

Emiel Krahmer Mariët Theune (Eds.)

Empirical Methods in Natural Language Generation

Data-Oriented Methods
and Empirical Evaluation

Preface

Natural language generation (NLG) is a subfield of natural language processing (NLP) that is often characterized as the study of automatically converting non-linguistic representations (e.g., from databases or other knowledge sources) into coherent natural language text. NLG is useful for many practical applications, ranging from automatically generated weather forecasts to summarizing medical information in a patient-friendly way, but is also interesting from a theoretical perspective, as it offers new, computational insights into the process of human language production in general. Sometimes, NLG is framed as the mirror image of natural language understanding (NLU), but in fact the respective problems and solutions are rather dissimilar: while NLU is basically a disambiguation problem, where ambiguous natural language inputs are mapped onto unambiguous representations, NLG is more like a choice problem, where it has to be decided which words and sentences best express certain specific concepts.

Arguably the most comprehensive currently available text book on NLG is Reiter and Dale's [7]. This book offers an excellent overview of the different subfields of NLG and contains many practical insights on how to build an NLG application. However, in recent years the field has evolved substantially, and as a result it is fair to say that the book is no longer fully representative of the research currently done in the area of NLG. Perhaps the most important new development is the current emphasis on data-oriented methods and empirical evaluation. In 2000, data-oriented methods to NLG were virtually non-existent and researchers were just starting to think about how experimental evaluations of NLG systems should be conducted, even though many other areas of NLP already placed a strong emphasis on data and experimentation. Now the situation has changed to such an extent that *all* chapters in this book crucially rely on empirical methods in one way or another.

Three reasons can be given for this important shift in attention, and it is instructive spelling them out here. First of all, progress in related areas of NLP such as machine translation, dialogue system design and automatic text summarization created more awareness of the importance of language generation, even prompting the organization of a series of multi-disciplinary workshops on Using Corpora for Natural Language Generation (UCNLG). In statistical machine translation, for example, special techniques are required to improve the grammaticality of the translated sentence in the target language. N-gram models can be used to filter out improbable sequences of words, but as Kevin Knight put it succinctly "automated language translation needs generation help badly" [6]. To give a second example, automatic summarizers which go beyond mere sentence extraction would benefit from techniques to combine and compress sentences. Basically, this requires NLG techniques which do not take non-linguistic information as input, but rather (possibly ungrammatical) linguistic information

(phrases or text fragments), and as a result this approach to NLG is sometimes referred to as text-to-text generation. It bears a strong conceptual resemblance to text revision, an area of NLG which received some scholarly attention in the 1980s and 1990s (e.g., [8, 9]). It has turned out that text-to-text generation lends itself well for data-oriented approaches, in part because textual training and evaluation material are easy to come by.

In contrast, text corpora are of relatively limited value for “full” NLG tasks which are about converting concepts into natural language. For this purpose, one would prefer to have so-called semantically transparent corpora [4], which contain both information about the available concepts as well as human-produced realizations of these concepts. Consider, for instance, the case of referring expression generation, a core task of many end-to-end NLG systems. A corpus of human-produced referring expressions is only useful if it contains complete information about the target object (what properties does it have?) and the other objects in the domain (the distractors). Clearly this kind of information is typically not available in traditional text corpora consisting of Web documents, newspaper articles or comparable collections of data. In recent years various researchers have started collecting semantically transparent corpora (e.g., [5, 10]), and this has given an important boost to NLG research. For instance, in the area of referring expression generation, the availability of semantically transparent corpora has made it possible for the first time to seriously evaluate traditional algorithms and to develop new, empirically motivated ones.

The availability of suitable corpora also made it feasible to organize shared tasks for NLG, where different teams of researchers develop and evaluate their algorithms on a shared, held out data set. These kinds of shared tasks, including the availability of benchmark data sets and standardized evaluation procedures, have proven to be an important impetus on developments in other areas of NLP, and already a similar effect can be observed for the various NLG shared tasks (“generation challenges”) for referring expression generation [1], for generation of references to named entities in text [2] and for instruction giving in virtual environments [3]. These generation challenges not only resulted in new-generation research, but also in a better understanding of evaluation and evaluation metrics for generation algorithms.

