Automatic Adaptation of Annotation Standards for Dependency Parsing — Using Projected Treebank as Source Corpus

Wenbin Jiang and Qun Liu

Key Lab. of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{jiangwenbin, liuqun}@ict.ac.cn

Abstract

We describe for dependency parsing an annotation adaptation strategy, which can automatically transfer the knowledge from a *source corpus* with a different annotation standard to the desired *target parser*, with the supervision by a *target corpus* annotated in the desired standard. Furthermore, instead of a hand-annotated one, a projected treebank derived from a bilingual corpus is used as the source corpus. This benefits the resource-scarce languages which haven't different hand-annotated treebanks. Experiments show that the target parser gains significant improvement over the baseline parser trained on the target corpus only, when the target corpus is smaller.

1 Introduction

Automatic annotation adaptation for sequence labeling (Jiang et al., 2009) aims to enhance a tagger with one annotation standard by transferring knowledge from a source corpus annotated in another standard. It would be valuable to adapt this strategy to parsing, since for some languages there are also several treebanks with different annotation standards, such as Chomskian-style Penn Treebank (Marcus et al., 1993) and HPSG LinGo Redwoods Treebank (Oepen et al., 2002) for English. However, we are not content with conducting annotation adaptation between existing different treebanks, because it would be more valuable to boost the parsers also for the resource-scarce languages, rather than only for the resource-rich ones that already have several treebanks.

Although hand-annotated treebanks are costly and scarce, it is not difficult for many languages to collect large numbers of bilingual sentence-pairs aligned to English. According to the word alignment, the English parses can be projected across to their translations, and the projected trees can be leveraged to boost parsing. Many efforts are devoted to the research on projected treebanks, such as (Lü et al., 2002), (Hwa et al., 2005) and (Ganchev et al., 2009), etc. Considering the fact that a projected treebank partially inherits the English annotation standard, some hand-written rules are designed to deal with the divergence between languages such as in (Hwa et al., 2002). However, it will be more valuable and interesting to adapt this divergence automatically and boost the existing parsers with this projected treebank.

In this paper, we investigate the automatic annotation adaptation strategy for Chinese dependency parsing, where the source corpus for adaptation is a projected treebank derived from a bilingual corpus aligned to English with word alignment and English trees. We also propose a novel, error-tolerant tree-projecting algorithm, which dynamically searches the project Chinese tree that has the largest consistency with the corresponding English tree, according to an alignment matrix rather than a single alignment. Experiments show that when the target corpus is smaller, the projected Chinese treebank, although with inevitable noise caused by non-literal translation and word alignment error, can be successfully utilized and result in significant improvement over the baseline model trained on the target corpus only.

In the rest of the paper, we first present the tree-projecting algorithm (section 2), and then the annotation adaptation strategy (section 3). After discussing the related work (section 4) we show the experiments (section 5).

2 Error-Tolerant Tree-Projecting Algorithm

Previous works making use of projected corpus usually adopt the direct-mapping method for structure projection (Yarowsky and Ngai, 2001; Hwa et al., 2005; Ganchev et al., 2009), where

some filtering is needed to eliminate the inaccurate or conflicting labels or dependency edges. Here we propose a more robust algorithm for dependency tree projection. According to the alignment matrix, this algorithm dynamically searches the projected Chinese dependency tree which has the largest consistency with the corresponding English tree.

We briefly introduce the alignment matrix before describing our projecting algorithm. Given a Chinese sentence $C_{1:M}$ and its English translation $E_{1:N}$, the alignment matrix A is an $M \times N$ matrix with each element $A_{i,j}$ denoting the probability of Chinese word C_i aligned to English word E_j. Such structure potentially encodes many more possible alignments.

Using $\mathcal{C}(T_C|T_E, A)$ to denote the degree of Chinese tree T_C being consistent with English tree T_E according to alignment matrix A, the projecting algorithm aims to find

$$\hat{T_C} = \underset{T_C}{\operatorname{argmax}}\, \mathcal{C}(T_C|T_E, A) \quad (1)$$

$\mathcal{C}(T_C|T_E, A)$ can be factorized into each dependency edge $x \rightarrow y$ in T_C, that is to say

$$\mathcal{C}(T_C|T_E, A) = \prod_{x \rightarrow y \in T_C} \mathcal{C}_e(x \rightarrow y|T_E, A) \quad (2)$$

We can obtain $\mathcal{C}_e$ by simple accumulation across all possible alignments

$$\mathcal{C}_e(x \rightarrow y|T_E, A) = \sum_{1 \le x', y' \le |E|} A_{x,x'} \times A_{y,y'} \times \delta(x', y'|T_E) \quad (3)$$

where $\delta(x', y'|T_E)$ is a 0-1 function that equals 1 only if $x' \rightarrow y'$ exists in T_E.

