Relaxed Cross-lingual Projection of Constituent Syntax

Wenbin Jiang and **Qun Liu** and **Yajuan Lü**

Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
{jiangwenbin, liuqun, lvyajuan}@ict.ac.cn

Abstract

We propose a relaxed correspondence assumption for cross-lingual projection of constituent syntax, which allows a supposed constituent of the target sentence to correspond to an unrestricted treelet in the source parse. Such a relaxed assumption fundamentally tolerates the syntactic non-isomorphism between languages, and enables us to learn the target-language-specific syntactic idiosyncrasy rather than a strained grammar directly projected from the source language syntax. Based on this assumption, a novel constituency projection method is also proposed in order to induce a projected constituent treebank from the source-parsed bilingual corpus. Experiments show that, the parser trained on the projected treebank dramatically outperforms previous projected and unsupervised parsers.

1 Introduction

For languages with treebanks, supervised models give the state-of-the-art performance in dependency parsing (McDonald and Pereira, 2006; Nivre et al., 2006; Koo and Collins, 2010; Martins et al., 2010) and constituent parsing (Collins, 2003; Charniak and Johnson, 2005; Petrov et al., 2006). To break the restriction of the treebank scale, lots of works have been devoted to the unsupervised methods (Klein and Manning, 2004; Bod, 2006; Seginer, 2007; Cohen and Smith, 2009) and the semi-supervised methods (Sarkar, 2001; Steedman et al., 2003; McClosky et al., 2006; Koo et al., 2008) to utilize the unannotated text. In recent years, researchers have also conducted many investigations on syntax projection (Hwa et al., 2005; Ganchev et al., 2009; Smith and Eisner, 2009; Jiang et al., 2010), in order to borrow syntactic knowledge from another language.

Different from the bilingual parsing (Smith and Smith, 2004; Burkett and Klein, 2008; Zhao et al., 2009; Huang et al., 2009; Chen et al., 2010) that improves parsing performance with bilingual constraints, and the bilingual grammar induction (Wu, 1997; Kuhn, 2004; Blunsom et al., 2008; Snyder et al., 2009) that induces grammar from parallel text, the syntax projection aims to project the syntactic knowledge from one language to another. This seems especially promising for the languages that have bilingual corpora parallel to resource-rich languages with large treebanks. Previous works mainly focus on dependency projection. The dependency relationship between words in the parsed source sentences can be directly projected across the word alignment to words in the target sentences, following the direct correspondence assumption (DCA) (Hwa et al., 2005). Due to the syntactic non-isomorphism between languages, DCA assumption usually leads to conflicting or incomplete projection. Researchers have to adopt strategies to tackle this problem, such as designing rules to handle language non-isomorphism (Hwa et al., 2005), and resorting to the quasi-synchronous grammar (Smith and Eisner, 2009).

For constituency projection, however, the lack of isomorphism becomes much more serious, since a constituent grammar describes a language in a more detailed way. In this paper we propose a relaxed correspondence assumption (RCA) for constituency

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1192–1201,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

