Automatic Acquisition of Subcategorization Frames from Tagged Text

Michael R. Brent and Robert C. Berwick

MIT Artificial Intelligence Lab
545 Technology Square
Cambridge, MA 02139

ABSTRACT

This paper describes an implemented program that takes a tagged text corpus and generates a partial list of the subcategorization frames in which each verb occurs. The completeness of the output list increases monotonically with the total occurrences of each verb in the training corpus. False positive rates are one to three percent. Five subcategorization frames are currently detected and we foresee no impediment to detecting many more. Ultimately, we expect to provide a large subcategorization dictionary to the NLP community and to train dictionaries for specific corpora.

INTRODUCTION

Accurate parsing requires knowing the subcategorization frames of verbs, as shown by (1).

(1) a. I expected [NP the man who smoked NP] to eat ice-cream
b. I doubted [NP the man who liked to eat ice-cream NP]

Current high-coverage parsers tend to use either custom, hand-generated lists of subcategorization frames (e.g., [7]), or published, hand-generated lists like the *Oxford Advanced Learner's Dictionary of Contemporary English*, [9] (e.g., [5]). In either case, such lists are expensive to build and to maintain in the face of evolving usage. In addition, they tend not to include rare usages or specialized vocabularies like financial or military jargon. Further, they are often incomplete in arbitrary ways. For example, Webster's Ninth New Collegiate Dictionary lists the sense of *strike* meaning "to occur to", as in "it struck him that...", but it does not list that same sense of *hit*. (Our program discovered both.) To address these problems we have implemented a program that takes a tagged text corpus and generates a partial list of the subcategorization frames in which each verb occurs. The program uses only a small, finite-state grammar for a fragment of English. The completeness of the output list increases monotonically with the total number of occurrences of each verb in the training corpus.

Automatically learning subcategorization frames (SFs) is impeded by a bootstrapping problem — you can't parse without knowing SFs and you can't learn from examples without parsing them. For instance, the obvious approach to identifying verbs that take infinitival complements would be to look for a verb followed by an infinitive. Unfortunately, as shown by (1), finding such a case does not license any definite conclusions. Our system bootstraps by recognizing those sentences that it can parse without already knowing the SFs — mainly sentences involving pronouns or proper names rather than full noun-phrases in certain argument positions. It simply ignores other sentences. The distributional constraints on pronouns and full noun-phrases are almost identical, so lessons learned in the easy-to-parse cases apply to all cases.

The remainder of this paper consists of a section describing and quantifying our results, a section describing the methods used to obtain them, and a section discussing related work.

RESULTS

So far, we have concentrated on the five subcategorization frames shown in Table 1. Table 2 shows the results obtained on a 2.6 million-word Wall Street Journal corpus

SF Description	Good Example	Bad Example
direct object	hit them	* arrive them
direct object & clause	tell him he's a fool	* slap him he's a fool
direct object & infinitive	want him to attend	* hope him to attend
clause	know I'll attend	* want I'll attend
infinitive	hope to attend	* hit to attend

Table 1: The five subcategorization frames detected so far

provided by Penn Treebank project ([2]).[1]

SF	tokens found	% false positives & source of error
DO	8,606	1.0% Subj of comp clause taken for DO 0.5% Adv taken for DO
DO & clause	381	1.0% Rel. clause taken as comp. clause 0.5% Fronted adjunct taken as main clause 0.5% Comp belonged to a higher verb
DO & inf	3,597	1.5% Purposive adjuncts taken for inf.
clause	14,144	Demonstrative "that" taken for comp.
inf comp	11,880	2.0% Purposive adjuncts taken for inf. 1.0% Adjective comp like "hard to take"

Quantity of text processed	=	2,644,618	words of WSJ
Total time (tagged, SPARC 2)	=	192.5	seconds
Throughput Rate	=	13738.3	words/seconds

Table 2: Top: Lexicographic results, error rates, and sources of error. Bottom: speed and volume.

METHODOLOGY

Our program uses a finite-state grammar for recognizing the auxiliary, and determining subcategorization frames. The English auxiliary system is known to be finite state and our treatment of it is standard, so the first subsection discusses the determination of subcategorization frames. The second subsection describes a planned statistical approach to the one to three percent error rates described above.

