Judith Wagner
Dr. sc. hum.

Multilingual Natural Language Generation from a Medical Concept Representation

Geboren am 30.07.1963 in Weinheim
Reifeprüfung am 12.05.1982 in Weinheim
Studiengang der Fachrichtung Medizinische Informatik vom WS 1982 bis SS 1988
Vordiplom am 10.01.1985 an der Universität Heidelberg
Diplom am 9.09.1988 an der Universität Heidelberg

Promotionsfach: Medizinische Biometrie und Informatik
Doktorvater: Prof. Dr. rer. biol. hum. R. Haux

A number of compositional Medical Concept Representation systems are being developed in order to enable re-use and sharing of medical knowledge and data. These need to be presentable as surface language phrases for use by end-users and applications. This is achieved by generation of natural language, providing the translation of conceptual representations to one or several 'natural' languages. It is a potential way of reconciling two requirements, both of which are essential for many of today's medical informatics developments, and especially for a computer-based patient record, but which are sometimes conflicting: effective communication and the sharing and reuse of medical information.
The main goals of this work were to describe and represent the knowledge necessary for multilingual generation from a Medical Concept Representation, as well as to develop a method and to design and implement a tool for multilingual generation, which can be adapted specifically to the requirements and the language of the medical domain. Typical medical language has become a telegraphic sublanguage with particular characteristics and conventions, and with a high density of information expressed in short and complex phrases: the medical 'jargon'.
This thesis presents a multilingual, semantic-oriented sublanguage approach to natural language generation. The approach is based on the hypothesis that a conceptual model of the domain and linguistic knowledge have to be clearly distinguished, and to be linked by defined structures. The knowledge base includes the domain model, the linguistic knowledge, and the links between both. The process comprises transformation operations, selection operations, and the linguistic realisation in a particular language, with a focus on how to parametise these operations by the individual language parameters. The link between conceptual and linguistic structures is then established on two levels: on the one hand, annotations link conceptual entities (concepts, relations) to linguistic entities (words, syntactic structures). On the other hand, transformation operations adapt conceptual representations by using definitions to a level where they can be directly translated to linguistic structures.
This approach has been applied to the generation of noun phrases in several European languages for conceptual representations of an existing medical concept representation: the GALEN Common Reference Model. A large-scale experiment has been done on the Urology part of the French ‘Nomenclature Commune des Actes Médicaux’. This part includes 522 surgical procedures, which have been conceptually

represented following an overall surgical procedures scheme, and phrases have been re-generated in different languages. The generation tool has also been used for a Structured Clinical User Interface collecting information about urinary infections: information is collected in one language by forms, and can be summarised in phrases of the same or a different language.

These experiments have demonstrated that the generation tool can be adapted to different source modelling schemes and to different destination languages, or sublanguages, of a domain. The introduction of new languages was facilitated by the generic approach. The generation of results in multiple languages for one domain has shown to be relatively simple once a domain is covered on the conceptual side. In addition, the effort for covering a new medical subdomain converges.

The transformation operations have been a central means for bridging distinct ways of representing the same things conceptually - distinct modelling styles - and distinct, but different, ways of expressing the same concepts in language - different languages and language styles. The operations are based on the Conceptual Graphs formalism, and they are parameterised by the availability of language-specific annotations, language-independent definitions, and overall generation strategies. They enable not only bridging of terminological gaps and the lack of exact translation equivalents, but also the creation of different generation strategies, which allow the style of the generated language to be adapted to application purposes and users. Together with specific definitions which mirror a conceptual scheme, they allow for tailoring to the specific characteristics of a modelling scheme.

The experience demonstrated that it is important to rely on an existing domain model for generation in order to arrive at an application-relevant domain coverage. It also showed the importance of the representation of conceptual relations, and of distinguishing conceptual and linguistic models. The usefulness of Natural Language Generation tools as a validation tool for complex conceptual modelling in medical concept representation systems came also to the fore. Finally, it has been shown to be possible for different applications to rely on a single common conceptual representation, adapting the generation of surface language to match those applications.