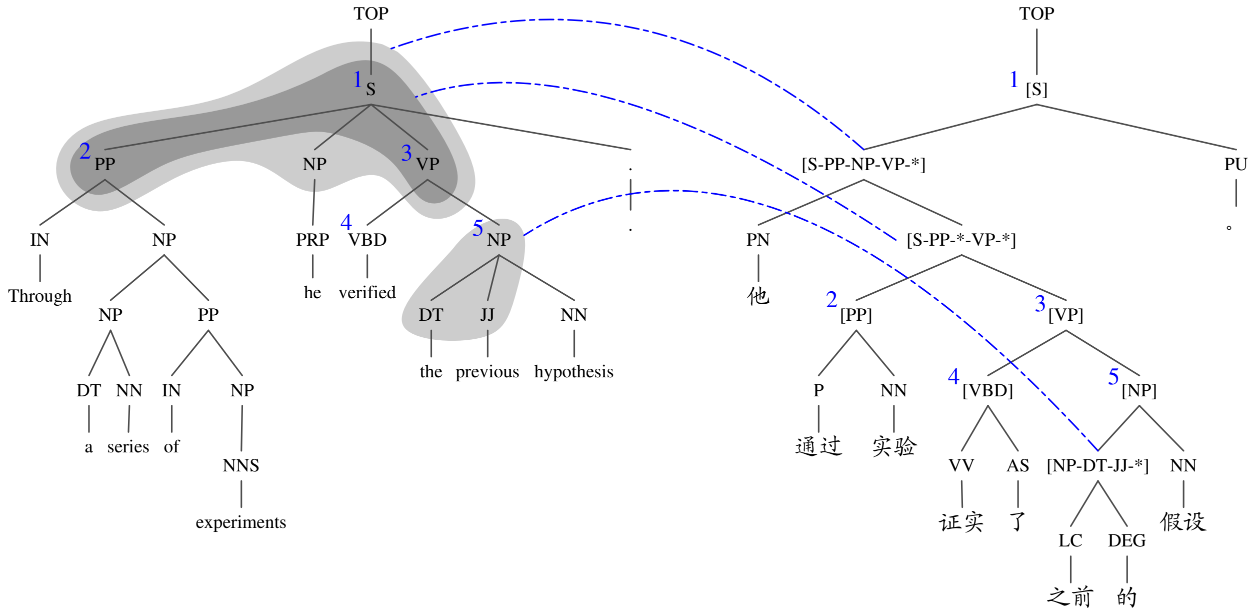

Figure 1: An example for constituency projection based on the RCA assumption. The projection is from English to Chinese. A dash dot line links a projected constituent to its corresponding treelet, which is marked with gray background; An Arabic numeral relates a directly-projected constituent to its counter-part in the source parse.

projection. It allows a supposed constituent of the target sentence to correspond to an unrestricted treelet in the source parse. Such a relaxed assumption fundamentally tolerates the syntactic non-isomorphism between languages, and enables us to learn the target-language-specific syntactic idiosyncrasy, rather than induce a strained grammar directly projected from the source language syntax. We also propose a novel cross-lingual projection method for constituent syntax based on the RCA assumption. Given a word-aligned source-parsed bilingual corpus, a PCFG grammar can be induced for the target language by maximum likelihood estimation on the exhaustive enumeration of candidate projected productions, where each nonterminal in a production is an unrestricted treelet extracted from the source parse. The projected PCFG grammar is then used to parse each target sentence under the guidance of the corresponding source tree, so as to produce an optimized projected constituent tree.

Experiments validate the effectiveness of the RCA assumption and the constituency projection method. We induce a projected Chinese constituent treebank from the FBIS Chinese-English parallel corpus with English sentences parsed by the Charniak parser. The Berkeley Parser trained on the projected treebank dramatically outperforms the previous projected and unsupervised parsers. This provides an promising substitute for unsupervised parsing methods, to the resource-scarce languages that have bilingual corpora parallel to resource-rich languages with human-annotated treebanks.

In the rest of this paper we first presents the RCA assumption, and the algorithm used to determine the corresponding treelet in the source parse for a candidate constituent in the target sentence. Then we describe the induction of the projected PCFG grammar and the projected constituent treebank from the word-aligned source-parsed parallel corpus. After giving experimental results and the comparison with previous unsupervised and projected parsers, we finally conclude our work and point out several aspects to be improved in the future work.

2 Relaxed Correspondence Assumption

The DCA assumption (Hwa et al., 2005) works well in dependency projection. A dependency grammar describes a sentence in a compact manner where the syntactic information is carried by the dependency relationships between pairs of words. It is reasonable to audaciously assume that the relationship of

