COMBINING RULE-BASED AND DATA-DRIVEN TECHNIQUES FOR GRAMMATICAL RELATION EXTRACTION IN SPOKEN LANGUAGE

Kenji Sagae and Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{sagae, alavie}@cs.cmu.edu

Abstract

We investigate an aspect of the relationship between parsing and corpus-based methods in NLP that has received relatively little attention: coverage augmentation in rule-based parsers. In the specific task of determining grammatical relations (such as subjects and objects) in transcribed spoken language, we show that a combination of rule-based and corpus-based approaches, where a rule-based system is used as the teacher (or an automatic data annotator) to a corpus-based system, outperforms either system in isolation.

1 Introduction

Corpus-based methods in natural language processing have advanced rapidly in the past decade. Their relevance to parsing and natural language analysis is vast, including lexical and structural disambiguation, and even purely data-driven parsers. In this paper we investigate an aspect of the relationship between parsing and corpus-based methods in NLP that has received relatively little attention: coverage augmentation in rule-based parsers. While probabilistic grammars have been widely used for disambiguation in rule-based parsers, it is less common to find data-driven methods designed to remedy the lack of grammatical coverage of a system. Although coverage issues are largely inexistent in modern treebank-trained statistical parsers, rule-based parsers driven by hand-written grammars are still widely used in a variety of applications, as they often provide for deeper linguistic analysis (i.e. in the form of detailed feature-structures) and are also easier to tune for high levels of accuracy on the data for which they were developed. Obtaining very broad coverage with such grammars is, however, a well-known problem (Sagae, Lavie & MacWhinney, 2001; Black, Lafferty & Roukos, 1992).

The specific task in our experiments is the identification of grammatical relations (GRs), such as subjects, objects and adjuncts, in transcripts of conversational language. The data used in our experiments were taken from the CHILDES database (MacWhinney, 2000) and consists of utterances spoken by parents to their children. Certain characteristics of this task make it particularly suitable for the illustration of the issues in discussion. First, a large corpus of domain-specific data annotated with grammatical relations is not available for training a data-driven system. Second, analysis of spontaneous conversational language is known to be a challenging area for rule-based parsers in terms of coverage. Our approach involves a high-precision rule-based parser, which is used as a

teacher for a simple system comprised of data-driven NLP tools. The combination of the knowledge encompassed in the grammar-driven system and domain-specific unlabelled data allows us to train a corpus-based system which, although less accurate than the original rule-based parser, does not suffer from the brittleness associated with hand-written grammars. We show that even a very simple combination of the two systems results in precision and recall of grammatical relations that are superior to those of either system in isolation.

2 Identifying Grammatical Relations

Because precision and recall of constituent bracketing are often used as parser evaluation metrics, it is common to envision a description of the syntactic constituent structure of sentences as the output of a parser. However, different kinds of parsers analyze sentences in different ways and with different purposes, and a diagram representing constituent structures is often not the most appropriate type of output. Carroll, Briscoe and Sanfillipo (1998) propose that precision and recall of grammatical relations be used for parser evaluation, and describe some advantages of using grammatical relations over other evaluation metrics. Our use of GRs is motivated by the crucial role of such information in the measurement of syntactic complexity in the field of child language research, through schemes such as IPSyn (Scarborough, 1990) and LARSP (Fletcher & Garman, 1988).

2.1 Using a Rule-Based Parser for Grammatical Relations

The starting point of our experiments was the system described in (Sagae, Lavie & MacWhinney, 2001) for syntactic analysis of data from the CHILDES database. The main part of this system is composed of a robust rule-based parser called LCFlex (Rosé & Lavie, 2001), and a handcrafted domain-specific unification grammar. LCFlex's robustness comes from strategies designed specifically for parsing spoken language, such as word skipping and limited constituent insertion.

