Few Inductive Models for Grammatical Relations Recovery

Vasile Rus and **Kirtan Desai**
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38120
email: vrus@memphis.edu

Abstract

This paper describes our experience with few models for recovering the grammatical relations in isolated English sentences. Our method starts with a set of linguistically motivated features and then uses machine learning methods to induce a classifier able to identify relations such as logical subject, direct object, etc. Our experiments show that our set of features is powerful and can deliver promising results.

Introduction

Grammatical relations are vital for advanced text understanding technologies such as information extraction, machine translation, question answering and others. Despite that Treebank II (Marcus, Santorini, & Marcinkiewicz 1993), a collection of text - corpus - manually annotated with syntactic information by experts, includes functional tags, which mark grammatical relations, in its annotation tags, modern automated parsing technologies generated from it only offer surface syntactic information in the form of a bracketed representation in which main constituents and major structural phrases in a sentence are identified, ignoring the functional tags.

To overcome the drawback of modern parsing technology to identify the underlying grammatical relations of English sentences, novel methods are necessary that offer accurate, robust and scalable solutions to the problem of finding syntactic functional information.

In this work a set of features is introduced which is then used to induce automated tools able to detect functional information in English sentences. The tools are obtained using the C4.5 (Quinlan 1996) package for decision tree induction.

Related Work

When syntactic information is needed to study a certain linguistic problem, people either use the bracketed form generated by state of the art parser and are happy with its surface level syntactic information or have their own pattern-based methods which lack generality and scalability (Stetina, Kurohashi, & Nagao 1998), (Lapata 1999), (Stevenson & Merlo 1999).

The Features

Our approach is to address the grammatical relations recovery task as a classification problem: given a verb in a sentence and a candidate phrasal head find the most appropriate grammatical relations (roles) the head plays for a verb. The set of possible roles contains: subject, direct object, indirect object, prepositional object or norole (a value which indicates that the candidate head does not play any role for the given verb). To preview our results, we demonstrate that combining a set of indicators automatically extracted from large text corpora provide good performance.

The key to any automatic classification task is to determine a set of useful features for discriminating the items to be classified. Observing the patterns of logic syntactic roles for verbs we derived the following features for our classification task: **Head**, **Lexical Category of Head**, **Voice**, **Type of Clause**, **Position of Head**, **Head-Verb Dependency**. Those features could be automatically extracted from a large corpus, either manually annotated or automatically generated.

Experimental Setup

There are three major issues that we need to address before performing any experiments: what verbs to focus on, where should we gather training data from and what machine learning algorithm(s) to use. In the next few paragraphs we provide answers for each of those issues.

Previous work on verb meaning research, such as (Korhonen, Gorrell, & McCarthy 2000) and (Ted & Carroll 1997), reported experiments on a set of 14 target verbs that exhibit multiple argument patterns: *ask, begin, believe cause, expect, find, give, help, like, move, produce, provide, seem, swing.* We adopted those 14 verbs since we believed it would be a good starting point to have a small set, on one hand, with many argument ambiguities, on the other hand, thus balancing challenges with manageability of the experiments.

Next, we looked for a corpus. Treebank (Marcus, Santorini, & Marcinkiewicz 1993) is a good candidate because it contains role annotations. We

Verb	Training Size	Errors(%)	Estimate Errors(%)	Errors before Pruning
all	14586	1.5	1.7	1.7
all+no-head	14586	1.6	1.7	1.7
all	14586	7.6	8.2	8.2
all+no-head	14586	7.1	7.7	7.7

Table 1: Errors when traces are solved and the dependency feature is added to the feature set.

started by developing patterns for tgrep, a tree retrieval pattern-based tool, to identify sentences containing target verbs from Wall Street Journal (WSJ) corpus (the version with part-of-speech tags) and used the online form to retrieve the data (http://www.ldc.upenn.edu/ldc/online/treebank/).
The training set is further processed: a stemmer is applied to obtain the stem of individual words and then the target verb is identified and the features extracted. One or more training examples (positive and negative) are generated from a sentence.

As learning paradigm we chose for decision trees. Here, we predict the accuracy of our induced classifiers using *10-fold cross validation*.

We present results for two major experiments: (1) using our set of features as a standard model and (2) use the dependency feature as a filter instead of being part of the model. They focus only on logical subjects, thus the task is to determine whether or not a given word is a logica subject or not.

We present results for two major experiments: (1) using our set of features as a standard model and (2) use the dependency feature as a filter instead of being part of the model. They focus only on logical subjects, thus the task is to determine whether or not a given word is a logica subject or not.

Table 1 presents a summary of the results. The upper half is about Experiment 1 and the lower about Experiment 2. The line having *all* in the verb column reports results when training examples of all target verbs were considered together in a single experiment, say *all-verb*. The last line in the table shows results when the head feature is ignored. There is a small increase in the error rate (from 1.6% to 1.7%) but a simpler, less-lexicalized model is obtained.

The last line in the table shows results when the lexical information about the word that plays the role is ignored. There is only a marginal increase in the error rate and it seems that in the absence of deeper semantic information about the word itself (for example the semantic class used to specify selectional restrictions) there is not much impact on the basic model by the lexical head feature although its value of being the head of a phrase is extremely important.

An alternative way to use the *head-verb dependency* feature is as a filter. In Experiment 2 we keep our base model (similar to the one in Experiment 1) but generate a smaller training set by filtering out training examples for words that are not directly related to the verb. Bottom half of Table 1 illustrates the new training set sizes and errors for the target verbs. We notice that the training sets are significantly smaller and that the errors range from 18.7% to 6.0%. A smaller training set leads to smaller decision trees which may be an advantage.

Conclusions

The results reported in this work form an upper bound of the performance of our model since the tagging and parsing are accurate (tag or parse errors are present, though at a very low rate, in Treebank).

The models presented yield high performance and they can form reliable components in larger text understanding systems such as logic form idenfication, automated textual inference engines, intelligent tutoring and others.

References

Korhonen, A.; Gorrell, G.; and McCarthy, D. 2000. Statistical filtering and subcategorization frame acquisition.

Lapata, M. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Meeting of the North American Chapter of the Association*, 397–404.

Marcus, M.; Santorini, B.; and Marcinkiewicz. 1993. Building a large annotated coprus of english: the penn treebank. *Computational Linguistic* 19(2):313–330.

Quinlan, R. 1996. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research* vol 4:77–90.

Stetina, J.; Kurohashi, S.; and Nagao, M. 1998. General word sense disambiguation method based on A full sentential context. In *Use of WordNet in Natural Language Processing Systems: Proceedings of the Workshop*. Somerset, New Jersey: Association for Computational Linguistics. 1–8.

Stevenson, S., and Merlo, P. 1999. Automatic verb classification using distributions of grammatical features.

Ted, B., and Carroll, J. 1997. Automatic extraction of subcategorization from corpora.