Algorithm 1 Treelet Extraction Algorithm.

```
1:  Input: $\mathbf{T_f}$: parse tree of source sentence $\mathbf{f}$
2:         $\mathbf{e}$: target sentence
3:         $\mathbf{A}$: word alignment of $\mathbf{e}$ and $\mathbf{f}$
4:  for $i$, $j$ s.t. $1 \leq i < j \leq |\mathbf{e}|$ do        ▷ all spans
5:      $\mathbf{t} \leftarrow$ EXTTREELET($\mathbf{e}, i, j, \mathbf{T_f}, \mathbf{A}$)
6:      $\mathbb{T}_{\langle i,j \rangle} \leftarrow$ PRUNETREE($\mathbf{t}$)
7:  Output: treelet set $\mathbb{T}$ for all spans of $\mathbf{e}$
8:  function EXTTREELET($\mathbf{e}$, $i$, $j$, $\mathbf{T}$, $\mathbf{A}$)
9:      if $\mathbf{T}$ aligns totally outside $\mathbf{e}_{i:j}$ then
10:         return $\emptyset$
11:     if $\mathbf{T}$ aligns totally inside $\mathbf{e}_{i:j}$ then
12:         return $\{\mathbf{T} \cdot root\}$
13:     $\mathbf{t} \leftarrow \{\mathbf{T} \cdot root\}$        ▷ partly aligned inside $\mathbf{e}_{i:j}$
14:     for each subtree $\mathbf{s}$ of $\mathbf{T}$ do
15:         $\mathbf{t} \leftarrow \mathbf{t} \cup$ EXTTREELET($\mathbf{e}, i, j, \mathbf{s}, \mathbf{A}$)
16:     return $\mathbf{t}$
17: function PRUNETREE($\mathbf{T}$)
18:     for each node $n$ in $\mathbf{T}$ do
19:         merge $n$'s successive empty children
20:     $\mathbf{t} \leftarrow \mathbf{T}$
21:     while $\mathbf{t}$ has only one non-empty subtree do
22:         $\mathbf{t} \leftarrow$ the non-empty subtree of $\mathbf{t}$
23:     return $\mathbf{t}$
```

a word pair in the source sentence also holds for the corresponding word pair in the target sentence. Compared with dependency grammar, constituent grammar depicts syntax in a more complex way that gives a sentence a hierarchically branched structure. Therefore the lack of syntactic isomorphism for constituency projection becomes much more serious, it will be hard and inappropriate to directly project the complex constituent structure from one language to another.

For constituency projection, we propose a relaxed corresponding assumption (RCA) to eliminate the influence of syntactic non-isomorphism between the source- and target languages. This assumption allows a supposed constituent of the target sentence to correspond to an unrestricted treelet in the source parse. A treelet is a connected subgraph in the source constituent tree, which covers a discontiguous sequence of words of the source sentence. This property enables a supposed constituent of the target sentence not necessarily to correspond to exactly a constituent of the source parse, so as to fundamentally tolerate the syntactic non-isomorphism between languages. Figure 1 gives an example of relaxed correspondence.

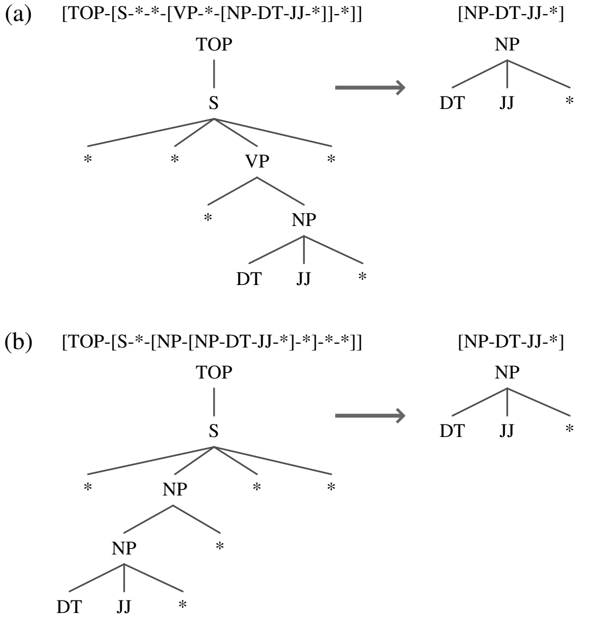

Figure 2: Two examples for treelet pruning. Asterisks indicate eliminated subtrees, which are represented as empty children of their parent nodes.

2.1 Corresponding Treelet Extraction

According to the word alignment between the source and target sentences, we can extract the treelet out of the source parse for any possible constituent span of the target sentence. Algorithm 1 shows the treelet extraction algorithm.

