Automatic Processing of Large Corpora for the Resolution of Anaphora References

Ido Dagan * Alon Itai
Computer Science Department
Technion, Haifa, Israel
dagan@techunix.bitnet, itai@cs.technion.ac.il

Abstract

Manual acquisition of semantic constraints in broad domains is very expensive. This paper presents an automatic scheme for collecting statistics on cooccurrence patterns in a large corpus. To a large extent, these statistics reflect semantic constraints and thus are used to disambiguate anaphora references and syntactic ambiguities. The scheme was implemented by gathering statistics on the output of other linguistic tools. An experiment was performed to resolve references of the pronoun "it" in sentences that were randomly selected from the corpus. The results of the experiment show that in most of the cases the cooccurrence statistics indeed reflect the semantic constraints and thus provide a basis for a useful disambiguation tool.

1 Introduction

The use of selectional constraints is one of the most popular methods in applying semantic information to the resolution of ambiguities in natural languages. The constraints typically specify which combinations of semantic classes are acceptable in subject-verb-object relationships and other syntactic structures. This information is used to filter out some analyses of ambiguous constructs or to set preferences between alternatives.

Though the use of selectional constraints is very popular, there is very little success (if any) in implementing this method for broad domains. The major problem is the huge amount of information that must be acquired in order to achieve a reasonable representation of a large domain. In order to overcome this problem, our project suggests an alternative to the traditional model, based on automatic acquisition of constraints from a large corpus. The rest of the paper describes how this method is used to resolve anaphora references. Similarly, the constraints are used also to resolve syntactic ambiguities, but this will not be described here. The reader should bare in mind that like the conventional use of selectional constraints, our method is intended to work in conjunction with other disambiguation means. These, such as various syntactic and pragmatic constraints and heuristics [Carbonell and Brown 1988, Hobbs 1978], represent additional levels of knowledge and are essential when selectional constraints are not sufficient.

*Part of this research was conducted while visiting IBM T. J. Watson Research Center, Yorktown Heights, NY

2 The Statistical Approach

According to the statistical model, cooccurrence patterns that were observed in the corpus are used as selection patterns. Whenever several alternatives are presented by an ambiguous construct, we prefer the one corresponding to more frequent patterns.

When using selectional constraints for anaphora resolution, the referent must satisfy the constraints which are imposed on the anaphor. If the anaphor participates in a certain syntactic relation, like being an object of some verb, then the substitution of the anaphor with the referent must satisfy the selectional constraints. In the statistical model, we substitute each of the candidates with the anaphor and approve only those candidates which produce frequent cooccurrence patterns. Consider, for example, the following sentence, taken from the Hansard corpus of the proceedings of the Canadian parliament [Brown et al. 1988]:

(1) They know full well that the companies held tax money aside for collection later on the basis that the government said it was going to collect it.

There are two occurrences of "it" in this sentence. The first serves as the subject of "collect" and the second as its object. We gathered the statistics for three candidates which occur in the sentence: "collection", "money" and "government". According to the syntactic structure of the sentence, each of them may serve as the referent for each of the occurrences of the pronoun. The following table lists the patterns that were produced by substituting each can-

didate with the anaphor, and the number of times each of these patterns occurred in the corpus:

subject-verb	collection	collect	0
subject-verb	money	collect	5
subject-verb	government	collect	198
verb-object	collect	collection	0
verb-object	collect	money	149
verb-object	collect	government	0

According to these statistics "government" is preferred as the referent of the first "it", and "money" of the second.

This example demonstrates the case of definite semantic constraints which eliminate all but the correct alternative. In other cases, several alternatives may satisfy the selectional constraints, and may be observed in the corpus a significant number of times. In such cases the final selection between the approved candidates should be performed by other means, such as syntactic heuristics or asking the user. Another possibility may be to use statistical preferences, and prefer the relatively more frequent patterns. However, at this stage it is not clear to us how useful the statistical preference can be, and we use the statistics only relative to a certain threshold, approving any patterns that pass this threshold.

3 Implementing the Acquisition Phase

The use of the statistical model involves two separate phases. The first is the acquisition phase, in which the corpus is processed and the statistical database is built. The second is the disambiguation phase, in which the statistical database is used to resolve ambiguities.

The statistical database contains cooccurrence patterns for various syntactic relations. In the experiment reported here we have used constraints for the "subject-verb", "verb-object" and "adjective-noun" relations. To locate these relations in the sentences of the corpus, each sentence is parsed by the PEG parser [Jensen 1986]. Then, a post-processing algorithm identifies the various relations in the parse tree. As was noted in [Grishman et al. 1986], the cooccurrence patterns reflect regularized or canonical structure. Therefore the post-processing algorithm has to map surface structures into the normalized relations. During our experiments we have used two different implementations for this algorithm [Lappin et al. 1988] [Jensen 1989], which take into account structures like passives, subclauses, questions and relative and infinitive clauses.

The use of an automatic procedure for extracting information from a corpus that was not preprocessed manually raises a basic problem of circularity. Since the corpus was not disambiguated, it is not possible to distinguish the semantically correct patterns from the incorrect ones. Both types of ambiguity, syntactic and lexical, may cause the system to acquire or use inappropriate patterns. This problems is considered very important when dealing with a corpus: it was the reason for the substantial human intervention in the procedure of [Grishman et al. 1986], and it is the reason why other techniques use manually tagged corpora (e.g. [Church 1988]).

In practice, however, we have discovered that the problem is not so crucial: semantically valid patterns have occurred many more times in syntactically unambiguous constructs than in ambiguous ones. Thus, they could be identified without the need of first disambiguating the sentences. Semantically non-valid patterns indeed occurred in the inappropriate parses but they were too rare to pass the threshold. As for lexical ambiguities, the chance that one sense of a word will be confused with another during disambiguation seems to be very small, and it never happened in our experiment.

