Automatic generation and simplification of written documents

Dana Dannélls

Göteborg University, Department of Swedish language,
SE-405 30 Gothenburg, Sweden

Abstract. This paper is intended to introduce a doctoral thesis that aims to suggest ways in which existing language technology tools and methods can be utilized in order to present and explain complex information to people who lack either the background or the time required to comprehend the original text form. This paper incorporates a background of previous research within which the present research project may be situated while the particular rationale and motivation behind the proposed research is also outlined.

Keywords text simplification, language generation, electronic health records, semantic web, knowledge representation.

1 Introduction

Published material originally produced for a specific group of readers sometimes cause readability problems for readers outside (and even inside) the profession. Although electronic information is available, it is not valuable since users are not able to understand it. This deficiency gives rise to the need to find new ways of presenting different sorts of textual information in a simple and understandable language.

Automatic simplification of technical and other kinds of documents has been studied to some extent, yet almost exclusively for English. In this study we plan to investigate whether the proposed approaches are applicable to Swedish. Many authors who have dealt with simplification tasks have tended to focus on target readers with specific difficulties such as dyslexia and aphasia (Canning and Taito, 1999; Carroll et al., 1998; Siobhan and Canning, 1999; Inui et al., 2003). Potential readers targeted in the present study are professionals, academics and students with different reading abilities and needs. Our aim is to design a system with the ability to process and extract important information from different documents and to use the extracted information to populate a structured knowledge representation (KR) from which different levels of output texts will be produced. Figure 1 illustrates the system architecture we propose.

This paper represents an ongoing project, whereby the present paper describes the first year of four years of doctoral research. In the course of our

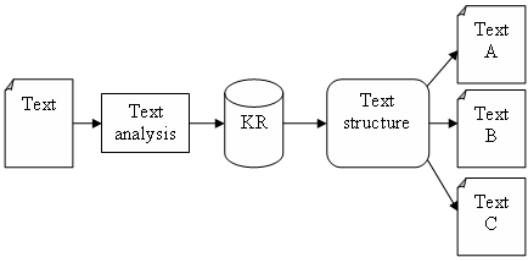

Fig. 1. A simplification system architecture

research we will explore language technology tools and methods that reduce linguistic complexity in syntax, vocabulary and rhetorical text structure in order to adapt the presentation of text content to a specific readership. Further to this, we will suggest and describe new approaches from the perspective of what is involved in the construction of a complete simplification system to fulfil the goal of producing different levels of output documents. We are currently looking at texts from the medical domain, however we hope to propose solutions that are applicable to other domains.

In the following sections we outline the project aims and objectives and the rationale behind the selection of the research problem. We then present a few approaches that have been applied to solve different simplification tasks by using generation techniques. We describe the methods we use to achieve the objectives of this research and finish with suggestions for future work.

Research aim and objective

This work aims to identify optimal ways in which natural language processing techniques can be brought to bear upon the problem of adapting the presentation of text content for a specific readership. The main objective of this research project is to explore different approaches for generating documents for lay readers from a structured representation. The research questions which I am planning to pursue are as follows:

1. What types of domain models and associated reasoning are required for producing readable documents from one knowledge representation?
2. How should knowledge be presented in order for it to be informative and understandable for a specific readership?
3. What constitutes 'readable' text?

In this first year of research I will concentrate on structural variations of written information and the ways in which documents should be structured. We have adapted an existing prototype multilingual generator that presents simulated breast cancer Electronic Health Records (EHRs) in English to Swedish and are currently studying the system's flexibility for generating different levels of output documents which will then be evaluated using real readers.

2 Readability problems

Readability problems are often related to content organization, vocabulary variety, lexical ambiguities, syntactic complexity, idioms etc. (Glöckner et al., 2006). McLaughlin (1969) defines readability as "the degree to which a given class of people find certain reading matter compelling and comprehensible". His definition stresses the interaction between the context and the class of readers with certain reading skills and prior knowledge. This is the same class of readers that we are concerned with in this thesis and deal with readability problems which relate to linguistic complexity in syntax, vocabulary and rhetorical text structure.

Recent experiments have shown that a large amount of readers encounter readability problems due to a lack of language knowledge in a specific field (Clinton-Davies and Fassil, 1992; Free et al., 1998; Hogan, 1998; Jones and Gill, 1998; Fassil, 2000; Somers and Lovel, 2003). Åhlfeldt et al. (2006) compared the language usage in expert and non-expert Swedish texts in the domain of cardiovascular disorders and showed that readability problems also exist among Swedish readers.

