Machine Translation and the Lexicon

Petra Steffens (editor)
(IBM Deutschland Informationssysteme GmbH)

Heidelberg: Springer-Verlag (Lecture
Notes in Artificial Intelligence, edited
by Jaime G. Carbonell and J. Siekmann,
volume 898), 1995, x+251 pp;
paperbound, ISBN 3-540-59040-4,
DM 58.00

Reviewed by
Inderjeet Mani
The MITRE Corporation

The practical success of machine translation (MT) depends on the ability to acquire, share, and manage lexical data. Rather than reinventing lexicons for each new system and application, it is preferable to leverage common lexical resources. Increasingly, researchers are using pre-existing resources such as machine-readable dictionaries (MRDs) and corpora to acquire lexicons and term banks for MT, as well as developing new resources in such a way as to facilitate their sharing and reuse. *Machine Translation and the Lexicon* offers practical perspectives on these activities, from the standpoint of researchers and of commercial developers and users in Europe. The book consists of revised versions of a subset of the papers presented at the Third International Workshop of the European Association for Machine Translation (EAMT) held in Heidelberg in April 1993. The book's 15 papers are spread over three sections: Part I, Acquiring Lexical Data (5 papers); Part II, Managing Lexical Data (7 papers); and Part III, Describing Lexical Data (3 papers). The editor, Petra Steffens of IBM Deutschland, presents an excellent introduction, which includes a useful bibliography of recent lexical work related to MT.

In recent years, attention has shifted from machine-readable dictionaries towards corpora as a source for acquiring lexical information. Part I kicks off with a well-written article by Ide and Véronis illustrating some of the well-known problems involved in attempting to extract class hierarchies from dictionaries, such as inconsistencies, circularities, and incompleteness in dictionary sense definitions, as well as the knowledge-intensive nature of the extraction task (many patterns have to be coded, and word-sense disambiguation may require substantial world knowledge). The authors conclude that "all of this means that in order to create resources for use in NLP from MRDs, it is necessary to have full NLP capabilities—including full knowledge bases—already at hand" (p. 27). However, they do not make clear what sorts of NLP requirements they have in mind; their statement (with its vague use of "full") does not do justice to the sophistication of bootstrapping techniques used in various dictionary extraction projects. For example, the approach of Wilks et al. (1993) relies on expanding out from seed word senses identified in definitions of "controlled vocabulary" words used in dictionary definitions, and Vanderwende (1995) describes a multipass approach that defers processing of ambiguous patterns to later passes.

For the future, Ide and Véronis recommend backing off from fully automatic extraction and focusing on merging information from multiple lexical resources, such as corpora and multiple dictionaries. These are in fact the directions taken by numerous lexical acquisition projects. Among the successes of the dictionary extraction work, Ide and Véronis cite the Text Encoding Initiative's development of a dictionary encoding format (Sperberg-McQueen and Burnard 1994), which is a significant step towards dictionary reuse, and the increased synergy between electronic publishing, NLP, and lexicography. Evidence of this synergy is found in Procter's paper on the exploitation of corpora by lexicographers in the Cambridge Language Survey (CLS), which is a large-scale multilingual project involving publishers, industrial labs, and universities in several European countries, aimed at building lexical databases to support both dictionary publishing and NLP lexicons. Procter describes how the CLS plans to collect various corpora, including non-native language corpora. Each word in a corpus will be annotated with codes for part of speech, semantic features, and subjects, and will be linked (in the case of English) to a record in the *Cambridge International Dictionary of English*. Procter doesn't discuss what standards, if any, will apply to these different CLS coding schemes.

The other papers in Part I include those by Storrer and Schwall on the acquisition of multiword lexemes, Ahmad on the acquisition of technical terms from corpora, and Daelemans on using machine learning for lexical acquisition. I was surprised to find no articles discussing the use of parallel corpora. Storrer and Schwall discuss some highly informal feasibility studies investigating acquisition from dictionaries and corpora of verbal idioms (e.g., *kick the bucket*) and support-verb constructions (e.g., *take into consideration*). Ahmad describes statistical techniques to identify technical terms, assuming that various "specialist" corpora, rich in technical terms, are available. Daelemans argues against the notion that there could ever be anything like a common NLP lexicon for a language, taking the position that the types of lexical information needed are highly task-specific. Instead, he envisages reuse arising from the reapplication of a single learning method to different problems. However, he provides little by way of quantitative results.

