Multilingual generation of medical information

Dana Dannélls
Department of Swedish Language
Göteborg University
SE-405 30 Gothenburg, Sweden
dana.dannells@svenska.gu.se

Louise Deléger
INSERM, UMR_S 872 Eq. 20
Université René Descartes
Paris, F-75006, France
louise.deleger@spim.jussieu.fr

Abstract

Multilingual generation systems aim to produce understandable texts in multiple languages from one knowledge representation. We adapted an existing prototype multilingual generator that presents simulated breast cancer Electronic Health Records (EHRs) in English to French and Swedish. The purpose of this work was to test how much effort it would require to modify this limited-domain, template-based English generator to enable it to generate in French and Swedish. We describe the adaptation to both languages, viewing the grammatical aspects involved and explaining the modifications performed. This work illustrates how the same underlying knowledge representation can be used to generate output texts in multiple languages with only minor linguistic modifications.

1 Introduction

As Electronic Health Records (EHRs) become more widely adopted throughout the European Community and legislation allows people to access their own records, it is important that technology is developed to present EHRs in the languages of the community and in a patient-friendly manner. Our project is part of the European Semantic Mining Consortium and aims towards the generation of patient-friendly reports from EHRs (Åhlfeldt et al., 2006).

The advantage of multilingual Natural Language Generation (NLG) systems is that they can produce texts in different languages from the same underlying knowledge representation. Most of such systems require a grammar of the particular language (Elhadad, 1992) and knowledge about how lexical elements are expressed in a particular domain (Not

and Pianta, 1995). In this project we examined and compared a number of language characteristics of medical texts between three languages in the cancer subdomain and one of our goals was to find a representation which allows generating informative documents in different languages.

We adapted an existing prototype multilingual generator that presents simulated breast cancer EHRs in English (Williams et al., 2007) to French and Swedish. The purpose of this work was to test how much effort it would require to modify the limited-domain, template-based English generator, which at present is used for generating patient summaries.

This paper is structured as follows. Section 2 gives a brief background for multilingual generation systems. In section 3, the language generation system is presented. Sections 4 and 5 give an overview of the differences between the languages and describe the modifications we performed. Section 6 provides examples of the generated French and Swedish output texts and continues with a discussion in Section 7. Finally, Section 8 ends with conclusions and suggestions for future work.

2 Background

NLG is the subfield of artificial intelligence and computational linguistics that focuses on computer systems that can produce understandable texts in English or other human languages from a structured data or a knowledge base which serves as the system input (Reiter and Dale, 2000). An NLG system basic architecture contains three main components viz. content selector, sentence planner and surface realizer. Each component is used to solve diverse tasks such as content selection and text organiza-

tion which includes aggregation[1] and lexicalization. Multilingual Natural Language Generation is a specialization of NLG whose task is to produce versions of a text in several languages from one knowledge representation, typically from a non-linguistic representation of information.

In general, it can be said that the tasks of a generation system require language-independent and domain specific components, but also language-specific components. The former is used by the content selector and the document planner whereas the later is used by the surface realizer. One of the most important NLG resources is the grammar that a surface realizer uses to produce linearized text from syntactic sentence plan. Multilingual generation systems in particular require a careful separation of language-specific processes and resources from language independent ones, and thus it is important to explore similarities and differences between languages across different domains in order to efficiently convert a surface realizer from one language into another (Callaway et al., 1999).

Experiences with multilingual generation systems (Svenberg, 1994; Rösner and Stede, 1994; Paris et al., 1995; Lee et al., 2001; Novello and Callaway, 2003b) have shown languages share commonalities in their grammatical descriptions which can be used for maintaining multiple language generation by maximizing resource sharing. For example, Okumura et al., (1991) show that many English verb patterns are available in French and it is possible to use similar syntactic rules for both English and French. They point out that more work is needed on the grammar and the morphology.

3   The language generation system

Figure 1 shows the architecture of the English prototype we utilized, see Williams et al. (2007) for more details. The EHR repository is produced by the CLEF simulator (Rogers et al., 2006) which simulates records of breast cancer patients. These are stored in a relational database which serves as the input knowledge resource for the generator. The generator consists of a content selection module, a document planning module and a surface realizer.