Readability and accessibility in general are important issues today especially because of the increase in technical publications using complex language that appear on the Internet. The need to simplify different kinds of textual information gains impetus from the belief that well informed readers will be able to take part in society and become more active in their daily actions and healthcare. It is also vital to ensure that each member of society has an equal awareness of human rights and that potential misunderstandings caused by readability difficulties are lessened.

Natural Language Processing (NLP) applications have the potential to facilitate reader access to a wide range of documents and enhance readability by providing a method whereby informative documents are automatically generated. Some of the techniques that have been applied to identify solutions to different types of reading difficulties are used within the Natural Language Generation (NLG) field. NLG techniques focus on generating written texts in natural languages from some underlying non-linguistic representation of information, generally from a database or a knowledge source. Numerous NLG systems have been developed to deal with readability problems. These systems utilize different simplification methods to produce several types of texts from the same input. In the next section we give a short overview of text simplification methods and motivate why NLG technology is appealing to use in order to solve various simplification tasks.

3 Text simplification methods

Automatic simplification involves selecting, organizing and rewriting information from a linguistic source while preserving the meaning of the original input. The advantage of using text simplification methods is to make a text easier to comprehend for a human user and more easily processed by a computer program (for

more details about the needs for using simplification methods see Chandrasekar et al., 1996). Text simplification methods focus on reducing syntactical and lexical complexity and utilize simplification techniques to help people with language disabilities like aphasia and other language difficulties that may be encountered by non-native speakers.

Simplification methods are normally applied to a linguistic input. The approaches which are taken to reduce syntactic complexity are mainly rule-based and use information from a tagger, a chunker or a parser to coordinate, correlate and subordinate clauses. Furthermore, these methods operate on sentence and paragraph levels.

Text simplification methods aim to rewrite sentences in order to reduce their syntactic or lexical complexity. However, when dealing with different levels of reading skills, this is not always necessary. Consider the following examples[1]:

Example 1. The axillary lymphnodes are rounded masses of tissue under the arm. The axillary lymphnodes contain white blood cells.

Example 2. The axillary lymphnodes are rounded masses of tissue under the arm which contain white blood cells.

Certain readers may find simple and short sentences tedious and may tend to prefer a context which contains more complex sentences, e.g. Example 2. Others may take offense from the sense that their knowledge abilities have been underestimated if the context presented to them only contains simple sentences, e.g. Example 1. The target readers that we are concerned with in the present study have rather good reading skills and should therefore have the opportunity to choose the contextual degree most suited to their skills. Hence, content organization, lexical variations and collocation are the aspects of simplifications we are interested in.

4 Natural Language Generation

NLG is the sub-field 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). The NLG system's basic architecture contains three main components viz. content selector, sentence planner and surface realizer. Each component is used to solve diverse tasks such as text organization, aggregation and lexicalization. This architecture makes generation systems easy to build by allowing modularisation flexibility and easy integration with other modalities such as graphics. Additional capabilities of generation systems include their flexibility regarding user modelling, content variance and their ability to produce versions of a text in several languages from one

[1] Example 1 is taken from a text generated by the NLG system we are working with, example 2 is our own modification of example 1.

knowledge representation, this specialisation is known as Multilingual natural Language Generation (MLG).

Language generation systems are often built with a certain application context in mind, they are structured with respect to the kind of input they accept and the output they should produce. This gives rise to certain problems such as how to encode the information so it will allow generating coherent output documents with different contents and forms? And also, how to select the content and decide on the content structure? – Such problems have received a lot of attention within the NLG community and there have been new approaches proposed to address problems which are related to the input and output specification of a language generation system.

Hirst et al. (1997) and Bouayad-Agha et al. (2002) present a master model from which documents of various types can be generated automatically. The master model is a specification of all the information that is related to a particular topic. The elements of such a master model can be pieces of a language independent structure in KR formalism, which would in turn be selected for the content, as appropriate for a particular reader. The selected elements will then pass through a complete language-generation system that would decide how to organize and express the content, based on given information about the form best suited for the target reader.

5 Knowledge Representation

Ontology is "a branch of metaphysics which aims to discover what entities exist and attempts to sort these entities into categories".[2] Examples of such categories are individuals, events, properties, relations, facts etc. For an ontology to be used within an application it must be represented in an appropriate KR language, that is, a representation suitable for computer processing.