Taken together these three developments (progress in related areas, availability of suitable corpora, organization of shared tasks) have had a considerable impact on the field, and this book offers the first comprehensive overview of recent empirically oriented NLG research. It brings together many of the key researchers and describes the state of the art in text-to-text generation (with chapters on modeling text structure, statistical sentence generation and sentence compression), in NLG for interactive applications (with chapters on learning how to generate appropriate system responses, on developing NLG tools that automatically adapt to their conversation partner, and on NLG as planning under uncertainty, as applied to spoken dialogue systems), in referring expression generation (with chapters on generating vague geographic descriptions, on realization of modifier orderings, and on individual variation), and in evaluation (with chapters

dedicated to comparing different automatic and hand-crafted generation systems for data-to-text generation, and on evaluation of surface realization, linguistic quality and affective NLG). In addition, this book also contains extended chapters on each one of the generation challenges organized so far, giving an overview of what has been achieved and providing insights into the lessons learned.

The selected chapters are mostly thoroughly revised and extended versions of original research that was presented at the 12th European Workshop on Natural Language Generation (ENLG 2009) or the 12th Conference of the European Association for Computational Linguistics (EACL 2009), both organized in Athens, Greece, between March 30 and April 3, 2009. Both ENLG 2009 and EACL 2009 were preceded by the usual extensive reviewing procedures and we thank Regina Barzilay, John Bateman, Anja Belz, Stephan Busemann, Charles Callaway, Roger Evans, Leo Ferres, Mary-Ellen Foster, Claire Gardent, Albert Gatt, John Kelleher, Geert-Jan Kruijff, David McDonald, Jon Oberlander, Paul Piwek, Richard Powers, Ehud Reiter, David Reitter, Graeme Ritchie, Matthew Stone, Takenobu Tokunaga, Kees van Deemter, Manfred Stede, Ielka van der Sluis, Jette Viethen and Michael White for their efforts.

April 2010

Emiel Krahmer
Mariët Theune

References

1. Belz, A., Gatt, A.: The attribute selection for GRE challenge: Overview and evaluation results. In: Proceedings of UCNLG+MT: Language Generation and Machine Translation, Copenhagen, Denmark, pp. 75–83 (2007)
2. Belz, A., Kow, E., Viethen, J., Gatt, A.: The GREC challenge 2008: Overview and evaluation results. In: Proceedings of the Fifth International Natural Language Generation Conference (INLG 2008), Salt Fork, OH, USA, pp. 183–191 (2008)
3. Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., Oberlander, J.: Report on the first NLG challenge on generating instructions in virtual environments (GIVE). In: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), Athens, Greece, pp. 165–173 (2009)
4. van Deemter, K., van der Sluis, I., Gatt, A.: Building a semantically transparent corpus for the generation of referring expressions. In: Proceedings of the 4th International Conference on Natural Language Generation (INLG 2006), Sydney, Australia, pp. 130–132 (2006)
5. Gatt, A., van der Sluis, I., van Deemter, K.: Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: Proceedings of the 11th European Workshop on Natural Language Generation (ENLG 2007), Saarbrücken, Germany, pp. 49–56 (2007)
6. Knight, K.: Automatic language translation generation help needs badly. Or: “Can a computer compress a text file without knowing what a verb is?” In: Proceedings of UCNLG+MT: Language Generation and Machine Translation, Copenhagen, Denmark, pp. 1–4 (2007)

7. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, Cambridge (2000)
8. Robin, J.: A revision-based generation architecture for reporting facts in their historical context. In: Horacek, H., Zock, M. (eds.) *New Concepts in Natural Language generation: Planning, Realization and Systems*. Frances Pinter, London (1993)
9. Vaughan, M.M., McDonald, D.D.: A model of revision in natural language generation. In: *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL 1986)*, New York, NY, USA, pp. 90–96 (1986)
10. Viethen, J., Dale, R.: Algorithms for generating referring expressions: Do they do what people do? In: *Proceedings of the 4th International Conference on Natural Language Generation (INLG 2006)*, Sydney, Australia, pp. 63–70 (2006)