The output of this rule-based system is a syntactic feature structure corresponding to the input sentence. Extracting GRs from the feature structure produced by LCFlex is simple: there is a grammatical relation between the head word of each sub-structure and the head word of the outer structure containing the sub-structure in question. Each grammatical relation is named after the syntactic function of the sub-structure in relation to its outer structure. This process is illustrated in figure 1.

A central challenge in this work was the trade-off between the grammar's coverage of the corpus, and the accuracy of the analyses produced by the parser. The well-known problem of grammatical coverage found in high precision rule-based parsers is accentuated in spoken language, because of the common deviations from traditional rules of grammar found in casual verbal interactions. In addition to the conversational characteristics commonly found in sentences in casual spoken language (vocatives and communicators, elided subjects, false starts), the language found in the CHILDES database also features the absence of auxiliaries in places where their use is clearly intended.

Providing coverage for casual language with conversational features requires that a grammar should handle syntactic constructions found in "standard" language, as well as the many variations found mostly in spoken language. By increasing grammar coverage with the addition of rules, more ambiguity is introduced, causing the search for the correct analysis to be more difficult. This trade-off between recall and precision was addressed in the rule-based system with a constraint relaxation approach, supported by robustness features of the parser, and implemented as a multi-pass parsing strategy. The first passes were designed to provide the least amount of ambiguity, while sacrificing coverage. When these failed, coverage was increased with the introduction of robust parsing methods that permitted word skipping, word insertion, and part-of-speech ambiguity. At each pass, only the sentences for which the system fails to find an analysis are identified and sent for reparsing in further passes. Each pass thus provides for an increasing amount of coverage (at the expense of accuracy). The system attempts to parse a sentence until either an analysis is found, or it decides that the likelihood of finding a correct analysis by further constraint relaxation is too low to consider, in which case it reports a failed parse.

Using a strict evaluation metric for complete feature structure matches, the rule-based system achieved 78.5% accuracy (or 90.2% accuracy, if only sentences for which an analysis was reported are considered), according to the original evaluation in (Sagae, Lavie & MacWhinney, 2001). Of these complete feature structure matches, 36.5% (or 21% of all sentences) were obtained only after

