Robust Extraction of Subcategorization Data from Spoken Language

Jianguo Li & Chris Brew
Department of Linguistics
The Ohio State University, USA
{jianguo|cbrew}@ling.ohio-state.edu

Eric Fosler-Lussier
Department of Computer Science & Engineering
The Ohio State University, USA
fosler@cse.ohio-state.edu

1 Introduction

Subcategorization data has been crucial for various NLP tasks. Current method for automatic SCF acquisition usually proceeds in two steps: first, generate all SCF cues from a corpus using a parser, and then filter out spurious SCF cues with statistical tests. Previous studies on SCF acquisition have worked mainly with written texts; spoken corpora have received little attention. Transcripts of spoken language pose two challenges absent in written texts: uncertainty about utterance segmentation and disfluency.

Roland & Jurafsky (1998) suggest that there are substantial subcategorization differences between spoken and written corpora. For example, spoken corpora tend to have fewer passive sentences but many more zero-anaphora structures than written corpora. In light of such subcategorization differences, we believe that an SCF set built from spoken language may, if of acceptable quality, be of particular value to NLP tasks involving syntactic analysis of spoken language.

2 SCF Acquisition System

Following the design proposed by Briscoe and Carroll (1997), we built an SCF acquisition system consisting of the following four components: Charniak's parser (Charniak, 2000); an SCF extractor; a lemmatizer; and an SCF evaluator. The first three components are responsible for generating SCF cues from the training corpora and the last component, consisting of the Binomial Hypothesis Test (Brent, 1993) and a back-off algorithm (Sarkar & Zeman, 2000), is used to filter SCF cues on the basis of their reliability and likelihood.

We evaluated our system on a million word written corpus and a comparable spoken corpus from BNC. For type precision and recall, we used 14 verbs selected by Briscoe & Carroll (1997) and evaluated our results against SCF entries in COMLEX (Grishman *et al.*, 1994). We also calculated token recall and the results are summarized in the following table.

Corpus	Written	Spoken
type precision	93.1%	91.2%
type recall	48.2%	46.4%
token recall	82.3%	80%

Table 1: Type precision, recall and token recall

3 Detecting Incorrect SCF Cues

We examined the way segmentation errors and disfluency affects our acquisition system – the statistical parser and the extractor in particular – in proposing SCF cues and explored ways to detect incorrect SCF cues. We extracted 500 SCF cues from the ViC corpus (Pitt, *et al*, 2005) and identified four major reasons that seem to have caused the extractor to propose incorrect SCF cues: multiple utterances; missing punctuation; disfluency; parsing errors.

Error analysis reveals that segmentation errors and disfluencies cause the parser and the extractor to tend to make systematic errors in proposing SCF cues – incorrect SCF cues are likely to have an extra complement. We therefore proposed the following two sets of linguistic heuristics for automatically detecting incorrect SCF cues:

Linguistic Heuristic Set 1: The following SCF cues are extremely unlikely whatever the verb. Reject an SCF cue as incorrect if it contains the following patterns:

- [(NP) PP **NP**]: We reach out [to your friends] [**your neighbor**].
- [NP PP-to **S**]: Would I want them to say [that][to me] [**would I want them to do that to me**].
- [NP NP **S**]: They just beat [Indiana in basketball] [the- Saturday] [**I think it was um-hum**].

Proceedings of the Ninth International Workshop on Parsing Technologies (IWPT), pages 204–205,
Vancouver, October 2005. ©2005 Association for Computational Linguistics

- [**PP-p** PP-p]: He starts living [**with the**] [with the guys].

Linguistic Heuristic Set 2: The following SCF cues are all possibly valid SCFs: for SCF cues of the following type, check if the given verb takes it in COMLEX. If not, reject it:

- [(NP) **S**]: When he was dying [**what did he say**].
- [PP-to **S**]: The same thing happened [to him] [**uh he had a scholarship**].
- [(NP) **NP**]: OU had a heck of time beating [them] [**uh-hum**].
- [(NP) **INF**]: You take [the plate] from the table [**rinse them off**] and put them by the sink.

Given the utilization of a gold standard in the heuristics, it would be improper to build an end-to-end system and evaluate against COMLEX. Instead, we evaluate by seeing how often our heuristics succeed producing results agreeable to a human judge.

To evaluate the robustness of our linguistic heuristics, we conducted a cross-corpora and cross-parser comparison. We used 1,169 verb tokens from the ViC corpus and another 1,169 from the Switchboard corpus.

Cross-corpus Comparison: The purpose of the cross-corpus comparison is to show that our linguistic heuristics based on the data from one spoken corpus can be applied to other spoken corpora. Therefore, we applied our heuristics to the ViC and the Switchboard corpus parsed by Charniak's parser. We calculated the percentage of incorrect SCF cues before and after applying our linguistic heuristics. The results are shown in Table 2.

Charniak's parser	ViC	Switchboard
before heuristics	18.8%	9.5%
after heuristics	6.4%	4.6%

Table 2: Incorrect SCF cue rate before and after heuristics

Table 2 shows that the incorrect SCF cue rate has been reduced to roughly the same level for the two spoken corpora after applying our linguistic heuristics.

Cross-parser Comparison: The purpose of the cross-parser comparison is to show that our linguistic heuristics based on the data parsed by one parser can be applied to other parsers as well. To this end, we applied our heuristics to the Switchboard corpus parsed by both Charniak's parser and Bikel's parsing engine (Bikel, 2004). Again, we calculated the percentage of incorrect SCF cues before and after applying our heuristics. The results are displayed in Table 3.

Although our linguistic heuristics works slightly better for data parsed by Charniak' parser, the incorrect SCF cue rate after applying heuristics remains at about the same level for the two different parsers we used.

Switchboard	Charniak	Bikel
before heuristics	9.5%	9.2%
after heuristics	4.6%	5.4%

Table 3: Incorrect SCF cue rate before and after heuristics

4 Conclusion

We showed that it should not be assumed that standard statistical parsers will fail on language that is very different from what they are trained on. Specifically, the results of Experiment 1 showed that it is feasible to apply current SCF extraction technology to spoken language. Experiment 2 showed that incorrect SCF cues due to segmentation errors and disfluency can be recognized by our linguistic heuristics. We have shown that our SCF acquisition system as a whole will work for the different demands of spoken language.

5 Acknowledgements

This work was supported by NSF grant 0347799 to the second author, and by a summer fellowship from the Ohio State Center for Cognitive Science to the first author.

References

Biekl, D. 2004. Intricacies of Collins' Parsing Model. *Computational Linguistics*, 30(4): 470-511

Brent, M. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. Computational Lingusitics: 19(3): 243-262

Briscoe, E. & Carroll, G. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC. 356-363

Chaniak, E. 2000. A Maximum-Entropy-Inspired Parser. In Proceedings of the 2000 Conference of the North American Chapter of ACL. 132-139

Grishman, R., Macleod, C. & Meyers, A. 1994. COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of the International Conference on Computational Lingusitics*, COLING-94, Kyoto, Japan. 268-272

Pitt, M., Johnson, K., Hume, E., Kiesling, S., Raymond, W. 2005. They Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. *Speech Communication*, 45: 89-95

Roland, D. & Jurafsky, D. 1998. How Verb Subcategorization Frequency Affected by the Corpus Choice. In *Proceedings of 17th International Conference on Computational Lingusitics*, 2: 1122-1128

Sarkar, A. & Zeman, D. 2000. Automatic Extraction of Subcategorization Frames for Czech. In Proceedings of the 19th International Conference on Computational Lingusitics. 691-697