific Texts is Decreasing Over Time. *Elife*, 6:e27725.

Agnieszka Przybyla-Wilkin. 2016. Easy-to-read in english, german and polish. *Ed. NATHALIE MÄLZER. Barrierefreie Kommunikation–Perspektiven aus Theorie und Praxis. Berlin: Frank & Timme*, pages 135–150.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47:919–944.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Pengjie Ren, Furu Wei, Zhumin Chen, Jun Ma, and Ming Zhou. 2016. A redundancy-aware sentence regression framework for extractive summarization. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 33–43.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres*, pages 12–21.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Abigail See, 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)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Svetlana Sheremetyeva. 2014. Automatic text simplification for handling intellectual property (the case of multiple patent claims). In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 41–52.

Sanja Štajner. 2021. Automatic text simplification for social good: progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Josef Steinberger, Karel Jezek, et al. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4(93-100):8.

Morgan Stills. 2016. Language sample length effects on various lexical diversity measures: An analysis of spanish language samples from children.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. *arXiv preprint arXiv:2110.05071*.

Rui Sun, Zhenchao Wang, Yafeng Ren, and Donghong Ji. 2016. Query-biased multi-document abstractive summarization via submodular maximization using event guidance. In *Web-Age Information Management: 17th International Conference, WAIM 2016, Nanchang, China, June 3-5, 2016, Proceedings, Part I 17*, pages 310–322. Springer.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based automatic text simplification for german.

Sho Takase and Naoaki Okazaki. 2020. Multi-task learning for cross-lingual abstractive summarization. *arXiv preprint arXiv:2010.07503*.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. X-SCITLDR: cross-lingual extreme summarization of scholarly documents. *ArXiv preprint*, abs/2205.15051.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1171–1181.

Robin Tillman and Ludvig Hagberg. 2014. Readability algorithms compability on multiple languages.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018. When science journalism meets artificial intelligence : An interactive demonstration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 163–168, Brussels, Belgium. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of german targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 380–393.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for german language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Linting Xue, 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, pages 483–498, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif Radi Aljohani, and Raheel Nawaz. 2020. Htss: A novel hybrid text summarisation and simplification architecture. *Information Processing & Management*, 57(6):102351.

Damián H Zanette. 2014. Statistical patterns in written language. *arXiv preprint arXiv:1412.3336*, page 11.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A

new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual 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)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. 2016. The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1535–1545, Osaka, Japan. The COLING 2016 Organizing Committee.