The searching procedure, argmax operation in equation 1, can be effectively solved by a simple, bottom-up dynamic algorithm with cube-pruning speed-up (Huang and Chiang, 2005). We omit the detailed algorithm here due to space restrictions.

3 Annotation Adaptation for Dependency Parsing

The automatic annotation adaptation strategy for sequence labeling (Jiang et al., 2009) aims to strengthen a tagger trained on a corpus annotated in one annotation standard with a larger assistant corpus annotated in another standard. We can define the purpose of the automatic annotation adaptation for dependency parsing in the same way.

Similar to that in sequence labeling, the training corpus with the desired annotation standard is called the *target corpus* while the assistant corpus annotated in a different standard is called the *source corpus*. For training, an intermediate parser, called the *source parser*, is trained directly on the source corpus and then used to parse the target corpus. After that a second parser, called the *target parser*, is trained on the target corpus with guide features extracted from the source parser's parsing results. For testing, a token sequence is first parsed by the source parser to obtain an intermediate parsing result with the source annotation standard, and then parsed by the target parser with the guide features extracted from the intermediate parsing result to obtain the final result.

The design of the guide features is the most important, and is specific to the parsing algorithm of the target parser. In this work we adopt the maximum spanning tree (MST) algorithm (McDonald et al., 2005; McDonald and Pereira, 2006) for both the source and the target parser, so the guide features should be defined on dependency edges in accordance with the edge-factored property of MST models. In the decoding procedure of the target parser, the degree of a dependency edge being supported can be adjusted by the relationship between this edge's head and modifier in the intermediate parsing result of the source parser. The most intuitionistic relationship is whether the dependency between head and modifier exists in this intermediate result. Such a bi-valued relationship is similar to that in the stacking method for combining dependency parsers (Martins et al., 2008; Nivre and McDonald, 2008). The guide features are then defined as this relationship itself as well as its combinations with the lexical features of MST models.

Furthermore, in order to explore more detailed knowledge from the source parser, we redefine the relationship as a four-valued variable which covers the following situations: *parent-child*, *child-parent*, *siblings* and *else*. With the guide features, the parameter tuning procedure of the target parser will automatically learn the regularity of using the source parser's intermediate result to guide its decision making.

4 Related Works

Many works have been devoted to obtain parsing knowledge from word aligned bilingual cor-

pora. (Lü et al., 2002) learns Chinese bracketing knowledge via ITG alignment; (Hwa et al., 2005) and (Ganchev et al., 2009) induces dependency grammar via projection from aligned English, where some filtering is used to reduce the noise and some hand-designed rules to handle language heterogeneity.

Just recently, Smith and Eisner (2009) gave an idea similar to ours. They perform dependency projection and annotation adaptation with Quasi-Synchronous Grammar (QG) Features. Although both related to projection and annotation, there are still important differences between these two works. First, we design an error-tolerant alignment-matrix-based tree-projecting algorithm to perform whole-tree projection, while they resort to QG features to score local configurations of aligned source and target trees. Second, their adaptation emphasizes to transform a tree from one annotation standard to another, while our adaptation emphasizes to strengthen the parser using a treebank annotated in a different standard.

5 Experiments

The source corpus for annotation adaptation, that is, the projected Chinese treebank, is derived from 5.6 millions LDC Chinese-English sentence pairs. The Chinese side of the bilingual corpus is word-segmented and POS-tagged by an implementation of (Jiang et al., 2008), and the English sentences are parsed by an implementation of (McDonald and Pereira, 2006) which is instead trained on WSJ section of Penn English Treebank (Marcus et al., 1993). The alignment matrixes for sentence pairs are obtained according to (Liu et al., 2009). The English trees are then projected across to Chinese using the algorithm in section 2. Out of these projected trees, we only select 500 thousands with word count l s.t. $6 \leq l \leq 100$ and with projecting confidence $c = \mathcal{C}(T_C|T_E, A)^{1/l}$ s.t. $c \geq 0.35$. While for the target corpus, we take Penn Chinese Treebank (CTB) 1.0 and CTB 5.0 (Xue et al., 2005) respectively, and follow the traditional corpus splitting: chapters 271-300 for testing, chapters 301-325 for development, and else for training.