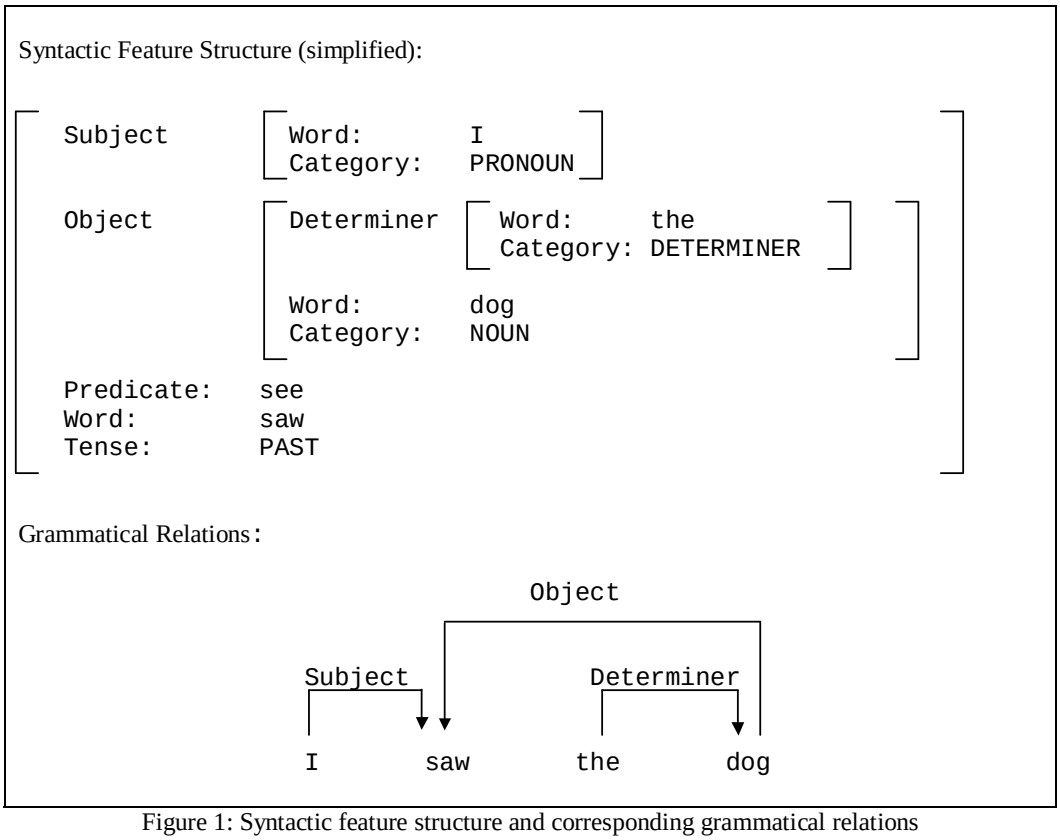

Figure 1: Syntactic feature structure and corresponding grammatical relations

the introduction of one or more robustness techniques for handling spoken language. Because the system was evaluated on complete matches only, these results cannot be mapped directly into precision and recall figures of grammatical relations. However, because about 12% of the test sentences received no analysis due to parse failure, it is clear that recall of grammatical relations was sub-optimal, in spite of high precision.

2.2 A Simple Data-Driven System for Grammatical Relations

As an alternative to the rule-based parser that suffers from coverage limitations that can result in complete parsing failure, we designed a simple and robust data-driven method for extracting GRs without the need of manually annotated training material. A key observation is that while the rule-based parser can occasionally fail to parse, the analyses found when parsing succeeds are of high precision. We can therefore use the rule-based parser in order to automatically create a large volume of labeled training data for a data-driven approach. We first parse a large corpus of in-domain data using the rule-based parser. Text that is successfully parsed by the rule-based parser is then used to create "labeled" examples for training the data-driven approach. One obvious weakness of this idea is that we are training the data-driven approach *solely* on material that can already be parsed by the rule-based parser, and it is thus questionable whether the resulting trained data-driven parser can learn how to correctly parse data that the original rule-based parser was unable to parse. Our conjecture was that for the task of extracting GRs, we could in fact develop a data-driven approach that extends the coverage of the rule-based parser using the above approach.

A simple data-driven system for assigning GRs to words was built as follows. We trained an off-the-shelf part-of-speech tagger with part-of-speech/GR-label pairs, instead of the usual word/part-of-speech-tag pairs. The GR labels associated with the words in the training data were extracted from the analyses generated by the rule-based parser. Because valuable information contained in the words themselves is lost in the assignment of GR tags to part-of-speech tags alone, the output of this tagger is further refined by an error-driven transformation-based learning strategy that takes the actual words into account, implemented using the fnTBL toolkit (Ngai & Florian, 2001).

While the above GR "tagger" can assign grammatical relation labels to words, we still have to determine the target of the directional link established by the grammatical relation indicated by the label. To accomplish this, the raw text input sentence was also parsed using a statistical parser (Charniak, 2000) trained on the Penn Treebank (Marcus et al., 1993), yielding an approximation of the skeletal constituent structure (parse tree) of the sentence. A slightly modified version of the "treebank constituent head table" originally designed by Magerman (1995) is then used to determine the heads of constituents in the parse tree. By stipulating that there is a directional link from every word in a constituent (except for the head) to the head of the constituent, and applying that notion to the entire parse-tree, we determine a set of unlabeled dependency links for the sentence. The

combination of the unlabeled links and the GR labels results in our target output of grammatical relations.