Part II is mainly about standards for reuse and management of lexical data in commercial systems. I will confine myself to a few relevant and better-written papers. Calzolari describes standardization efforts by the Expert Advisory Group on Language Engineering Standards. In focusing on architectures for reuse of lexical and terminological resources, IBM Deutschland's TransLexis system for managing MT system lexicons and term banks counts as a fairly advanced framework. As Bläser describes it, their goal is to build a "theory-neutral representation of multilingual lexical and terminological data" (p. 159). The system supports four different formats for the exchange of lexicons and term banks, including two SGML formats. Terminology and lexical entries can be imported and merged automatically with existing entries using a statistical algorithm. Another collaborative effort is the CEC-funded ESPRIT project Translator's Workbench (TWB), which brings together several European translation departments, tool vendors, and research groups to develop tools for professional translators. Mayer's paper describes the design of the TWB multilingual term bank, which contains terminology on automotive engineering in English, German, and Spanish.

In Part III, Caroli provides a fairly extensive classification of German multiword lexemes, including idioms, collocations, support-verb constructions, metaphors, etc., in terms of a scale of compositionality. Unfortunately, examples of idiomatic expressions in German are accompanied by corresponding English idiomatic translations, but without word-for-word English glosses, making it difficult for readers unfamiliar with German. The papers by Ostler and Heid describe the overall approach of the DELIS project, another large-scale collaborative effort whose goal, like that of the CLS, is to develop a system to build lexical databases for both dictionary publishing and NLP applications. In particular, they aim for a corpus-based examination of the syntactic and semantic properties of perception and speech-act vocabulary in five languages:

English, Danish, Dutch, French, and Italian. Their use of Fillmore's frame semantics makes possible certain fine semantic distinctions, for example, the source of the percept (e.g., *hear a dog*) is distinguished from the stimulus perceived (e.g., *hear a bark*). It will be interesting to see what sorts of cross-linguistic generalizations will emerge from such distinctions and the corpora in use (which, by the way, aren't identified, except for a footnote reference to an 18-million-word corpus used by Oxford University Press). Their project expects to eventually link the different monolingual lexicons in this vocabulary domain. With regard to structuring these multilingual lexicons, Heid discusses the potential relevance of classifications of lexical differences in terms of divergences and mismatches (e.g., Dorr 1990, Barnett, Mani, and Rich 1994); the corpora to which these lexicons are to be linked could provide useful data for testing these and other classifications.

Overall, despite the fine introduction and several interesting papers, the book offers an uneven mix. While a high-level project report or system overview may work well in the ambiance of a workshop setting, it becomes less attractive in the pages of a book. I think the book will be of interest primarily to readers seeking an overview of some of the issues of lexicon data management and reuse that various groups in Europe are addressing through collaborative efforts. Although it is quite helpful in indicating what sorts of products we can expect from this collaboration, it will be less useful to readers with more specialized needs, for example, MT researchers examining techniques for lexical acquisition.

References

Barnett, James, Inderjeet Mani, and Elaine Rich. 1994. "Reversible machine translation: What to do when the languages don't match up." In Tomek Strzalkowski, editor, *Reversible Grammar in Natural Language Processing.* Dordrecht: Kluwer Academic Publishers.

Dorr, Bonnie. 1990. "Solving thematic divergences in machine translation." In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics,* Pittsburgh, pages 127–134.

Sperberg-McQueen, C. M. and Lou Burnard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford: Text Encoding Initiative. *http://www.uic.edu/orgs/tei/info/guide.html.*

Vanderwende, Lucy. 1995. "Ambiguity in the acquisition of lexical information." In *Working Notes of the AAAI Spring Symposium on the Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity,* pages 174–179, Stanford University, March.

Wilks, Yorick, Dan Fass, Cheng-Ming Guo, James McDonald, Tony Plate, and Brian Slator. 1993. "Providing machine tractable dictionary tools." In James Pustejovsky, editor, *Semantics and the Lexicon,* Dordrecht: Kluwer Academic Publishers.

Inderjeet Mani is a Principal Scientist at MITRE. His research has addressed problems in lexical semantics, machine translation, text generation, and information extraction and retrieval. His doctoral dissertation is on the semantics of nominalizations. Recently, he has explored applications of NLP and machine learning to on-line news browsing and summarization. Mani's address is: Artificial Intelligence Technical Center, W640, The MITRE Corporation, 11493 Sunset Hills Road, Reston, VA 22090; e-mail: imani@mitre.org