According to Sowa (1999) there are several principles behind building a knowledge representation for natural language systems: (1) It should be able to provide answers to questions within a domain; (2) It should be flexible and allow pragmatically efficient computation; (3) The specification language should be designed for the Semantic Web; (4) There should be support tools that are able to descibe it. Furthermore, one of the key requirements towards the semantic web technology is that it should be applied to existing software projects and require little human involvement in its creation and maintenance (Angelova and Bontcheva, 1996).

A common framework for expressing information is the Resource Description Framework (RDF)[3], and currently OWL (Web Ontology Language)[4] is becoming standard way to express ontologies in RDF syntax. Bontcheva and Wilks (2004)

[2] The encyclopedia of language and linguistic, volume 5.
[3] RDF vocabulary description language 1.0, at http://www.w3.org/TR/rdf-schema
[4] OWL Web Ontology Language Reference, In World Wide Web Consortium, 2004, at http://www.w3.org/TR/owlref/

show how these representations can be used as input to generate textual summaries and point out the content of the ontology itself as a major factor for the quality of the output. Typically, language generation systems have access to a full semantic representation of the domain, yet unfortunately the quality of the output (which depends on the encoded information in the input) is not satisfactory. Hence, part of this project is to refine formal knowledge needs and semantic representations from which informative documents can be generated.

6 Methodologies and experiments

We have already begun with a series of experiments to test an NLG system's ability to produce different levels of output texts and to investigate the effects of the output choices on different readers. Our present work focuses on techniques for rewriting and presenting different levels of documents with the same content.

At present we are working with the CLEF (Rogers et al., 2006) language generation system prototype (Hallett and Scott, 2005). The prototype is a template-based English generator which presents a simulation of breast cancer EHRs (Williams et al., 2007). The overall approach in CLEF is based on ontology anchored knowledge bases which cover domain knowledge about the concepts represented. The simulator simulates the history of a patient's illness and links occurrences of events[5] which are stored in a relational database. There exists six types of events (such as interventions, investigations, problems) each of which has a variable number of attributes, and fourteen relations between events, for example *Investigation Has-Indication Problem*, where *Investigation* is *lumpectomy* and *problem* is *cancer*. This data-encoded chronicle of the patient's medical history is the input for the generator.

The language generation prototype was adapted from English to Swedish and French (Dannélls and Deléger, 2007). The adaptation was basically a translation process which required grammatical and syntactical modifications. The generation gave fairly good results, however a common problem in all three languages was that the output documents were rudimentary and contained redundant information. The Swedish version of the prototype is currently under development and is used as a test bed for our experiments in which we examine the system's flexibility in generating different levels of output documents. These will later be evaluated using real readers. The system architecture is shown in Figure 2.

Content selection selects events according to their semantic relations using clustering methods. These are represented as *messages* for the document planner, an example of such message is shown in Figure 3. Document planning builds a discourse 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 different output languages, but there is a different realization component for each language where rules fill sentence templates from the discourse structure and realize them syntactically.

[5] An 'event' reflects on the type and date of discovery of for instance a problem.

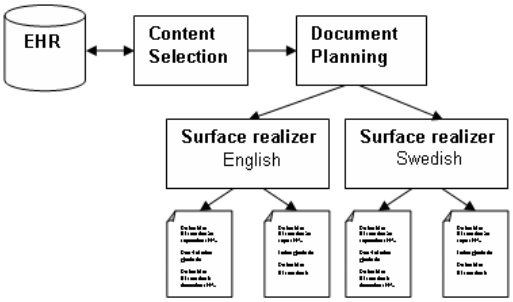

Fig. 2. The NLG system architecture

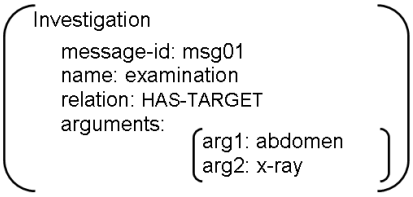

Fig. 3. Example of *Investigation* message

Part of the document planning process is the task of aggregation. Aggregation is a process which concerns the composition of several logical assertions that share information into a single natural language utterance with coordinated or omitted parts (Horacek, 2002). It removes redundancies during generation without losing any information. Aggregation, which has been called ellipsis or coordination in linguistics, makes a text more fluent and easy to read. Aggregation usually takes place in the Sentence Planning module that normally follows on from the Document Planning module. This module is used to determine how to pack the discourse structure into sentence specifications and usually lexical items that should be used to express the content are also chosen. In the system we are using these processes are not fully implemented and the generated texts show there is room for improvement.

