Transplanting Supertags from English to Spanish

Srinivas Bangalore
AT&T Labs–Research
180 Park Avenue
Florham Park, NJ09732
srini@research.att.com

Abstract

In this paper, we present an approach to quickly develop supertags for a target language given supertags for another language (reference language), along with a sentence-aligned parallel corpus between reference language and target language pairs. Our method can be interpreted as composing the alignment relation with dependency relation of the reference sentence to obtain the dependency relation for the target sentence. This dependency relation is then used to induce the supertags for the target words.

1 Introduction

Supertags localize lexical and structural ambiguity by associating rich and complex descriptions to words of a language. This localization allows us to compute lexical and contextual distributional properties of supertags. In earlier work (JS94; Sri97a; Sri97b) we have shown that this distributional information can be used in a novel way to perform *almost parsing*. Trained on a million words of correctly supertagged Wall Street Journal Text, a simple trigram based supertagger assigns the same supertags to 92% of the words as they would have been assigned in the intended parse of a sentence. In subsequent work we have demonstrated the utility of supertags in a variety of applications including, Language Modeling (Sri96), Information Filtering (CS97b; CS97c), Information Extraction (DNB+97) and Sentence Simplification (CS97a).

2 An issue in Supertagging approach

However, constructing a rich repertoire of supertags for a language is a time consuming and tedious task as exemplified by the history of development of the English XTAG Grammar (XTA95) at University of Pennsylvania and the French XTAG Grammar at University of Paris.[1] In this paper, our attempt is to provide a solution to alleviate the task of building a supertag collection for a language (*target language*) based on the set of supertags of another language (*reference language*). In particular, we present a method of transplanting the set of supertags from the XTAG Grammar for English to Spanish using a parallel corpus of sentence-aligned English-Spanish sentences.

3 Grammar Induction vs Grammar Transplantation

Previous proposals (Res92; Sch92) for learning LTAG grammars involved inducing elementary trees from unannotated corpora. However, these proposals require training of a large number of parameters on even larger collections of corpora and yet the resulting structures may not be linguistically motivated. In contrast, our approach is based on the premise that elementary trees of natural language grammars are related and that these structures can be inherited *almost* as is, from the reference language to the target language. We use the term *grammar transplantation* as opposed to *grammar induction* in order to differentiate the amount effort involved in the development of supertags for the target language. However, a limitation of our approach is that the target language is imposed with structures that closely resemble the source language structure.

4 Methodology

Our approach to transplanting supertags involves applying the following steps to each sen-

[1] But this should not be regarded as a limitation exclusively of the supertag-based parsing paradigm. Treebank-based statistical parsing methods are limited by the effort involved in constructing a treebank.

tence pair in the reference-target parallel corpus. We have applied this method to an English-Spanish ATIS corpus.

- We first obtain a word alignment for each sentence pair using the alignment algorithm described in (ABD98). The alignment algorithm is completely unsupervised and only requires a sentence aligned corpus in two languages. It uses a correlation metric among reference-target word-pairs as a cost of reference-target word pairing and performs an alignment search that minimizes the sum of the costs of a set of pairings which map the reference sentence to its target sentence.
- The words of the English sentence are supertagged using a supertagger. The supertagger used for the ATIS domain was trained on 2000 word-supertag pairs and performs at 92% accuracy on a 500 word test set.
- The supertagged English sentence is further annotated with dependency links using the Lightweight Dependency Analyzer described in (Sri97b).
- The dependency links are then migrated to the target sentence as follows: if words w_i and w_j are linked in the reference sentence, w_i is aligned with v_p and w_j is aligned with v_q, then a dependency link is posited between v_p and v_q.
- Finally, the dependency structure migrated on to the target sentence is used to recover the correct ordering of arguments of each word. This information is used to construct the supertag for the word.

Our method can be interpreted as composing the alignment relation with dependency relation of the reference sentence to obtain the dependency relation for the target sentence. This dependency relation is then used to induce the supertags for the target words.

5 Example

Consider the following pair of sentences from the sentence-aligned English-Spanish ATIS corpus.

English: SHOW BUSINESS CLASS FARES ON U S AIR FROM BOSTON TO TORONTO

Spanish: MUESTRE LAS TARIFAS EN CLASE DE NEGOCIOS EN U S AIR DE BOSTON A TORONTO

The result of the alignment algorithm is shown below. Notice that the result contains alignments between one word in the source string (FARES) to two words in the target string (LAS:TARIFAS). Multi-word alignments are shown separated by a ":". The alignment algorithm allows mapping between at most two words in the source string to two words in the target string.

English: SHOW BUSINESS CLASS FARES ON U S AIR FROM BOSTON TO TORONTO

Spanish: MUESTRE LAS:TARIFAS EN CLASE DE NEGOCIOS EN U S AIR DE BOSTON A TORONTO

Target Position	Source Position
1	1
2 3	4
4	
5	3
6	
7	2
8	5
9	6
10	7
11	8
12	9
13	10
14	11
15	12

The output of the supertagger for the English string is in Table 1. The supertagger assigns to each word the part-of-speech and supertag information. The supertag information is used to assign dependency information among the words of the sentence.