2.3 Identifying Grammatical Relations with Rule-Based and Data-Driven Methods

Once we have established a corpus-based procedure that makes GR assignments to every input sentence, regardless of whether or not it can be parsed by the rule-based system, we can attempt to determine how much a system trained on the output of rule-based parser can improve the precision and recall of GRs obtained with the rule-based system alone. Although several ways of combining the outputs of the rule-based and data-driven systems can be imagined, based on the strengths and weaknesses of each system in identifying specific GRs, we combined the two systems in a very simple way: the output of the data-driven system is used when parsing with the rule-based system fails. This should serve as a lower bound on the possible improvements in recall and f-measure. It is also worth noting that we make no claim that the corpus-based system we used in our experiments provides the best results we could achieve training on the output of the LCFlex-based system. It is, rather, a simple combination of readily available off-the-shelf NLP tools, and serves to illustrate how corpus-based techniques may be used to improve the performance of a rule-based system. A well developed statistical system for identifying GRs would surely perform better, and the design of such a system is planned as future work.

3 Results and Discussion

To evaluate the systems described in section 2, we took an unused portion of the CHILDES database consisting of 505 words (118 sentences) as a test set, hand-labeled it with four grammatical relations (subject, object, adjunct, and predicate nominal), and obtained GR assignments for this test set using three different setups: running the rule-based system alone, the data-driven system alone, and the combination of the two systems. The number of instances of each GR in the test set is shown in table 1. Each setup was evaluated on precision and recall of the four GRs. The results can be seen in tables 2, 3, and 4, respectively (F-score is the harmonic mean of precision and recall).

While the rule-based system does achieve reasonably high precision in the recognition of these grammatical relations, its overall F-score (harmonic mean of precision and recall) is somewhat low

Grammatical Relation	Number of instances in test set
Subject	76
Object	50
Adjunct	51
Predicate nominal	16

Table 1: Number of instances of each GR in the test set

Grammatical Relation	Precision	Recall	F-score
Subject	0.93	0.68	0.79
Object	0.78	0.56	0.65
Adjunct	0.77	0.75	0.76
Predicate nominal	0.91	0.67	0.77

Table 2: Results using only the rule-based system.

Grammatical Relation	Precision	Recall	F-score
Subject	0.75	0.74	0.74
Object	0.67	0.64	0.65
Adjunct	0.69	0.43	0.53
Predicate nominal	0.24	0.33	0.28

Table 3: Results using the statistical system

Grammatical Relation	Precision	Recall	F-score
Subject	0.84	0.84	0.84
Object	0.78	0.72	0.75
Adjunct	0.77	0.80	0.79
Predicate nominal	0.77	0.67	0.71

Table 4: Results using the rule-based/data-driven combination

due to the lack of recall caused by parse failures, where no grammatical relation information is generated. Conversely, while the simple corpus-based system has better recall measures on two of the four GRs tested, its precision is lower than the rule-based system's on all four GRs. Of the 118 sentences in the test set, the rule-based system failed on 20, or 16.9%. Table 5 shows the number of instances of each GR in the 20 test sentences for which the rule-based parser failed to report an analysis. The performance of the data-driven system on those 20 sentences can be seen in table 6.

The combination of the two systems produced improved recall (and F-scores) on three out of the four GRs tested. The exception was the predicate nominal relation, with which the corpus-based system clearly had problems, most likely due to the relatively lower frequency of that relation. Note that while there are only three instances of that relation in the 20 sentences where the output of the data-driven parser is used, overall precision drops significantly from what the rule-based system produces. This is due to the data-driven system erroneously finding a number of instances of the predicate nominal relation. The use of a validation set could determine if a situation such as this occurs, so that the sharp decline in precision (which caused the decline in F-score) is avoided. It remains to be seen how the use of a larger amount of unlabelled data, a more comprehensive manually annotated test set, and a development set would affect the overall performance of the

Grammatical Relation	Number of instances in test set
Subject	20
Object	14
Adjunct	8
Predicate nominal	3

Table 5: Number of instances of each GR in failed sentences

Grammatical Relation	Precision	Recall	F-score
Subject	0.60	0.60	0.60
Object	0.80	0.57	0.67
Adjunct	0.75	0.37	0.50
Predicate nominal	0.00	0.00	0.00

Table 6: Results using the data-driven system on failed sentences

combined system. However, the results as they stand already show that this combination of the rule-based and data-driven systems outperforms either system in isolation.