Conceptual rules use the semantics of individual messages to form a discourse structure (Hallett and Scott, 2005), for example two *Investigation* messages with two different target fields such as the *breast* and *abdomen* are collapsed into one message with two target fields $\{breast; abdomen\}$.

We enhanced the aggregation process with two simple syntactic rules that use semantic information to combine two messages into one single message, these rules are shown in Figure 4.

The rules conjoin two messages that succeed each other in the history log as follows:

```
(1). ([input, (messageM(relationR1(Y,X), arg(X), arg(Y), date(Y,Z));
              messageM(relationR1(Y,X), arg(X), arg(Y), date(Y,W)))],
     [output, (messageM(relationR1(Y,X), arg(X), arg(Y), date(Y,Z), date(Y,W)))])

(2). ([input, (messageM(relationR1(Y,X), date(Y,W));
               messageM(relationR1(P,Q), date(P,W)))],
     [output, (messageM(relationR1(Y,X), relationR1(P,Q), date(Y,W), date(P,W)))])
```

Fig. 4. Syntactic aggregation rules

1. Two messages with identical relations, identical target fields and different date fields are collapsed into one message with two date fields.
2. Two messages with different relations, different target fields and identical date fields are collapsed into one message with two messages with one date field.

The resulting message is then passed to the surface realizer which builds a sentence in the corresponding language by inserting a connective, e.g. *and* between dates or events. The aggregation process led to an additional set of rules that were necessary in order to correct the syntax, for example, we defined a rule to replace the order of the verb and the noun (*hade du* vs. *du hade*) because the noun *du* becomes the sentence subject after aggregation. The following is a fragment of the generation results of the Swedish and the English output texts:

Original text (Swedish):

> Den 18 januari gjorde du en självundersökning och du upptäckte att du hade en knöl i det vänstra bröstet. Den 25 januari gjorde du ytterligare en självundersökning och du upptäckte att du hade en knöl i det vänstra bröstet. Du besökte läkaren den 1 februari 1999. Den 1 februari hade du en radikal mastektomi för att behandla cancer i det vänstra bröstet.

Aggregated text (Swedish):

> Den 18 januari **och** den 25 januari gjorde du en självundersökning och du upptäckte att du hade en knöl i det vänstra bröstet. Du besökte läkaren den 1 februari 1999 **och** du hade en radikal mastektomi för att behandla cancer i det vänstra bröstet.

Original text (English):

> On January 18th you did a self examination and you found that you had a lump in your left breast. On January 25th you did another self examination and you found that you had a lump in your left breast. You had a consultation with your doctor on February 1st 1999. On February 1st you had a radical mastectomy to treat cancer in your left breast.

Aggregated text (English):

> On January 18th **and** on January 25th you did a self examination and you found that you had a lump in your left breast. You had a consultation with your doctor on February 1st 1999 **and** you had a radical mastectomy to treat cancer in your left breast.

As the results show, there is a difference in the syntactic realization of some generated sentences. However there are certain modifications which are necessary to apply before we can investigate the effects of the output choices on different readers. For example, there is no automatic procedure for choosing between different concepts, synonyms words and deciding whether a concept requires an explanation. Aggregation is applied at sentence level and there is no major difference in the content organization.

The evaluation of the aggregated output documents will be in the form of pilot experiments in which users will be asked to read the generated documents and answer questions about their content. We will judge human users' success or failure in performing these tasks by calculating the percentage of correctly answered questions, with a focus on sentence- and text- level analysis. We hope these evaluations will increase our understanding of what is required to improve the output.

7 Future directions

Additional improvements to the current system include enhancing the Document Planning module with a component that allows for the selection of synonym words and varying contents. Kokkinakis et al. (2007) show there is a difference in the usage of terminology between experts and lay-readers. In the near future we intend to cover necessary, though potentially confusing, medical concepts in order to facilitate reading.

To overcome grammatical difficulties we intend to add a grammar module that will follow on directly from the knowledge source. Since the EHR repository is based on the idea of representing clinical information as a form of semantic net and has the same characteristics as our desired knowledge base, it serves as an excellent domain knowledge source to experiment with. There already exists work that has been carried out by a research group at our University on handling functional framework for writing and working with grammar (Ranta, 2004). The Grammatical Framework (GF) is a type-theoretical logical framework and is able to describe general ontologies. This formalism will allow us to enhance the system with a grammar realisation module that maps out semantic information and linguistic presentation.