Content selection searches the repository for a patient's records. Document planning builds a dis-

---

[1]Aggregation is a process which concerns composition of several logical assertions that share information into a single natural language utterance with coordinated or omitted parts (York, 2002).
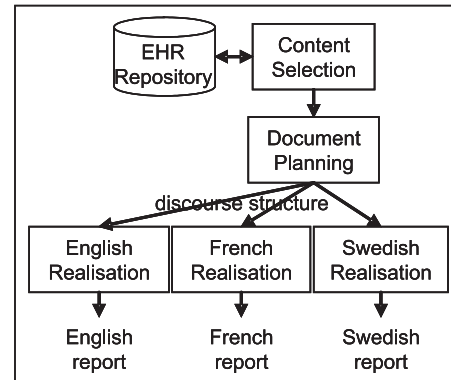
Figure 1: System architecture.

Figure 2: Summary output.

course structure (a rhetorical graph) from these and extends it by adding *Explanation* relations pointing to glossary entries which explain medical concepts. Content selection and document planning are common for all languages, but there is a different realisation component for each language where rules fill sentence templates from the discourse structure and realise them syntactically.

English output, shown in figure 2, is fairly rudimentary and machine-like at the current stage of development. However, our initial purpose was to demonstrate ease of conversion to other languages and future work will improve the output.

4   Swedish language generation

4.1   Language characteristics

Swedish is a North Germanic language which has a rather large morphological inflection for nouns and adjectives. By contrast to English, Swedish noun has inherent gender: utrum and neuter. The gram-

	English	Swedish
definite articles	the	den, det, de +en, +et
indefinite articles	a, an	en, ett
possessive pronouns	your	din, ditt, dina
gender	n.a	utrum, neuter

Table 1: Some differences of grammar between Swedish and English.

	English	French
definite articles	the	le, la, l', les
indefinite articles	a, an	un, une, des
possessive pronouns	your	votre, vos
gender	n.a	feminine, masculine

Table 2: Some grammar differences between French and English.

mar makes use of the genders neuter and utrum, determiners and adjectives must agree with number, definiteness and the head noun in order for the analysis grammar to work. An example of agreement between a determiner, adjective and the head noun is: **det** *vänstr***a** *bröst***et** (your left breast).

In Swedish, the choice of the article and the pronoun depends on the gender of the noun, shown in Table 1. Obvious divergence seen is how definiteness is realized in Swedish, i.e. (1) by adding a determiner before the noun, depending on whether the noun is plural or singular, or/and (2) by adding a suffix to the noun. For example, the phrase *the breast* can either be translated as: *det bröstet*, or *bröstet*. Correspondingly, *ditt bröst* (singular) and *dina bröst* (plural) are translations for *your breast*.

Another common characteristic of Swedish is the use of compounds which means words are put together to build one word, for example the Swedish translation of the phrase *doctor examination* is *läkarundersökning*.

4.2 Adaptation method

We started by translating the medical terms and glossaries manually, using the Swedish MeSH [2], a controlled vocabulary thesaurus of the NLM (U.S. National Library of Medicine). For each of the translated terms we assigned a tag with additional information about its gender, for example *röntgen* (Xray) was presented as a tuple in the lexicon, i.e., *(röntgen, utr)*.

The next step was to adapt the templates of the English generator, a task which required several linguistic modifications including: (1) syntactic arrangement of the words e.g. *hade du* (you had), i.e., use of "verb+pronoun" instead of "pronoun+verb".; (2) choice of lexical categories (part-of-speech),

e.g., the translation of the "adjective" *another* was the "adverb" *ytterligare*; (3) semantic interpretation of verbs and prepositions, e.g. the preposition *in* have a wide variety of meanings in Swedish and can correspond to three different prepositions, i.e., *av*, *på* and *i*.

For the purpose of this work we implemented a small grammar which included simple grammatical rules to correlate the agreement between nouns and adjectives. These rules were written using neuter or utrum gender only. An example of such grammar rule is: "IF *article-den-det* ADD *adj + adj-suffix* AND *noun + noun-suffix*".

5 French language generation

5.1 Language characteristics

French is a Roman language which presents several specific grammatical characteristics as compared to English. We will focus here on morphological inflections such as gender and number of nouns, adjectives and determiners. In French, adjectives and determiners agree in gender and number with nouns. For instance, the word *radiographie* (Xray) is a feminine singular noun and must be used with corresponding determiners like *la* (the) or *une* (a). Table 2 gives correspondences between French and English articles, and we can see that for a given English determiner several French ones are possible according to the gender and number of the noun.

As for the syntactic organization of sentences, given the fact that we work in a restricted domain such as medical language for patients, the structure is quite similar in French and English, with only minor differences such as the position of adjectives (before the noun in English, after the noun in French).

5.2 Adaptation method

The adaptation of the system was done semi-automatically. The adaptation was performed in several steps:

(1) medical terms translation;
(2) morpho-syntactic tagging of medical terms;
(3) templates translation and adaptation;
(4) acquisition of definitions of medical terms.