Given the target sentence $\mathbf{e}$, the parse tree $\mathbf{T_f}$ of the source sentence $\mathbf{f}$, and the word alignment $\mathbf{A}$ between $\mathbf{e}$ and $\mathbf{f}$, the algorithm extracts the corresponding treelet out of $\mathbf{T_f}$ for each candidate span of $\mathbf{e}$ (line 4-6). For a given span $\langle i, j \rangle$, its corresponding treelet in $\mathbf{T_f}$ can be extracted by a recursive top-down traversal in the tree. If all nodes in the current subtree $\mathbf{T}$ align outside of source subsequence $\mathbf{e}_{i:j}$, the recursion stops and returns an empty tree $\emptyset$, indicating that the subtree is eliminated from the final treelet (line 9-10). And, if all nodes in $\mathbf{T}$ align inside $\mathbf{e}_{i:j}$, the root of $\mathbf{T}$ is returned as the concise representation of the whole subtree (line 11-12). For the third situation, that is to say $\mathbf{T}$ aligns partly inside $\mathbf{e}_{i:j}$, the recursion has to continue to investigate the subtrees of $\mathbf{T}$ (line 14-15). The recursive traversal finally returns a treelet $\mathbf{t}$ that exactly corresponds to the candidate constituent span $\langle i, j \rangle$ of the source sentence.

We can find that even for a smaller span, the recursive extraction procedure still starts from the root of the source tree. This leads to a expatiatory treelet with some redundant nodes on the top. Function PRUNETREE takes charge of the treelet pruning (line 6). It traverses the treelet to merge the successive empty sibling nodes (marked with asterisks) into one (line 18-19), then conducts a top-down pruning to delete the redundant branches until meeting a branch with more than one non-empty subtrees (line 20-22). Figure 2 shows the effect of the pruning operation with two examples. The pruning operation maps the two original treelets into the same simplified version, that is, the pruned treelet. The branches pruned out of the original treelet serve as the context of the pruned treelet. The bracketed representations of the pruned treelets, as shown above the treelet graphs, are used as the nonterminals of the projected target parses.

Since the overall complexity of the algorithm is $O(|\mathbf{e}|^3)$, it seems inefficient to collect the treelets for all spans in the target sentence. But in fact it runs fast on the realistic corpus in our experiments, we assume that the function EXTTREELET doesn't always consume $O(|\mathbf{e}|)$ because of the more or less isomorphism between two languages.

3 Projected Grammar and Treebank

This section describes how to build a projected constituent treebank based on the RCA assumption. According to the last section, each span of the target sentence could correspond to a treelet in the source parse. If a span $\langle i, j \rangle$ has a corresponding treelet $\mathbf{t}$, a candidate projected constituent can be defined as a triple $\langle i, j, \mathbf{t} \rangle$. For an n-way partition of this span,

$$\langle i, k_1 \rangle, \langle k_1 + 1, k_2 \rangle, .., \langle k_{n-1} + 1, j \rangle$$

if each sub-span $\langle k_{p-1}+1, k_p \rangle$ corresponds to a candidate constituent $\langle k_{p-1}+1, k_p, \mathbf{t}_p \rangle$, a candidate projected production can then be defined, denoted as

$$\langle i, j, \mathbf{t} \rangle \rightarrow \langle i, k_1, \mathbf{t}_1 \rangle \langle k_1+1, k_2, \mathbf{t}_2 \rangle .. \langle k_{n-1}+1, j, \mathbf{t}_n \rangle$$

There may be many candidate projected constituents because of arbitrary combination, the tree projection procedure aims to find the optimum tree from the parse forest determined by these candidate constituents. Each production in the optimum tree should satisfy this principle: the rule used in this production appears in the whole corpus as frequently as possible.

However, due to translation diversity and word alignment error, the real constituent tree of the target sentence may not be contained in the candidate projected constituents. We propose a relaxed and fault-tolerant tree projection strategy to tackle this problem. First, based on the distribution of candidate projected constituents over each single sentence, we estimate the distribution over the whole corpus for the rules used in these constituents, so as to obtain a projected PCFG grammar. Then, using a PCFG parser and this grammar, we parse each target sentence under the guidance of the candidate projected constituent set of the target sentence, so as to obtain the optimum projected tree as far as possible. In the following, we first describe the estimation of the projected PCFG grammar and then show the tree projection procedure.