References

Åhlfeldt, H., Borin, L., Daumke, P., Grabar, N., Hallett, C., Hardcastle, D., Kokkinakis, D., Mancini, C., Mark, K., Merkel, M., Pietsch, C., R. Power, R., Scott, D.,

Silvervarg, A., Toporowska Gronostaj, M., Williams, S., Willis, A.: Literature Review on Patient-Friendly Documentation Systems. Network of Excellence Semantic Mining, Work Package 27 Deliverable 1 (2006).

Angelova, G. and Bontcheva, K.: DB-MAT: Knowledge Acquisition, Processing and NL Generation. In P.W. Eklund, G. Ellis, and G. Mann, editors, Conceptual Structures: Knowledge Representation as Interlingua, number 1115 in Lecture Notes in AI. Springer-Verlag, Berlin, (1996).

Bontcheva, K. and Wilks, Y.: Automatic Report Generation from Ontologies: the MIAKT approach. Nineth International Conference on Applications of Natural Language to Information Systems (NLDB'2004). Manchester, UK. (2004).

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

Clinton-Davies, L. and Fassil,Y.: Health and social problems of refugees, Social Science in Medicine **35** (1992) 507–13.

Canning, Y. and Taito, J.: Syntactic simplification of newspaper text for aphasic readers. In Proc. of the 22nd Annual International ACM SIGIR Conference (SIGIR), (1999).

Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, (1998).

Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and methods for text simplification. In Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING), Copenhagen, Denmark (1996).

Dannélls, D. and Deléger, L.: Multilingual generation of medical information. To appear in the 9th Bar-Ilan Symposium on the Foundations of Artificial Intelligence (BISFAI) Bar-Ilan University, Israel (2007).

Fassil, Y.: Looking after the health of refugees, British Medical Journal **321** (7252) (2000) 59.

Free, C., Bhui, K., Irwin, J., Martin, S., Carter, J., Hare-Cockburn K. : Breaking down language barriers. Some ethnic groups may have problems in getting as far as a consultation, British Medical Journal **317** (7161) (1998) 816–7.

Glöckner, I., Hartrumpf, S., Helbig H., Leveling, J., Osswald, R.: An Architecture for Rating and Controlling Text Readability. University in Hagen, Germany (2006).

Hallett, C. and Scott, D.: Structural variation in generated health reports. Proceedings of the 3rd International Workshop on Paraphrasing, Jeju Island, Republic of Korea (2005).

Hirst, G., DiMarco, C., Hovy, E., Parsons, K.: Authoring and Generating Health-Education Documents That Are Tailored to the Needs of the Individual Patient. User modelling: Proceedings of the Sixth International Conference. UM97. Vienna, New York: Springer Wien New York (1997).

Hogan, H.: Increasing health inequalities for refugees, British Medical Journal **317** (7170) (1998) 1444–6.

Helmut, H.: Aggregation with Strong Regularities and Alternatives. International Natural Language Generartion Conference (2002).

Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T.: Text simplification for reading assistance: A project note, In Proceedings of the Second International Workshop on Paraphrasing (2003) 9–16.

Jones, D. and Gill, P.: Breaking down language barriers, British Medical Journal **316(7127)** (1998) 1476–80.

Kokkinakis, D., Toporowska Gronostaj, M., Hallett, C., Hardcastle, D.: Lexical Parameters, Based on Corpus Analysis of English and Swedish Cancer Data, of Relevance for NLG. Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA), Tartu, Esthonia (to appear) (2007).

McLaughin, G. H.: SMOG grading - a new readability formula. Journal of reading **22** (1969) 639–646.

Ranta, A.: Grammatical Framework, a type-theoretical grammar formalism. Journal of Functional Programming **14(2)** (2004) 145–189.

Reiter, E. and Dale, R.: Building natural language generation systems. Cambridge University Press, New York, NY, USA (2000).

Rogers, J., Puleston, C., Rector, A.: The CLEF Chronicle: Patient Histories derived from Electronic Health Records. Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), IEEE (2006).

Siobhan, D. and Yvonne, C.: Automatic text simplification for readers with aphasia. University of Sunderland, School of Computing, Engineering and Technology, (1999).

Somers Harold, L. and Lovel, H.: Can AAC technology facilitate communication for patients with limited English?. Association for Computational Linguistics EACL (2003).

Sowa, F. J.: Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA (1999).

Williams, S., Piwek, P., Power, R.: Generating monologue and dialogue to present personalised. medical information to patients, Proceedings of the 11th European Workshop on Natural Language Generation (2007).