Medical terms were automatically translated into French by matching them with two vocabulary sources. The first one is an online medical French dictionary with English translations[3]. The second one is the UMLS[4], a resource gathering medical terminologies in several languages and linking equivalent terms through concept identifiers (CUIs). We retrieved French terms with the same CUIs as the English ones. The translations were reviewed manually so that no mistakes would be included.

In order to manage morphological issues involving agreement of gender and number, we morpho-syntactically tagged the terms using the part-of-speech tagger TreeTagger[5] and the French lemmatizer Flemm[6] (which gives more detailed information such as gender).

The templates of the English generator were manually adapted to French. The syntactic structure of the English sentences is quite similar to French and the adaptation mainly consisted in managing morphological issues such as having determiners and adjectives agree in gender and number with the nouns by using the morphosyntactic tags acquired in the previous step. The following is an example of how an English template was adapted to French:
"you had *indefinite-article* EXAMINATION of your BODY-LOCATION"
"vous avez eu *indefinite-article* EXAMINATION de *possessive-pronoun* BODY-LOCATION"

Finally, we also acquired definitions in French in order to explain the medical terms by using two methods. First, we retrieved definitions from the same dictionary as used for translating the terms. For those terms that could not be found in the dictionary, we acquired definitions from the Web. Two search engines were used for the web queries. The first one was CISMeF[7], a French health gateway that includes online definitions of medical terms. The second one was the Google Define tool. With

---

[3]Dictionnaire médical Masson, `http://www.atmedica.com`

[4]`http://umlsinfo.nlm.nih.gov/`

[5]`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[6]`http://www.univ-nancy2.fr/pers/namer/Telecharger\_Flemm.htm`

[7]http://www.cismef.org/

Figure 3: Swedish output.

Figure 4: French output.

CISMeF, there is good reliability since it is specialized in medicine, but with Google, where the whole Web is queried, definitions need to be filtered. We filtered them manually but this will be done automatically in future work.

6 Results

The results shown in Figures 3 and 4 illustrate the Swedish and French versions of the same output text given in Figure 2.

In general, the generated information is comprehensible for both French and Swedish. Grammatically, the output in French is good, the sentences being all grammatically correct. The Swedish output is also good, seeing that grammatical mistakes were corrected by our rules. An example of a phrase

that was corrected by our rules is: *en transfusion* (a transfusion), the gender of the word *transfusion* is utrum and the chosen article is therefore *en*.

For French, the only remaining grammatical issue is the choice of the article used in the definition of cancer ("*Un cancer est le (la) développement...*"). This could be resolved in future work by acquiring morphosyntactic information about the definitions. While for Swedish the remaining problems to overcome relates to the syntactic and morphological aspects of the language. Our modification does not cover morphological inflection and the grammatical rules are not complete.

At present, 58% of the terms have been translated into French and Swedish of which 51% have been defined, and thus we cannot provide full outputs for all of the information contained in the system yet.

7 Discussion

Difficulties encountered by previous authors and in this work relates to the syntactic variations of the target language and the morphological complexity. We observed that the errors which occur in Swedish are due to grammatical incorrectness, irregularities in syntax and large morphological inflection. Solutions to these problems require more effort in constructing a grammar and applying necessary morphological modifications. Morphological changes in particular require extensive work to cover irregular nouns and adjectives.

No major problems were encountered for French, except for the fact that morphological inflections had to be dealt with, which was done by tagging the words. This can be explained by the fact that we only generated a limited amount of texts and we believe that in a context of a broader area with more numerous and complex sentences to be generated, additional problems linked to morphology and syntax would have arisen and been more difficult to solve.

Common problems related to both languages are: (1) translations difficulties, such as choosing the correct interpretation of a word in the context, and (2) appropriateness and readability of the sentences, that is related to sentence structure and style.

With respect to generating patient-friendly documents by means of the medical terms involved and the content organization, we followed the suggestions made by Åhlfeldt et al. (2006) and consequently chose content words which are common and understandable. Parallel production of one document resulted in coherent and readable patient-friendly texts in different languages.

However, template-based approaches such as presented in this paper are rather simplistic and have been discussed by the NLG community (Bateman, 1997). Solutions that uses more elaborate multilingual generation systems (such as presented by Bouayad-Agha et al. (2002)) are preferred because they are more powerful. Yet, our aim was not to deal with such state-of-the-art systems, but merely to test the adaptation to other languages of a prototype system able to generate patient-friendly documents.