3.1 Projected PCFG Grammar

From a human-annotated treebank, we can induce a PCFG grammar by estimating the frequency of the production rules, which are contained in the productions of the trees. But for each target sentence we don't know which candidate productions consist the correct constituent tree, so we can't estimate the frequency of the production rules directly.

A reasonable hypothesis is, if a candidate projected production for a target sentence happens to be in the correct parse of the sentence, the rule used in this production will appear frequently in the whole corpus. We assume that each candidate projected production may be a part of the correct parse, but with different probabilities. If we give each candidate projected production an appropriate probability and use this probability as the appearance frequency of this production in the correct parse, we can achieve an approximation of the PCFG grammar hidden in the target sentences. In this work, we restrict the productions to be binarized to reduce the computational complexity. It results in a binarized PCFG grammar, similar to previous unsupervised works.

To estimate the frequencies of the candidate productions in the correct parse of the target sentence, we need first estimate the frequencies of the candidate spans, which are described as follows:

$$p(\langle i,j\rangle|\mathbf{e}) = \frac{\#\ of\ trees\ including\ \langle i,j\rangle}{\#\ of\ all\ trees} \quad (1)$$

The count of all binary trees of a target sentence $\mathbf{e}$ can be calculated similar to the β value calculation in the inside-outside algorithm. Without confusion, we adopt the symbol $\beta(i,j)$ to denote the count of binary tree for span $\langle i,j\rangle$:

$$\beta(i,j) = \begin{cases} 1 & i=j \\ \sum_{k=i}^{j-1} \beta(i,k)\cdot\beta(k+1,j) & i<j \end{cases} \quad (2)$$

$\beta(1,|\mathbf{e}|)$ is the count of binary trees of target sentence $\mathbf{e}$. We also need to calculate the count of binary tree fragments that cover the nodes outside span $\langle i,j\rangle$. This is similar to the calculation of the α value in the inside-outside algorithm. We also adopt the symbol $\alpha(i,j)$ here:

$$\alpha(i,j) = \begin{cases} 1 & i=1, j=|\mathbf{e}| \\ \sum_{k=j+1}^{|\mathbf{e}|} \alpha(i,k)\cdot\beta(k+1,|\mathbf{e}|) & \\ +\sum_{k=1}^{i-1} \alpha(k,j)\cdot\beta(k,j-1) & \textbf{else} \end{cases} \quad (3)$$

For simplicity we omit some conditions in above formulas. The count of trees containing span $\langle i,j\rangle$ is $\alpha(i,j)\cdot\beta(i,j)$. Equation 1 can be rewritten as

$$p(\langle i,j\rangle|\mathbf{e}) = \frac{\alpha(i,j)\cdot\beta(i,j)}{\beta(1,|\mathbf{e}|)} \quad (4)$$

On condition that $\langle i,j\rangle$ is a span in the parse of $\mathbf{e}$, the probability that $\langle i,j\rangle$ has two children $\langle i,k\rangle$ and $\langle k+1,j\rangle$ is

$$p(\langle i,k\rangle\langle k+1,j\rangle|\langle i,j\rangle) = \frac{\beta(i,k)\cdot\beta(k+1,j)}{\beta(i,j)} \quad (5)$$

Therefore, the probability that $\langle i,j\rangle$ is a span in the parse of $\mathbf{e}$ and has two children $\langle i,k\rangle$ and $\langle k+1,j\rangle$ can be calculated as follows:

$$\begin{aligned} & p(\langle i,j\rangle \to \langle i,k\rangle\langle k+1,j\rangle|\mathbf{e}) \\ &\quad = p(\langle i,j\rangle|\mathbf{e})\cdot p(\langle i,k\rangle\langle k+1,j\rangle|\langle i,j\rangle) \\ &\quad = \frac{\alpha(i,j)\cdot\beta(i,k)\cdot\beta(k+1,j)}{\beta(1,|\mathbf{e}|)} \end{aligned} \quad (6)$$

Since each candidate projected span aligns to one treelet at most, this probability is also the frequency of the candidate projected production related to the three spans.

