Discovering the Lexical Features of a Language

Eric Brill *
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
email: brill@unagi.cis.upenn.edu

1 Introduction

This paper examines the possibility of automatically discovering the lexical features of a language. There is strong evidence that the set of possible lexical features which can be used in a language is unbounded, and thus not innate. Lakoff [Lakoff 87] describes a language in which the feature *±woman-or-fire-or-dangerous-thing* exists. This feature is based upon ancient folklore of the society in which it is used. If the set of possible lexical features is indeed unbounded, then it cannot be part of the innate Universal Grammar and must be learned. Even if the set is not unbounded, the child is still left with the challenging task of determining which features are used in her language.

If a child does not know a priori what lexical features are used in her language, there are two sources for acquiring this information: semantic and syntactic cues. A learner using semantic cues could recognize that words often refer to objects, actions, and properties, and from this deduce the lexical features: noun, verb and adjective. Pinker [Pinker 89] proposes that a combination of semantic cues and innate semantic primitives could account for the acquisition of verb features. He believes that the child can discover semantic properties of a verb by noticing the types of actions typically taking place when the verb is uttered. Once these properties are known, says Pinker, they can be used to reliably predict the distributional behavior of the verb. However, Gleitman [Gleitman 90] presents evidence that semantic cues are not sufficient for a child to acquire verb features and believes that the use of this semantic information in conjunction with information about the subcategorization properties of the verb may be sufficient for learning verb features.

This paper takes Gleitman's suggestion to the extreme, in hope of determining whether syntactic cues may not just aid in feature discovery, but may be all that is necessary. We present evidence for the sufficiency of a strictly syntax-based model for discovering the lexical features of a language. The work is based upon the hypothesis that whenever two words are semantically dissimilar, this difference will manifest itself in the syntax via lexical distribution (in a sense, playing out the notion of distributional analysis [Harris 51]). Most, if not all, features have a semantic basis. For instance, there is a clear semantic difference between most count and mass nouns. But while meaning specifies the core of a word class, it does not specify precisely what can and cannot be a member of a class. For instance, *furniture* is a mass noun in English, but is a count noun in French. While the meaning of *furniture* cannot be sufficient for determining whether it is a count or mass noun, the distribution of the word can.

Described below is a fully implemented program which takes a corpus of text as input and outputs a fairly accurate word class list for the language in question. Each word class corresponds to a lexical feature. The program runs in $O(n^3)$ time and $O(n^2)$ space, where n is the number of words in the lexicon.

2 Discovering Lexical Features

The program is based upon a Markov model. A Markov model is defined as:

1. A set of states
2. Initial state probabilities **init(x)**
3. Transition probabilities **trans(x,y)**

An important property of Markov models is that they have no memory other than that stored in the current state. In other words, where X(t) is the value given by the model at time t,

$$Pr(X(t) = x_t \mid X(t-1) = x_{t-1} \ldots X(0) = x_0) =$$
$$Pr(X(t) = x_t \mid X(t-1) = x_{t-1})$$

In the model we use, there is a unique state for each word in the lexicon. We are not concerned with initial state probabilities. Transition probabilities represent the probability that word b will follow a and are estimated by examining a large corpus of text. To estimate the transition probability from state a to state b:

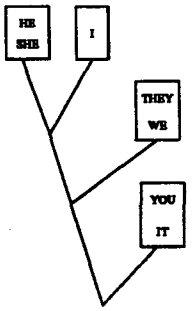
*The author would like to thank Mitch Marcus for valuable help. This work was supported by AFOSR jointly under grant No. AFOSR-90-0066, and by ARO grant No. DAAL 03-89-C0031 PRI.

1. Count the number of times b follows a in the corpus.

2. Divide this value by the number of times a occurs in the corpus.

Such a model is clearly insufficient for expressing the grammar of a natural language. However, there is a great deal of information encoded in such a model about the distributional behavior of words with respect to a very local context, namely the context of immediately adjacent words. For a particular word, this information is captured in the set of transitions and transition probabilities going into and out of the state representing the word in the Markov model.

Once the transition probabilities of the model have been estimated, it is possible to discover word classes. If two states are sufficiently similar with respect to the transitions into and out of them, then it is assumed that the states are equivalent. The set of all sufficiently similar states forms a word class. By varying the level considered to be sufficiently similar, different levels of word classes can be discovered. For instance, when only highly similar states are considered equivalent, one might expect animate nouns to form a class. When the similarity requirement is relaxed, this class may expand into the class of all nouns. Once word classes are found, lexical features can be extracted by assuming that there is a feature of the language which accounts for each word class. Below is an example actually generated by the program:



With very strict state similarity requirements, *HE* and *SHE* form a class. As the similarity requirement is relaxed, the class grows to include *I*, forming the class of singular nominative pronouns. Upon further relaxation, *THEY* and *WE* form a class. Next, *(HE, SHE, I)* and *(THEY, WE)* collapse into a single class, the class of nominative pronouns. *YOU* and *IT* collapse into the class of pronouns which are both nominative and accusative. Note that next, *YOU* and *IT* merge with the class of nominative pronouns. This is because the program currently deals with bimodals by eventually assigning them to the class whose characteristics they exhibit most strongly. For another example of this, see *HER* below.

3 Results and Future Directions

This algorithm was run on a Markov model trained on the Brown Corpus, a corpus of approximately one million words [Francis 82]. The results, although preliminary, are very encouraging. These are a few of the word classes found by the program:

- CAME WENT
- THEM ME HIM US
- HER HIS
- FOR ON BY IN WITH FROM AT
- THEIR MY OUR YOUR ITS
- ANY MANY EACH SOME
- MAY WILL COULD MIGHT WOULD CAN SHOULD MUST
- FIRST LAST
- LITTLE MUCH
- MEN PEOPLE MAN

This work is still in progress, and a number of different directions are being pursued. We are currently attempting to automatically acquire the suffixes of a language, and then trying to class words based upon how they distribute with respect to suffixes.

One problem with this work is that it is difficult to judge results. One can eye the results and see that the lexical features found seem to be correct, but how can we judge that the features are indeed the correct ones? How can one set of hypothesized features meaningfully be compared to another set? We are currently working on an information-theoretic metric, similar to that proposed by Jelinek [Jelinek 90] for scoring probabilistic context-free grammars, to score the quality of hypothesized lexical feature sets.

References

[Francis 82] Francis, W. and H. Kucera. (1982) *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin Co.

[Gleitman 90] Gleitman, Lila. (1990) "The Structural Sources of Verb Meanings." *Language Acquisition,* Volume 1, pp. 3-55.

[Harris 51] Harris, Zelig. (1951) *Structural Linguistics.* Chicago: University of Chicago Press.

[Jelinek 90] Jelinek, F., J.D. Lafferty & R.L. Mercer. (1990) "Basic Methods of Probabilistic Context Free Grammars." *I.B.M. Technical Report,* RC 16374.

[Lakoff 87] Lakoff, G. (1987) *Women, Fire and Dangerous Things: What Categories Reveal About the Mind.* Chicago: University of Chicago Press.

[Pinker 89] Pinker, S. *Learnability and Cognition.* Cambridge: MIT Press.