The POS, supertags and dependency links are transplanted on to the target string using the

Position	Words	POS	Supertag	Dependency links
1	SHOW	VB	A_Inx0Vnx1	4.
2	BUSINESS	NN	B_Nn	3*
3	CLASS	NN	B_Nn	4*
4	FARES	NNS	A_NXN	
5	ON	IN	B_nxPnx	4* 8.
6	U	NNP	B_Nn	7*
7	S	NNP	B_Nn	8*
8	AIR	NNP	A_NXN	
9	FROM	IN	B_nxPnx	8* 10.
10	BOSTON	NNP	A_NXN	
11	TO	IN	B_nxPnx	8* 12.
12	TORONTO	NNP	A_NXN	

Table 1: Result of applying the supertagger and the LDA on the English string

Position	Words	POS	Supertag	Dependency links
1	MUESTRE	NN	A_Inx0Vnx1	2:3.
2:3	LAS:TARIFAS	NNS	A_NXN	
4	EN			
5	CLASE	NN	B_Nn	2:3*
6	DE			
7	NEGOCIOS	NN	B_Nn	4*
8	EN	IN	B_nxPnx	2:3* 11.
9	U	NNP	B_Nn	10*
10	S	NNP	B_Nn	11*
11	AIR	NNP	A_NXN	
12	DE	IN	B_nxPnx	11* 13.
13	BOSTON	NNP	A_NXN	
14	A	TO	B_nxPnx	11* 15.
15	TORONTO	NNP	A_NXN	

Table 2: Result of combining the alignment information with the dependency information

alignment information and the result is in Table 2.

The target string dependency structure is examined for *completeness* and *consistency*. Completeness requires that each word is assigned a supertag and its dependency requirements are satisfied. Consistency requires that the direction of the head/dependent of a given word matches the direction of its dependency requirement.

In our example, the words at positions 4 and 6 are not assigned any supertags and hence violate completeness constraint and the words at positions 5 and 7 violate consistency constraints since the supertag (B_Nn) requires the head to appear to its right while the head appears on the left.

We solve the consistency and completeness problems by assigning to a word the most frequent supertag it is associated with, given the entire corpus, which can fit into the dependency context of the target string and at the same time respect the dependency constraints imposed by the source language. The corrected POS, supertag and dependency structure for the target string is shown in Table 3.

6 Evaluation

The system can be evaluated in a number of ways: in the context of an application, in terms of the supertags assigned, in terms of the dependency links assigned or in terms of time reduced in developing a full-fledged domain independent grammar. We are in the process of evaluating the system on its performance in assigning

Position	Words	POS	Supertag	Dependency links
1	MUESTRE	NN	A_Inx0Vnx1	2:3.
2:3	LAS:TARIFAS	NNS	A_NXN	
4	EN	IN	B_nxPnx	2:3* 5.
5	CLASE	NN	A_NXN	
6	DE	IN	B_nxPnx	5* 7.
7	NEGOCIOS	NN	A_NXN	
8	EN	IN	B_nxPnx	2:3* 11.
9	U	NNP	B_Nn	10*
10	S	NNP	B_Nn	11*
11	AIR	NNP	A_NXN	
12	DE	IN	B_nxPnx	11* 13.
13	BOSTON	NNP	A_NXN	
14	A	TO	B_nxPnx	11* 15.
15	TORONTO	NNP	A_NXN	

Table 3: Result of correcting the dependency structure based on completeness and consistency constraints.

supertags and dependency links to 1000 words of annotated test corpus from the ATIS domain. Preliminary results suggest that the performance in assigning supertags is about 80% accurate.

References

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Canada, 1998.

R. Chandrasekar and B. Srinivas. Automatic induction of rules for text simplification. *Knowledge-based Systems*, 10:183–190, 1997.

R. Chandrasekar and B. Srinivas. Gleaning information from the web: Using syntax to filter out irrelevant information. In *Proceedings of AAAI 1997 Spring Symposium on NLP on the World Wide Web*, 1997.

R. Chandrasekar and B. Srinivas. Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging. In *Proceedings of RIAO'97*, Montreal, June 1997.

Christine Doran, Michael Niv, Breckenridge Baldwin, Jeffrey Reynar, and B. Srinivas. Mother of Perl: A Multi-tier Pattern Description Language. In *Proceedings of the International Workshop on Lexically Driven Information Extraction*, Frascati, Italy, July 1997.

Aravind K. Joshi and B. Srinivas. Disambiguation of Super Parts of Speech (or Supertags): Almost Parsing. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan, August 1994.

Philip Resnik. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, Nantes, France, July 1992.

Yves Schabes. Stochastic lexicalized tree-adjoining grammars. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, Nantes, France, July 1992.

B. Srinivas. "Almost Parsing" Technique for Language Modeling. In *Proceedings of ICSLP96 Conference*, Philadelphia, USA, 1996.

B. Srinivas. *Complexity of Lexical Descriptions and its Relevance to Partial Parsing.* PhD thesis, University of Pennsylvania, Philadelphia, PA, August 1997.

B. Srinivas. Performance Evaluation of Supertagging for Partial Parsing. In *Proceedings of Fifth International Workshop on Parsing Technology*, Boston, USA, September 1997.

The XTAG-Group. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 95-03, University of Pennsylvania, 1995. Updated version available at http://www.cis.upenn.edu/ xtag/tr/techreport.html.