The counting approach above is based on the assumption that there is a uniform distribution over the projected trees for every target sentence. The inside and outside algorithms and the other counting formulae are used to calculate the expected counts under this assumption. This looks like a single iteration of EM.

A binarized projected PCFG grammar can then be easily induced by maximum likelihood estimation. Due to word alignment errors, free translation, and exhaustive enumeration of possible projected productions, such a PCFG grammar may contain too much noisy nonterminals and production rules. We introduce a threshold b_{RULE} to filter the grammar. A production rule can be reserved only if its frequency is larger than b_{RULE}.

3.2 Relaxed Tree Projection

The projected PCFG grammar is used in the procedure of constituency projection. Such a grammar, as a kind of global syntactic knowledge, can attenuate the negative effect of word alignment error, free translation and syntactic non-isomorphism for the constituency projection between each single sentence pair. To obtain as optimal a projected constituency tree as possible, we have to integrate two kinds of knowledge: the local knowledge in the candidate projected production set of the target sentence, and the global knowledge in the projected PCFG grammar.

The integrated projection strategy can be conducted as follows. We parse each target sentence with the projected PCFG grammar $\mathbf{G}$, and use the candidate projected production set $\mathbf{D}$ to guide the PCFG parsing. The parsing procedure aims to find an optimum projected tree, which maximizes both the PCFG tree probability and the count of productions that also appear in the candidate projected pro-

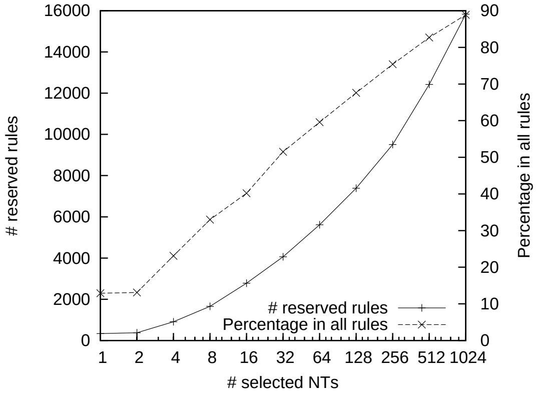

Figure 3: Rule counts corresponding to selected nonterminal sets, and their frequency summation proportions to the whole rule set.

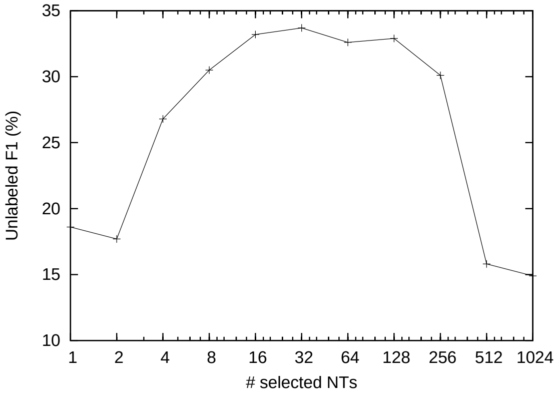

Figure 4: Performance curve of the projected PCFG grammars corresponding to different sizes of nonterminal sets.

duction set. The two optimization objectives can be coordinated as follows:

$$\tilde{y} = \operatorname*{argmax}_{y} \prod_{d \in y} (p(d|\mathbf{G}) \cdot e^{\lambda \cdot \delta(d, \mathbf{D})}) \tag{7}$$

Here, d represents a production; δ is a boolean function that returns 1 if d appears in $\mathbf{D}$ and returns 0 otherwise; λ is a weight coefficient that needs to be tuned to maximize the quality of the projected treebank.

4 Experiments

Our work focuses on the constituency projection from English to Chinese. The FBIS Chinese-English parallel corpus is used to obtain a projected constituent treebank. It contains 239 thousand sentence pairs, with about 6.9/8.9 million Chinese/English words. We parse the English sentences with the Charniak Parser (Charniak and Johnson, 2005), and tag the Chinese sentences with a POS tagger implemented faithfully according to (Collins, 2002) and trained on the Penn Chinese Treebank 5.0 (Xue et al., 2005). We perform word alignment by runing GIZA++ (Och and Ney, 2000), and then use the alignment results for constituency projection.

