Lexical Semantics and Knowledge Representation in Multilingual Text Generation

Manfred Stede
(Technische Universität Berlin)

Boston: Kluwer Academic Publishers
(The Kluwer international series in engineering and computer science, volume 492), 1999, xv+219 pp;
hardbound, ISBN 0-7923-8419-9,
$129.00, £84.00, Dfl 265.00

Reviewed by
Barbara Di Eugenio
University of Illinois at Chicago

Suppose you were to describe Tom draining oil from the engine of his car in 20 seconds. As a competent English speaker, depending on context, you could choose among at least the following subtly different descriptions, called paraphrases by Stede:

1. For 20 seconds, the oil drained from the engine.
2. The engine drained in 20 seconds.
3. Within 20 seconds, Tom drained the engine.
4. Tom drained the engine of the oil in 20 seconds.

In contrast, a state-of-the-art natural language generation (NLG) system would likely be able to produce only one of them, modulo variations introduced by, for example, passivization, topicalization, or pronominalization. The problem is not simply to choose among paraphrases, but to be able to produce them to start with.

In this book, an extended version of his dissertation, Stede provides the theoretical and practical means of endowing a text generator with such capability. He concentrates on **tactical planning**, namely, on choosing an appropriate verbalization for some content that another component of the NLG system (the **text planner**) has assembled in response to communicative goals. Stede's approach promotes lexicalization as the central step in this process and the lexicon itself as the link between the language-independent domain model and the language-specific resources necessary to generate the surface sentence. Stede's ultimate goal is ambitious: he seeks to provide an architecture that can be used as is to perform generation of the same content in different languages—even lexical entries are to be reused when possible. In his view, multilingual generation, including the problem of language divergences, can be seen as a mere extension of the monolingual paraphrase task. Whereas the methods proposed are very compelling with respect to monolingual generation, it is not clear whether the multilingual goal has been fully achieved. I will come back to this point at the end of this review.

The book is structured in three main parts:

- Chapters 1 through 3 provide the introduction and background on NLG, on lexicalization, and on lexical variation. Further background is sprinkled throughout the book, in particular regarding verbal aspect (Chapter 4) and lexical semantics (Chapter 6).

- Chapters 4 through 8 constitute the core of the book; I will discuss them below.
- Finally, Chapter 9 showcases the paraphrasing capabilities of the generator through a number of examples. Chapter 10 is an interesting although occasionally weak description of how the approach can be extended to generate paragraphs, not just sentences. Chapter 11 summarizes the relation between the author's and others' approaches, and speculates on a few directions for future research.

At first I found Chapters 4 through 8 slightly overwhelming, as they introduce several levels of representation, each with its own terminology and acronyms. However, at a second, more-careful, reading, everything falls into place. The resulting approach has at its center a lexicon that partly implements current theories of lexical semantics such as Jackendoff's (1990) and Levin's (1993). The lexicon is used to mediate and map between a language-independent domain model and a language-dependent ontology widely used in NLG, the Upper Model (Bateman 1990). Although the idea of a two-level representation accommodating language-neutral and language-specific requirements is not new (see for example Nirenburg and Levin [1992], Dorr and Voss [1993], and Di Eugenio [1998]), Stede is among the few who make effective use of those two levels in a complex system.

Chapter 4 presents the domain model built by means of the description logic system LOOM (MacGregor and Burstein 1991). Stede is specifically interested in verbalizations of situations, to use his own neutral term. Thus, the domain model contains a representation for categories such as states, activities, and events that Stede nicely synthesizes from linguistic work on aspect.[1]

Chapter 5 discusses the different levels of representation used by the generator: the language-neutral level of **situation specifications** (SitSpecs), built as instantiations of portions of the domain model; and the language-specific level of **semantic sentence specifications** (SemSpecs), written in terms of Upper Model concepts and relations. Importantly, SemSpecs are lexicalized. A SemSpec for a sentence is built by collecting all possible verbalization options that cover a subset of the meaning expressed by the SitSpec to be verbalized, and combining them appropriately. In the same way that the domain model provides the foundation for the well-formedness of SitSpecs, the Upper Model guarantees that a SemSpec can actually be correctly converted to linguistic output.