4 The Experiment

An experiment was performed to resolve references of the anaphor "it" in the Hansard corpus. The examples of the ambiguous sentences were selected in the following way: First, sentences containing the word "it" were extracted randomly from the corpus. Then, we manually filtered out sentences that were not relevant for the use of selectional constraints in resolving anaphoric references. Such cases were non-anaphoric occurrences of "it", cases where the referent was not a noun phrase and cases where the anaphor was not involved in one of the three relations that we used. In addition, we have excluded cases where there was only one possible referent, so that our results will reflect correctly the performance of the disambiguation method. The filtering process eliminated about two thirds of the original sentences, and we proceeded with 59 examples. The alternative candidates for the referent (which satisfy definite syntactic constraints such as number, gender and requirements for reflexives) were identified manually in each example. [1]

The statistics were collected from part of the corpus, of about 28 million words. For 21 out of the 59 examples the statistics were not meaningful (we used a threshold of 5 occurrences for each of the alternative patterns). In these cases the algorithm cannot approve any of the candidates, getting a "coverage" of 38/59 (64%).

As explained in Section 2, the output of the statistical method is used to represent the selectional

[1] The Hansard corpus, as maintained by the speech group at IBM Watson Research Center, does not contain consecutive sentences. Therefore, we identified only candidates within the same sentence as the anaphor. To provide enough candidates, we examined occurrences of "it" after the 15th word of the sentence. The examples provided between 2 to 5 candidates, with an average of 2.8 candidates per anaphor.

constraints. This is done by approving all patterns which appeared a significant number of times. Therefore, the output is considered correct if the appropriate candidate is approved. This happened in 33 cases, getting "accuracy" of 33/38 (87%). In 18 of these cases, the appropriate candidate was the only one which was approved, getting a complete resolution of the ambiguity.

This last result demonstrates the advantage of the statistical data over semantic constraints. While semantic constraints should approve any combination of arguments in a syntactic relation that *may* occur in the text, the statistics approve only those combinations that *actually* occur and reject others. Manual observation of the 18 sentences in which the statistics completely resolved the ambiguity showed that only in 7 cases the ambiguity could be eliminated by traditional selectional constraints. This is consistent with the evaluation in [Hobbs 1978], where only in 12 out of 132 sentences the ambiguity was eliminated by selectional constraints.

An additional note should be made concerning the technical methodology of the experiment. Within the limited resources of our research, it was not feasible to build the statistical database for the entire Hansard corpus, which contains about 60 million words. The expensive resources are the parsing time and the storage for the cooccurrence patterns and their statistics. [2] However, it turns out that parsing the entire corpus is not necessary to evaluate the success of the statistical model! As the evaluation relates to a limited number of examples, it is sufficient to collect the statistics only for patterns that are relevant for the disambiguation of these examples. Therefore, we have extracted from the corpus only those sentences that contained at least one cooccurrence of words from a relevant pattern. This procedure allowed us to parse only 10,000 sentences.

5 Conclusions

We have suggested using cooccurrence patterns, automatically acquired from a large corpus, as an alternative to selectional constraints. The initial results indicate that even in its basic form, as presented here, the approach is useful for disambiguation, and many times performs even better than the traditional model. This should be considered relative to the effort that would have been required to achieve such coverage and accuracy by manual acquisition of constraints, for the broad domain of parliament proceedings.

In a general perspective, this project promotes the use of a large corpus for linguistic research and applications. Processing such large corpora is a non-trivial engineering problem, the solution of which enables research to focus on complicated real world sentences. Our research demonstrates how statistical methods can be built on top of more 'traditional' linguistic tools, achieving a better and more feasible environment for the resolution of ambiguities.

6 Acknowledgements

We would like to thank Mori Rimon, Shalom Lappin, Wlodek Zadrozny, Slava Katz, John Justeson, Lisa Braden-Harder and Peter Brown for their fruitful advice and technical support.

7 References

[Brown et al. 1988] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R.L. and Roossin P.S., A statistical approach to language translation, *COLING 1988.*

[Carbonell and Brown 1988] Carbonell, J. G. and Brown, R. D. Anaphora resolution: A multi strategy approach, *COLING 1988.*

[Church 1988] Church, K. W., A stochastic parts program and noun phrase parser for unrestricted text, *ACL Conf. on Applied NLP,* 1988.

[Grishman et al. 1986] R. Grishman, L. Hirschman and Ngo Thanh Nhan, Discovery procedures for sublanguage selectional patterns initial experiments, *Computational Linguistics,* vol. 12, 205-214, 1986.

[Hobbs 1978] Hobbs, J. R. Resolving pronoun references, *Lingua,* vol. 44, 311-338, 1978.

[Jensen 1986] K. Jensen, PEG 1986: A broad-coverage computational syntax of English, *Technical Report, IBM T. J. Watson Research Center,* 1986.

[Jensen 1989] Jensen, K. PEGASOS: Deriving predicate-argument structures after a syntactic parse. Presented at the *International Workshop on Parsing Technologies,* Carnegie Mellon University, August 1989.

[Lappin et al. 1989] S. Lappin, I. Golan, M. Rimon, Computing grammatical functions from a configurational parse tree, *Technical Report 88.268, IBM Israel Center of Science and Technology,* 1989.

[2] Although the construction of the full size database is not feasible for us, it is clearly feasible for a large scale project. This is shown by a similar database that was implemented as part of the language model of the IBM speech recognition system. This database contains counters for occurrences of sequences of three words in large corpora (trigrams), which are much more numerous than our syntactic patterns.