Complement Grammar

The obvious approach to finding an SF like "V NP to V" is to look for occurrences of just that pattern in the training corpus, but the obvious approach fails to address the bootstrapping problem, as shown by (1) above. Our solution is based on the following insights:

- Some examples are clear and unambiguous.
- Observations made in clear cases generalize to all cases.
- It is possible to distinguish the clear cases from the ambiguous ones with reasonable accuracy.
- With enough examples, it pays to wait for the clear cases.

[1] Error rates computed by hand verification of 200 examples for each SF using the tagged mode. Error rates for verb detection are estimated separately below.

Rather than take the obvious approach of looking for "V NP to V", we look for clear cases like "V PRONOUN to V". The advantages can be seen by contrasting (2) with (1) (page 1).

(1) a. OK I expected him to eat ice-cream
b. * I doubted him to eat ice-cream

More generally, our system recognizes linguistic structure using a small finite-state grammar that describes only that fragment of English that is most useful for recognizing SFs. The grammar relies exclusively on closed-class lexical items such as pronouns, prepositions, determiners, and auxiliary verbs.

The complement grammar needs to distinguish three types of complements: direct objects, infinitives, and clauses. Figure 1 shows a substantial part of the grammar responsible for detecting these complements. Any verb followed immediately by matches for <DO>, <clause>, <infinitive>, <DO><clause>, or <DO><inf> is assigned the corresponding SF. Adverbs are ignored for purposes of adjacency. The notation "?" follows optional expressions, and DO is specified in context-sensitive notation for convenience.

```
<clause>        := that?
                   (<subj-pron> | <subj-obj-pron> |
                     his | <proper-name>)
                   <tensed-verb>
<subj-pron>     := I | he | she | we | they
<subj-obj-pron> := you, it, yours, hers, ours, theirs

<DO>            := <obj-pron>
<DO>            := (<subj-obj-pron> | <proper-name>):
                   <tensed-verb> ____
<obj-pron>      := me | him | us | them

<infinitive>    := to <uninflected-verb>
```

Figure 1: A non-recursive (finite-state) grammar for detecting certain verbal complements. "?" indicates an optional element. <DO> is specified in context-sensitive notation, for convenience. Any verb followed immediately expressions matching <DO>, <clause>, <infinitive>, <DO> <clause>, or <DO> <infinitive> is assigned the corresponding SF.

Robust Classification

Our system, like any other, occasionally makes mistakes. Error rates of one to three percent are a substantial accomplishment, but if a word occurs enough times in a corpus it is bound to show up eventually in some construction that fools the system. For that reason any learning

system that gets only positive examples and makes a permanent judgment on a single example will always degrade as the number of occurrences increases. In fact, making a judgment based on any fixed number of examples with any finite error rate will always lead to degradation with corpus-size. A better approach is to require a fixed percentage of the total occurrences of any given verb to appear with a given SF before concluding that random error is not responsible for these observations. Unfortunately, the cutoff percentage is arbitrary and sampling error makes classification unstable for verbs with few occurrences in the input. The sampling error can be dealt with ([1]) but the arbitrary cutoff percentage can't.[2] Rather than using fixed cutoffs, we are developing an approach that will automatically generate statistical models of the sources of noise using standard regression techniques. For example, purposive adjuncts like "Jon quit to pursue a career in finance" are quite rare, accounting for only two percent of the apparent infinitival complements. Furthermore, they are distributed across a much larger set of matrix verbs than the true infinitival complements, so any given verb occurs very rarely indeed with purposive adjuncts. In a histogram sorting verbs by their apparent frequency of occurrence with infinitival complements, those that in fact have appeared with purposive adjuncts and not true infinitival complements will be clustered at the low frequencies. The distributions of such clusters can be modeled automatically and the models used for identifying false positives.

RELATED WORK

Interest in extracting lexical and especially collocational information from text has risen dramatically in the last two years, as sufficiently large corpora and sufficiently cheap computation have become available. Three recent papers in this area are [3], [8], and [12]. The latter two are concerned exclusively with collocation relations between open-class words and not with grammatical properties. Church is also interested primarily in open-class collocations, but he does discuss verbs that tend to be followed by infinitives within his mutual information framework.

