The Lexicon II

Miriam Butt
November 2002

Hand-Coding vs. Lexicon Induction

Languages contain closed class as well as open class items

Closed Class: Auxiliaries, Prepositions, Modals, Determiners, Adverbs

Open Class: Verbs, Nouns (also Compound Nouns), Adjectives

For the closed class items, hand-coding is inevitable because they have such special functions (functional rather than lexical elements).

Lexicon Induction

But --- could one come up with an automatic process to build a lexicon for the open classes?

In particular, since language is always changing, can one identify the new items automatically by searching through corpora? (NB, the Web is one big corpus).

Lexicon Induction

Problems:

- Nouns are inflected and often compounded (German): how to find the right lemma?

- Verbs are also inflected and in addition one needs to specify information about their subcategorization (argument structure) properties.

Stemmers

Stemmers have been used since the 80s (Porter Stemmer, J&M p.83) to strip off morphology from the end of a world and arrive at the stem.

The algorithms tend to be quite simple (see Snowball algorithm for German) and therefore also computationally cheap.

They are effective if one is prepared to live with some noise, especially for languages other than English.

Stemmers For English

Ferber (2000):

Word	**Stem**	**Should Be**
rating	rat	rate
rats	rat	rat

Stemmers For German

Snowball Stemmer (http://snowball.tartarus.org/):

Word	**Stem**	**Should Be**
Kater (male cat)	Kat	Kater
katholische	kathol	katholisch

Becker and Konig (2002):

Word	**Stem**	**Should Be**
Buchen (kind of trees)	Buch (book)	Buche
Pfluge (plows)	Pflug	Pflug
Luge (lie)	Lug	Luge

Stemmers For German

Becker and Konig (2002) found that a stemmer **in combination with** an already existing **lexicon** worked quite well: 85.92%

This would appear to be generally true, which brings us back to the question: how does one get a big lexicon quick and fast?

Verb Lexicon Induction

General Procedure (with some variations):

1. Parse a given corpus with some available shallow parser (chunk parser, finite-state parser, stochastic parser).
2. Figure out the "head" or stems for the verbs using a relatively smart stemmer, morphological analyzer or a combination of a stemmer and an existing lexicon.
3. Identify the arguments of the verb (e.g., positional in English, by case in German).

Verb Lexicon Induction

Problems:

1. How to tell PP arguments from adjuncts?
2. Intransitives are hard to tell.
3. Dative Constructions in German.

General Success Rate:

Schulte im Walde (German): 62.30%

Brent (1993, English): 73.85%

Carroll and Rooth (1998, English): 76.95%

Verb Lexicon Induction

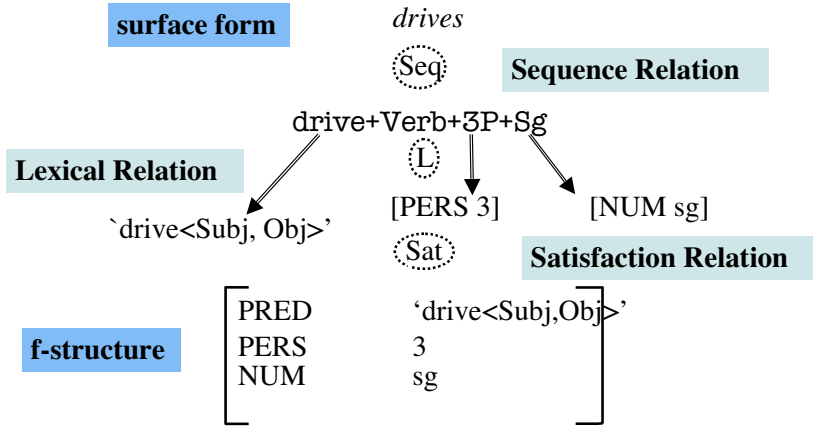
The Results of Automatic Lexicon Induction thus have to be hand-checked.

After a few rounds of this, one does get a useful and large lexicon (e.g., the ParGram German lexicon).

XLE

FST Morphological Analyzers

The morphology-syntax interface for an LFG grammar:



References

Becker, Tanja and Esther König. 2002. Lexikonfreie Lemmatisierung fur Substantive des Deutschen. In Proceedings of Konvens 2002, Saarbrucken, October.

Ferber, Reginald. 2000. Data Mining and Information Retrieval. Ms. Fernuniversität Hagen, January. http://teefix.fernuni-hagen.de/~ferber/kurse/dm-ir/vl/book_1.html

Hull, David A. and Gregory Grefenstette. 1996. A detailed analysis of English stemming algorithms. TR MLTT-023. RXRC, January. http://www.xrce.xerox.com/pblis/mltt/mltt-023.ps

Schulte im Walde, Sabine. 1998. Automatic Semantic Classification of Verbs According to their Alternation Behavior. Diplomthesis, IMS Stuttgart.