We adopt the 2nd-order MST model (McDonald et al., 2005) as the target parser for better performance, and the 1st-order MST model as the source parser for fast training. Both the two parsers are trained with averaged perceptron algorithm (Collins, 2002). The development set of CTB is also used to determine the best model for the source parser, conditioned on the hypothesis of larger isomorphisme between Chinese and English.

Model	P% on CTB 1	P% on CTB 5
source parser	53.28	53.28
target parser	**83.56**	**87.34**
baseline parser	82.23	87.15

Table 1: Performances of annotation adaptation with CTB 1.0 and CTB 5.0 as the target corpus respectively, as well as of the baseline parsers (2nd-order MST parsers trained on the target corpora).

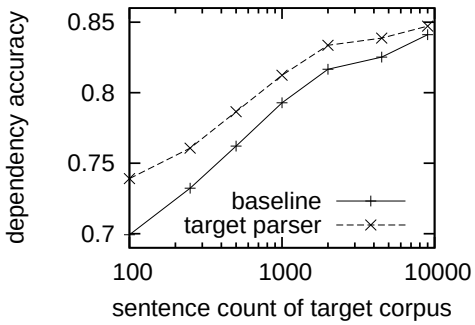

Figure 1: Performance of the target parsers with target corpora of different scales.

Table 1 shows that the experimental results of annotation adaptation, with CTB 1.0 and CTB 5.0 as the target corpus respectively. We can see that the source parsers, directly trained on the source corpora of projected trees, performs poorly on both CTB test sets (which are in fact the same). This is partly due to the noise in the projected treebank, and partly due to the heterogeneous between the CTB trees and the projected trees. On the contrary, automatic annotation adaptation effectively transfers the knowledge to the target parsers, achieving improvement on both target corpora. Especially on CTB 1.0, an accuracy increment of 1.3 points is obtained over the baseline parser.

We observe that for the much larger CTB 5.0, the performance of annotation adaptation is much lower. To further investigate the adaptation performances with target corpora of different scales, we conduct annotation adaptation on a series of target corpora which consist of different amount of dependency trees from CTB 5.0. Curves in Figure 1 shows the experimental results. We see that the smaller the training corpus is, the more significant improvement can be obtained. For example,

with a target corpus composed of 2K trees, nearly 2 points of accuracy increment is achieved. This is a good news to the resource-scarce languages.

6 Conclusion and Future Works

This paper describes for dependency parsing an automatic annotation adaptation strategy. What is more important, we use a projected treebank, rather than a hand-annotated one, as the source corpus for adaptation. This is quite different from previous works on projected trees (Hwa et al., 2005; Ganchev et al., 2009), and is also more valuable than previous works of annotation adaptation (Jiang et al., 2009). Experiments show that this strategy gains improvement over baseline parsers with target corpora of different scales, especially the smaller ones. This provides a new strategy for resource-scarce languages to train high-precision dependency parsers. In the future, we will adapt this strategy to constituent parsing, which is more challenging and interesting due to the complexity of projection between constituent trees, and due to the obscurity of annotation adaptation for constituent parsing.

Acknowledgement

This project was supported by National Natural Science Foundation of China, Contracts 60603095 and 60736014, and 863 State Key Project No. 2006AA010108. We are grateful to the anonymous reviewers for their valuable suggestions. We also thank Yang Liu for sharing his codes of alignment matrix generation, and Liang Huang and Haitao Mi for helpful discussions.

References

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the EMNLP*, pages 1–8, Philadelphia, USA.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the 47th ACL*.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proceedings of the IWPT*, pages 53–64.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the ACL*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. In *Natural Language Engineering*, volume 11, pages 311–325.

Wenbin Jiang, Liang Huang, Yajuan Lü, and Qun Liu. 2008. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the ACL*.

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic adaptation of annotation standards: Chinese word segmentation and pos tagging–a case study. In *Proceedings of the 47th ACL*.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation. In *Proceedings of the EMNLP*.

Yajuan Lü, Sheng Li, Tiejun Zhao, and Muyun Yang. 2002. Learning chinese bracketing knowledge based on a bilingual language model. In *Proceedings of the COLING*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. In *Computational Linguistics*.

André F. T. Martins, Dipanjan Das, Noah A. Smith, and Eric P. Xing. 2008. Stacking dependency parsers. In *Proceedings of EMNLP*.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of EACL*, pages 81–88.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of ACL*, pages 91–98.

Joakim Nivre and Ryan McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL*.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning Dan Flickinger, and Thorsten Brants. 2002. The lingo redwoods treebank: Motivation and preliminary applications. In *In Proceedings of COLING*.

David Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of EMNLP*.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. In *Natural Language Engineering*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the NAACL*.