8 Conclusions and future work

In this paper we demonstrated the feasibility of adapting an NLG system to French and Swedish. Considering the short time needed (approximately one week) to perform necessary modifications for the purpose of generating multilingual context , we conclude that our approaches gave successful results. The process of generating patient summaries in three different languages gave insights into the differences between the languages from which we can proceed with the necessary modifications required to accommodate well defined and employed resources, for example reuse a grammar resource such as described in Novello and Callaway (2003a).

Other improvements relate to the output text form. As we already mentioned, the output is rudimentary and machine-like. The generated output can be improved by adding a sentence planning module that follows on from the document planning module, which determines how to pack the discourse structure into sentence specifications and also chooses lexical items, this process is described in more detail in Williams (2004). The system we used is rather simple and does not use any kind of processes (e.g. aggregation) to make a discourse more fluent and easy to read. We wonder though whether in order to enable the generation of more fluent, grammatical texts, we would have to move away from templates and towards more powerful generation methods.

References

H. Åhlfeldt, L. Borin, P. Daumke, N. Grabar, C. Hallett, D. Hardcastle, D. Kokkinakis, C. Mancini, K. Mark, M. Merkel, C. Pietsch, R. Power, D. Scott, A. Silvervarg, M. Toporowska Gronostaj, S. Williams, and A. Willis. 2006. Literature review on patient-friendly documentation systems. Technical report, Network of Excellence Semantic Mining. Work Package 27 Deliverable 1.

John A. Bateman. 1997. Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(6):15–55.

N. Bouayad-Agha, R. Power, D. Scott, and A. Belz, editors. 2002. *PILLS: Multilingual generation of medical information documents with overlapping content.* Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC). pp. 2111–2114.

Charles B. Callaway, Brent H. Daniel, and James C. Lester. 1999. Multilingual natural language generation for 3D learning environments. In *Proceedings of the 1999 Argentine Symposium on Artificial Intelligence*, pages 177–190, Buenos Aires, Argentina.

Michael Elhadad. 1992. *Using argumentation to control lexical choice: a functional unification-based approach.* Ph.D. thesis, Graduate School of Arts and Sciences Columbia Universitya.

Young-Suk Lee, Wu Sok Yi, Stephanie Seneff, and Clifford J. Weinstein. 2001. Interlingua-based broad-coverage korean-to-english translation in cclinc. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–6, Morristown, NJ, USA. Association for Computational Linguistics.

Elena Not and Emanuele Pianta. 1995. Issues of multilinguality in the automatic generation of administrative instructional texts. In M. Gori and G. Soda, editors, *Topics in Artifical Intelligence, Proceedings of the AI*IA '95 Congress*. Springer.

Alessandra Novello and Charles Callaway. 2003a. Porting to an Italian surface realizer: A case study. In *Proceedings of the 9th European Workshop on NLG*, pages 71–78, Budapest, Hungary, April.

Alessandra Novello and Charles B. Callaway. 2003b. Multilingual generation for museum applications. In *Proceedings of the Italian Association for Artificial Intelligence (AI*IA)*, Pisa, Italy, September.

A. Okumura, K. Muraki, and S. Akamine. 1991. Multi-lingual sentence generation from the pivot interlingua. In *In Proceedings of MT SUMMIT III*.

Cécile Paris, Keith Vander Linden, Markus Fischer, Anthony Hartley, Lyn Pemberton, Richard Power, and Donia Scott. 1995. A Support Tool for Writing Multilingual Instructions. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 1995*, pages 1398 – 1404, Montréal, Canada.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems.* MIT Press and The McGraw-Hill Companies,Inc. Singapore.

J. Rogers, C. Puleston, and A. Rector. 2006. The clef chronicle: Transforming patient records into an e-science resource. In *Biohealth Informatics Group*.

D. Rösner and M. Stede. 1994. Generating multilingual documents from a knowledge base: The techdoc project. In *eprint arXiv:cmp-lg/9407018*, pages 7018–+, July. Provided by the Smithsonian/NASA Astrophysics Data System.

Stefan Svenberg. 1994. Representing conceptual and lingusitic knowledge for multilingual generation in a technical domain. In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 245–248, Kennebunkport.

Sandra Williams, Paul Piwek, and Richard Power. 2007. Generating monologue and dialogue to present personalised medical information to patients. In *Proceedings of the 11th European Workshop on Natural Language Generation*.

Sandra Williams. 2004. *Natural Language Generation (NLG) of Discourse Relations for Different Reading Levels.* Ph.D. thesis, University of Aberdeen.

New York, editor. 2002. *Aggregation with Strong Regularities and Alternatives.* In International Natural Language Generartion Conference.