Chapter 6 describes the lexicon. Among other information, each lexical entry provides a denotation, which is a small excerpt of domain-model concepts and relations and represents the meaning associated with the lexical entry, and a **partial SemSpec** (PSemSpec), which describes the contribution of this lexical entry to the SemSpec representing the sentence. A PSemSpec can point to lexical entries in different languages if they are equivalent in denotation (e.g., *rise* in English and *steigen* in German). The correspondence between denotation and PSemSpec is maintained by coindexed variables.

Chapter 7 presents the lexical resources to generate verbal alternations. Stede is interested in alternations that change verbal meanings, such as the resultative–causative

1 Stede argues that it is not contradictory to develop a language-independent ontology by exploiting some linguistic research.

alternation that transforms an activity into an event (e.g., paraphrase 4 in the example at the beginning of this review is the resultative–causative version of paraphrase 1). Stede encodes such transformations, which he derives from work by Jackendoff (1990) and Levin (1993), as rules that are attached to the lexical entry of each verb that can undergo that transformation. Stede assumes that these rules are monotonic, that is, they only add components of meaning to a base form but don't eliminate any. As Stede concedes, this assumption may be too strong, although it appears to be appropriate for the examples he presents.

Chapter 8 presents the computational architecture (his system is called MOOSE) and the procedures that first generate all possible verbalization options (interestingly, by exploiting LOOM subsumption check) and then combine them to find the best one. The architecture has a "plug" for preference parameters, such as formality, that are used to choose a paraphrase over another. The link between preference parameters and SemSpecs is through other aspects of lexical entries also discussed in Chapter 6, such as **connotation**, that concern stylistic features. As previously mentioned, Chapter 9 shows the computational architecture at work through a range of examples.

Two aspects of the book left me somewhat unsatisfied. I did not expect a formal evaluation, which is still an open problem for NLG, and that appears to be very difficult given the complex nature of Stede's work. However, I would have liked some information regarding the size of the knowledge base and of the lexicon. I would also have liked to know whether an exhaustive experimentation with the prototype has been carried out and answers to questions such as: does MOOSE ever fail to produce a verbalization? does MOOSE ever produce an infelicitous one?

The second aspect is whether the multilingual goal has been satisfactorily achieved, given the closeness of the two chosen languages, English and German. Obviously, different languages require different lexicons and at least different portions of the Upper Model. However, it is unclear whether the computational flow of information embodied in the architecture in Stede's Figure 8.1 can remain unchanged when tackling less closely related languages. Stede acknowledges this problem at the beginning of Chapter 3, but does not revisit it later; in fact he does not even mention it as an issue for future work.

I recommend the book not only to researchers interested in text generation and in machine translation, but to everybody interested in the relationship between language-independent knowledge representation and language-specific ontologies. Whereas the hypothesis that the missing link between the two is the lexicon is not surprising nowadays, Stede's specific proposal is well defined and effective.

References

Bateman, John A. 1990. Upper modeling: Organizing knowledge for natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Pittsburgh, PA.

Di Eugenio, Barbara. 1998. An action representation formalism to interpret natural language instructions. *Computational Intelligence*, 14(1):89–133.

Dorr, Bonnie J. and Clare R. Voss. 1993. Machine translation of spatial expressions: Defining the relation between an interlingua and a knowledge representation system. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-93)*.

Jackendoff, Ray. 1990. *Semantic Structures*. The MIT Press, Cambridge, MA.

Levin, Beth C. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

MacGregor, Robert and Mark H. Burstein. 1991. Using a description classifier to enhance knowledge representation. *IEEE Expert*, 6(3):41–46.

Nirenburg, Sergei and Lori Levin. 1992. Syntax-driven and ontology-driven lexical semantics. In James Pustejovsky and Sabine Bergler, editors, *Lexical Semantics and Knowledge Representation*. Springer Verlag.

Barbara Di Eugenio is an assistant professor in the department of Electrical Engineering and Computer Science at the University of Illinois at Chicago. She is interested in the interpretation and generation of instructional text and dialogue, specifically in issues of knowledge representation and lexical semantics concerning action verbs. Di Eugenio's address is: EECS Department, 1120 SEO (M/C 154), 851 S. Morgan St, University of Illinois at Chicago, Chicago, IL, 60607, USA; e-mail: bdieugen@eecs.uic.edu