Corpus-Based Induction of Lexical Representation and Meaning

Maria Lapata
School of Cognitive Science, University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
mlap@cogsci.ed.ac.uk

Motivation

The acquisition of linguistic knowledge, i.e., the identification, extraction, and encoding of linguistic information in a corpus, has been one of the main motivations for data-driven approaches to natural language. Methods have been developed for the acquisition of, for instance, parts of speech, noun compounds, collocations, support verbs, subcategorization frames, phrase structure rules, selectional restrictions and sense induction (cf. Armstrong (1993) for an overview).

Drawing on this body of research, I am investigating the acquisition of lexical semantic knowledge from corpora, thereby addressing the logical problem of language acquisition, one of the fundamental issues in linguistics and cognitive science. My guiding assumption is that syntactic as well as semantic representations are projected from information in the lexicon, and that a crucial part of the relevant lexical information is the result of language experience, and hence can be induced from corpora.

The proposed research includes three main subtasks: (a) induction of different types of ("low-level") lexical semantic information (i.e., subcategorization frames, selectional restrictions, semantic classes), using established corpus-based methods; (b) combination of the induced types of lexical semantic information into ("high-level") semantic representations, based on existing theories of the lexicon; (c) evaluation of the resulting model against human intuitions.

By applying corpus-based techniques to lexical semantics, i.e., a classical representational problem in linguistics, I hope to contribute to bridging the gap between current data-driven approaches to language and the knowledge-driven methods of traditional linguistics.

Results

I have carried out work on automatically acquiring subcategorization frames from the British National Corpus (BNC) and showing that subcategorization preferences can be modeled as differences in frame frequencies derived from corpora. A high correlation was obtained between frequencies acquired by the subcategorization learning model and frequencies reported in psycholinguistic studies (Lapata & Keller 1998).

I have also examined the extent to which diathesis alternations (changes in the realization of the argument structure of a verb accompanied by changes in meaning) are empirically attested in corpus data. The research focuses on the automatic acquisition of alternating verbs from the BNC by using partial-parsing methods and taxonomic information (i.e., WordNet) and demonstrates how type and token frequencies acquired from corpus data can be used to quantify linguistic generalizations such as the productivity of an alternation and the typicality of its members (Lapata 1999).

Finally, I have carried out some preliminary work on using subcategorization information to disambiguate verb semantic classes. This work casts the task of verb class disambiguation in a probabilistic framework which exploits Levin's (1993) taxonomy of verbs and frame frequencies acquired from the BNC (Lapata & Brew 1999).

These three studies support the following claims:

- lexical preferences can be reliably acquired from corpora and shown to correspond to human intuitions as measured by psycholinguistic experiments;
- corpus frequencies can be used to quantify lexical generalizations such as Levin's semantic classification;
- the framework of a semantic theory (i.e., Levin) allows for the acquisition of refined semantic features that do not emerge from purely corpus-based collocational analysis.

References

Armstrong, S., ed. 1993. *Using Large Corpora.* Cambridge, MA: MIT Press.

Lapata, M., and Brew, C. 1999. Using subcategorization to resolve verb class ambiguity. Unpubl. ms., School of Cognitive Science, University of Edinburgh.

Lapata, M., and Keller, F. 1998. Corpus frequency as a predictor of verb bias. Poster at *AMLAP-98*, Freiburg.

Lapata, M. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. Proceedings of *ACL-99*, University of Maryland, College Park.

Levin, B. 1993. *English Verb Classes and Alternations.* Chicago: University of Chicago Press.