Natural Language Generation and Summarization at RALI

Guy Lapalme
RALI - DIRO
Université de Montréal
C.P. 6128, Succ. Centre-Ville
Montréal, Québec, Canada, H3C 3J7
lapalme@iro.umontreal.ca

Processing language in written or spoken form, in a mother tongue or in another language is a very complex and important problem. Hence the idea of building automatic or semi-automatic tools to support people during their attempt to understand what they read or to translate a given message into an adequate linguistic form. Since the eighties, I have worked with my students on many NLP projects, this talk focusses on some of them, past and present, dealing with generation and summarization.

We have always thrived to produce *working* systems that deal with *real* texts or use data to produce texts that can be easily understood by humans. This fundamental motivation imposes some challenging constraints but also produces interesting payoffs. Given the fact that our lab is in French speaking university in a mostly English speaking country, we have often worked in either of these languages and often in both.

1 Generation

PRÉTEXTE (Gagnon and Lapalme, 1996) was a system for generating French texts conveying temporal information. Temporal information and localization expressed by temporal adverbial and verbal phrases was represented with DRT. Systemic Grammar Theory was used to translate the DRT representation into a syntactic form to produce the final text.

SPIN (Kosseim and Lapalme, 2000) deals with a fundamental problem in natural language generation: how to organize the content of a text in a coherent and natural way. From a corpus analysis of French instructional texts, we determined 9 senses typically communicated in these texts and 7 rhetorical relations used to present them. We then developed presentation heuristics to determine how the senses should be organized rhetorically to create a coherent and natural text.

POSTGRAPHE (Fasciano and Lapalme, 2000) generated a report integrating graphics and text from a set of writer's intentions. The system was given data in tabular form and a declaration of the types of values in the columns of the table. Also indicated were intentions to be conveyed in the graphics (e.g., compare two variables or show the evolution of a set of variables) and the system generated a report in LaTeX. PostGraphe also generated the accompanying text to help the reader focus on the important points of the graphics.

SIMPLENLG-EN-FR (Vaudry and Lapalme, 2013) is a bilingual adaptation of the English realizer SimpleNLG. Its French grammatical coverage is equivalent to the English one and covers the essential notions that are taught to learners of French as a second language as defined by *Le français fondamental (1er Degré)*. The French lexicon contains a commonly used French vocabulary, including function words. JSREAL is a work in progress describing a French text realizer in Javascript that can be easily embedded in a web browser. Its main originality is the fact that it produces DOM elements and not text strings so that they can easily produce parts of web pages from JSON inputs sent by the server for example.

In a project of interactive generation, we develop a cognitively inspired methodology to assist people during the production process, as the route between input and output can be full of hurdles and quite long. For each step, we want to develop web based applications that address a specific problem and help induce some pattern reaction in the production of language. For the moment we have produced two prototypes: DRILLTUTOR (Zock and Lapalme, 2010) which is goal-oriented multilingual phrasebook and WEBREG (Zock et al., 2012) to practice the generation of appropriate referring expressions.

Proceedings of the 14th European Workshop on Natural Language Generation, pages 92–93,
Sofia, Bulgaria, August 8-9 2013.

2 Summarization

Summarization is *in principle* strongly related to NLG because it implies reading and understanding one or many documents in order to produce a short text describing the main ideas of the original. Summarization approaches are often classified as either abstractive or extractive, the former being the selection of the most important sentences from the original documents.

In much the same way as NLG has *suffered* from the fact that it is often possible to trick the readers with canned text or formatted templates, abstractive summarization had to compete with acceptable results produced by scorers of sentences, the ones with the best scores being then concatenated to produce a summary. In our group, we tried to stay away from such approaches that in our view did not give any new insights even though it did not always allow us to *win* the summarization competitions at DUC or TAC.

SUMUM (Saggion and Lapalme, 2002) explored the idea of dynamic summarization by taking a raw technical text as input and produced an indicative-informative summary. The indicative part of the summary identifies the topics of the document, and the informative part elaborates on some of these topics according to the reader's interest. SumUM motivates the topics, describes entities, and defines concepts. This is accomplished through a process of shallow syntactic and semantic analysis, concept identification, and text regeneration.

LETSUM (Farzindar and Lapalme, 2004) developed an approach for the summarization of legal documents by helping a legal expert determine the key ideas of a judgment. It is based on the exploration of the document's architecture and its thematic structures in order to build a table style summary for improving coherency and readability of the text. Although LetSUM extracted full sentences from the original document, it reorganized, merged and displayed different parts in order to better give an idea of the document and focus the reader, a legal expert, to the important parts.

ABSUM (Genest and Lapalme, 2013) introduces a flexible and scalable methodology for abstractive summarization that analyzes the source documents using a knowledge base to identify patterns in the the source documents and generate summary text from them. This knowledge-based approach allows for implicit understanding and transformation of the source documents' content because it is carefully crafted for the summarization task and domain of interest.

3 Conclusion

These examples illustrate some links that we have established between generation and summarization over the last few years and that are promising for the future of these two research areas.

References

Atefeh Farzindar and Guy Lapalme. 2004. Legal texts summarization by exploration of the thematic structures and argumentative roles. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 27–34, Barcelona, Spain, July. Association for Computational Linguistics.

M. Fasciano and G. Lapalme. 2000. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge and Information Systems*, 2(3):310–339, Aug.

M. Gagnon and G. Lapalme. 1996. From conceptual time to linguistic time. *Computational Linguistics*, 22(1):91–127, March.

Pierre-Etienne Genest and Guy Lapalme. 2013. Absum: a knowledge-based abstractive summarizer. *Computational Linguistics*, page 30 pages, July. In preparation.

L. Kosseim and G. Lapalme. 2000. Choosing rhetorical structures to plan instructional texts. *Computational Intelligence*, 16(3):408–445, Aug.

Horacio Saggion and Guy Lapalme. 2002. Generating informative and indicative summaries with SumUM. *Computational Linguistics*, 28(4):497–526, Dec.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting SimpleNLG for bilingual English-French realisation. In *14th European Conference on Natural Language Generation*, Sofia, Bulgaria, Aug. This volume.

Michael Zock and Guy Lapalme. 2010. A generic tool for creating and using multilingual phrasebooks. In Bernadette Sharp and Michael Zock eds., editors, *Proceedings of NLPCS 2010 (Natural Language Processing and Cognitive Science)*, pages 79–89, Funchal, Madeira - Portugal, Jun.

Michael Zock, Guy Lapalme, and Mehdi Yousfi-Monod. 2012. Learn to speak like normal people do: the case of object descriptions. In *9th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2012)*, pages 120–129, Wraclow, jun.