Corpus-Based Methods in Natural Language Generation: Friend or Foe?

Extended Abstract

Owen Rambow
AT&T Labs – Research
Florham Park, NJ, USA
rambow@research.att.com

In computational linguistics, the 1990s were characterized by the rapid rise to prominence of corpus-based methods in natural language understanding (NLU). These methods include statistical and machine-learning and approaches. In natural language generation (NLG), in the mean time, there was little work using statistical and machine learning approaches. Some researchers felt that the kind of ambiguities that appeared to profit from corpus-based approaches in NLU did not exist in NLG: if the input is adequately specified, then all the rules that map to a correct output can also be explicitly specified. However, this paper will argue that this view is not correct, and NLG can and does profit from corpus-based methods. The resistance to corpus-based approaches in NLG may have more to do with the fact that in many NLG applications (such as report or description generation) the output to be generated is extremely limited. As is the case with NLU, if the language is limited, hand-crafted methods are adequate and successful. Thus, it is not a surprise that the first use of corpus-based techniques, at ISI (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998) was motivated by the use of NLG not in "traditional" NLG applications, but in machine translation, in which the range of output language is (potentially) much larger.

In fact, the situations in NLU and NLG do not actually differ with respect to the notion of ambiguity. Though it is not a trivial task, we can fully specify a grammar such that the generated text is not ungrammatical. But the problem for NLG is not specifying a grammar, but determining which part of the grammar to use: to give a simple example, a give-event can be generated with the double-object frame (*give Mary a book*) or with a prepositional object (*give a book to Mary*). We can easily specify the syntax of these two constrictions. What we need to know is when to choose which. But the situation is exactly the same in NLU: the problem is knowing which grammar rules to use when during analysis. Thus, just as the mapping from input to output is ambiguous in NLU, it is ambiguous in NLG, not because the grammar is wrong, but because it leaves too many options. The difference is that in NLG, different outputs differ not in whether they are correct (as is the case in NLU), but in whether they are appropriate or felicitous in a given context. Thus, the need for corpus-based approaches is less apparent.

Determining which linguistic forms are appropriate in what contexts is a hard task. The introspective grammaticality judgment that (perhaps) is legitimate in the study of syntax is methodologically suspect in the study of language use in context, and most work in linguistic pragmatics is in fact corpus-based, such as Prince's work using the Watergate transcripts and similar corpora (Prince, 1981). Thus, it is clear that the role of corpus-based methods in NLG is not to displace traditional methods, but rather to accelerate them. If indeed corpus-based methods are necessary in any case, we may as well use automated procedures for discovering regularities; we no longer need to use multi-colored pencils to mark up paper copies. For the researcher, there is enough left to do: the corpus-based techniques still require

linguistic research in order to determine which features to code for (i.e., what linguistic phenomena to count). To the extent that corpus-based methods fail currently, it is largely because we are substituting easily codable features for those that are more difficult to code, or because we are simply coding the wrong features. It is not because there is some hidden truth which traditional linguistic methodologies have access to but corpus-based methods do not, because they are not in fact in opposition to each other.

Finally, the emphasis on evaluation that the corpus-based techniques in NLU have brought with them have often aroused animosity in the NLG community. Evaluation is necessary for development purposes when using corpus-based techniques: it is easy to generate many different hypotheses, and we need to be able to choose among them. Since this is crucial, increased attention needs to be paid to evaluation in generation (Bangalore et al., 2000; Rambow et al., 2001). But again, the situation is in fact not different from a traditional linguistic methodology: theories about language use in context need to be defeasible on empirical grounds and hence need to be evaluated against a corpus. Of course, the choice of evaluation corpus is an important one, and the costs associated with compiling and annotating corpora can greatly impact the choice of evaluation corpus and hence the evaluation.

In conclusion, NLG has nothing to fear from corpus-based methods. Instead, the NLG community can continue to provide a test-bed for linguists to exercise their theories (to a much greater extent than can NLU). The difference is that everyone can now start using computers.

References

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the First International Natural Language Generation Conference (INLG2000)*, Mitzpe Ramon, Israel.

K. Knight and V. Hatzivassiloglou. 1995. Two-level, many-paths generation. In *33rd Meeting of the Association for Computational Linguistics (ACL'95)*.

Irene Langkilde and Kevin Knight. 1998. The practical value of n-grams in generation. In *Proceedings of the Ninth International Natural Language Generation Workshop (INLG'98)*, Niagara-on-the-Lake, Ontario.

Ellen F. Prince. 1981. Topicalization, focus movement and yiddish movement: A pragmatic dif ferentiation. In D. Alford, editor, *Proceedings of the Seventh Annual Meeting of the Berkely Linguistics Society*, pages 249–264. BLS.

Owen Rambow, Monica Rogati, and Marilyn Walker. 2001. Evaluating a trainable sentence planner for a spoken dialogue system. In *39th Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France.