Mutual information, as applied by Church, is a measure of the tendency of two items to appear near one-another — their observed frequency in nearby positions is divided by the expectation of that frequency if their positions were random and independent. As Church points out, having such statistics for word-pairs is useful for the predictive models used in optical character-recognition and speech recognition as well as for syntactic disambiguation. To measures the tendency of verbs to be followed within a few words by infinitives, Church uses his statistical disambiguator ([4]) to distinguish between *to* as an infinitive marker and *to* as a preposition. Then he measures the mutual information between occurrences of the verb and occurrences of infinitives following within a certain number of words. Unlike our system, Church's approach does not aim to decide whether or not a verb occurs with an infinitival complement — example (1) showed that being followed by an infinitive is not the same as taking an infinitival complement. It might be interesting to try building a verb categorization scheme based on Church's mutual information measure, but to the best of our knowledge no such work has been reported.

[2] Note that this is not an arbitrary confidence level, which would be less unsavory, but an actual percentage of verb occurrences. That is, there is a fact of the matter — a natural clustering, but no systematic characterization of it is available, so an eyeball estimate must be used instead.

CONCLUSIONS

The initial results reported above are only the beginning of what promises to be a be large and rewarding endeavor. In a forthcoming paper Brent reports on acquisition of subcategorization frames using raw, untagged text. Running on raw text, the program starts with only the grammar and a lexicon of some 200 closed-class words. This opens up the possibility of learning from literally hundreds of millions of words of text without worrying the possible major categories of all the words or their relative frequencies.

Along with implementing detection schemes for more SFs, our next major goal will be noise-reduction. If that is successful we hope to release to the community a substantial dictionary of verbs and subcategorization frames. We also hope to use the SF information for semantic categorization [6] using lexical-syntax/lexical-semantics constraints [10, 11]. A particularly clear example of how this can be done is provided by the verbs taking DO&clause with a non-pleonastic subject: all such verbs can describe communication [13]. The complete list of DO&clause verbs our program program found more than once, running in raw text mode on 2.6 million words of Wall Street Journal, supports Zwicky's observation (3).

(1) advise, assure, convince, inform, reassure, remind, tell, warn

ACKNOWLEDGMENTS

Thanks to Don Hindle, Leila Gleitman, and Jane Grimshaw for useful and encouraging conversations. Thanks also to Mark Liberman and the Penn Treebank project at the University of Pennsylvania for supplying tagged text. This work was supported in part by National Science Foundation grant DCR-85552543 under a Presidential Young Investigator Award to Professor Robert C. Berwick.

REFERENCES

1. M. Brent. Semantic Classification of Verbs from their Syntactic Contexts: An Implemented Case Study of Stativity.

In *Proceedings of the 5th European ACL Conference*. Association for Computational Linguistics, 1991.

2. E. Brill, D. Magerman, M. Marcus, B. Santorini. Deducing Linguistic Structure from the Statistics of Large Corpora. *Proceedings of the 3rd DARPA Speech and Natural Language Workshop*, 1990.

3. K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comp. Ling.*, 16, 1990.

4. K. Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the 2nd ACL Conference on Applied NLP*. ACL, 1988.

5. C. DeMarcken. Parsing the LOB Corpus. In *Proceedings of the ACL*. Association for Comp. Ling., 1990.

6. L. Gleitman. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–56, 1990.

7. D. Hindle. User Manual for Fidditch, a Deterministic Parser. Technical Report 7590-142, Naval Research Laboratory, 1983.

8. D. Hindle. Noun classification from predicate argument structures. In *Proceedings of the 28th Annual Meeting of the ACL*, pages 268–275. ACL, 1990.

9. A. Hornby and A. Covey. *Oxford Advanced Learner's Dictionary of Contemporary English*. Oxford University Press, Oxford, 1973.

10. B. Levin. English Verbal Diathesis. Lexicon Project working Papers no. 32, MIT Center for Cognitive Science, MIT, Cambridge, MA., 1989.

11. S. Pinker. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA, 1989.

12. F. Smadja and K. McKeown. Automatically extracting and representing collocations for language generation. In *28th Annual Meeting of the Association for Comp. Ling.*, pages 252–259. ACL, 1990.

13. A. Zwicky. In a Manner of Speaking. *Linguistic Inquiry*, 2:223–233, 1970.