4 Related Work

Carroll and Briscoe (2002) present a wide-coverage parser that outputs grammatical relations, and discuss the trade-off between precision and recall of grammatical relations, as well as useful ways to manipulate such trade-off to achieve high precision at the expense of recall. This trade-off is also observed in our experiments. However, our angle on this issue focuses on the combination of systems with different precision/recall behavior, to achieve a higher combined F-score.

Blaheta and Charniak (2000) discuss the assignment of Penn Treebank (Marcus et al., 1993) function tags to constituent structure trees. They use a statistical approach to assign tags (similar in many ways to grammatical relations) to parse tree nodes. Our corpus-based model also uses parse trees, but only to determine that a GR exists between two words. The work of Gildea and Palmer (2002) has shown that the use of constituent structure information is useful in determining predicate-argument structure. While their work involved propositions of a more semantic nature, we believe their results to be applicable to the identification of grammatical relations.

5 Conclusions and Future Work

We have presented a way to combine rule-based and data-driven NLP techniques in the extraction of grammatical relations. We have shown that starting with a rule-based system, we can use unlabeled data and a corpus-based system to improve recall (and F-score) of grammatical relations. While the

experiment presented included a sub-optimal corpus-based system and only a very simple combination scheme, the results were conclusively positive. As future work, we plan to develop a statistical model of GRs, including both links and labels, expand our systems to recognize a wider range of GRs, and explore different ways of combining results from multiple systems.

6 Acknowledgements

We would like to thank Brian MacWhinney for insightful discussions about automatic processing of transcribed child-parent interactions that led to this work.

This research work was supported in part by the National Institutes of Health under grant number 2R01HD23998.

References

Black, E., Lafferty, J., & Roukos, S. (1992). Development and Evaluation of a Broad-Coverage Probabilistic Grammar of English-Language Computer Manuals. In *Proceedings of the Association for Computational Linguistics* (pp. 185-192). Newark, DE.

Blaheta, D., & Charniak, E. (2000). Assigning function tags to parsed text. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 234-240). Seattle, WA.

Carroll, J., & Briscoe, E. (2002). High precision extraction of grammatical relations. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)* (pp. 134-140). Taipei.

Carroll, J., Briscoe, E., & Sanfillipo, J. (1998). Parser evaluation: A survey and a new proposal. In *Proceedings of the First International Conference on Language Resources and Evaluation* (pp. 447-454). Granada: LREC.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics* Seattle, WA.

Fletcher, P., & Garman, M. (1988). LARSPing by numbers. *British Journal of Disorders of Communication, 23*, 309-321.

Gildea, D., & Palmer, M. (2002). The necessity of parsing for predicate argument recognition. In *Proceedings of the Association for Computational Linguistics*. Philadelphia, PA.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Magerman, D. (1995). Statistical decision-tree models for parsing. In *Proceedings of the Association for Computational Linguistics*.

Marcus, M. P., Santorini, B., & Marcinkiewics, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*.

Ngai, G., & Florian, R. (2001). In *Proceedings of the 2nd Meeting of the North Amercian Chapter of the Association for Computational Linguistics* (pp. 40-47). Pittsburgh, PA.

Rosé, C. P., & Lavie, A. (2001). Balancing robustness and efficiency in unification-augmented context-free parsers for large practical applications. In A. van Noord & A. Junqua (Eds.), *Robustness in language and speech technology*. Amsterdam: Kluwer.

Sagae, K., Lavie, A., & MacWhinney, B. (2001). Parsing the CHILDES database: Methodology and lessons learned. In *Proceedings of the Seventh Inernational Workshop in Parsing Technologies*. Beijing, China.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1-22.