Following the previous works of unsupervised constituent parsing, we evaluate the projected parser on the subsets of CTB 1.0 and CTB 5.0, which contain no more than 10 or 40 words after the removal of punctuation. The gold-standard POS tags are directly used for testing. The evaluation for unsupervised parsing differs slightly from the standard PARSEVAL metrics, it ignores the multiplicity of brackets, brackets of span one, and the bracket labels. In all experiments we report the unlabeled F1 value which is the harmonic mean of the unlabeled precision and recall.

4.1 Projected PCFG Grammar

An initial projected PCFG grammar can be induced from the word-aligned and source-parsed parallel corpus according to section 3.1. Such an initial grammar is huge and contains a large amount of projected nonterminals and production rules, where many of them come from free translation and word alignment errors. We conservatively set the filtration threshold b_{RULE} as 1.0 to discard the rules with frequency less than one, the rule count falls dramatically from 3.3 millions to 92 thousands.

Figure 3 shows the statistics of the remained production rules. We sort the projected nonterminals according to their frequencies and select the top 2^N ($1 \leq N \leq 10$) best ones, and then discard the rules that fall out of the selected nonterminal set. The frequency summation of the rule set corresponding to 32 best nonterminals accounts for nearly 90% of the frequency summation of the whole rule set.

We use the developing set of CTB 1.0 (chapter 301-325) to evaluate the performance of a series of filtered grammars. Figure 4 gives the unlabeled F1 value of each grammar on all trees in the developing set. The filtered grammar corresponding to the set of top 32 nonterminals achieves the highest performance. We denote this grammar as $\mathbf{G}_{32}$ and use it

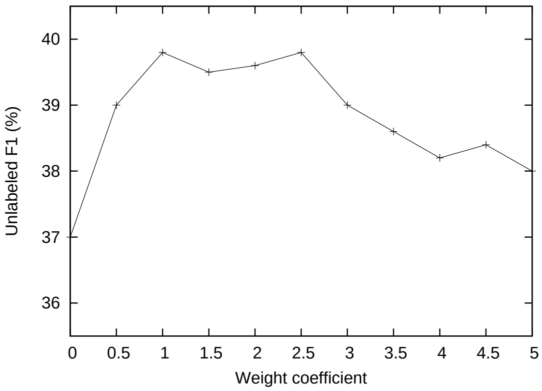

Figure 5: Performance curve of the Berkeley Parser trained on 5 thousand projected trees. The weight coefficient λ ranges from 0 to 5.

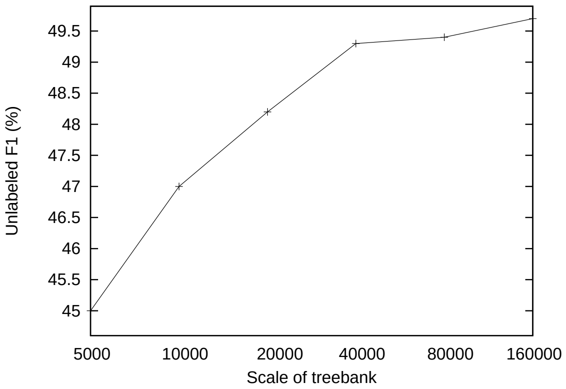

Figure 6: Performance curve of the Berkeley Parser trained on different amounts of best project trees. The scale of the selected treebank ranges from 5000 to 160000.

in the following tree projection procedure.

4.2 Projected Treebank and Parser

The projected grammar $\mathbf{G}_{32}$ provides global syntactic knowledge for constituency projection. Such global knowledge and the local knowledge carried by the candidate projected production set are integrated in a linear weighted manner as in Formula 7. The weight coefficient λ is tuned to maximize the quality of the projected treebank, which is indirectly measured by evaluating the performance of the parser trained on it.

We select the first 5 thousand sentence pairs from the Chinese-English FBIS corpus, and induce a series of projected treebanks using different λ, ranging from 0 to 5. Then we train the Berkeley Parser on each projected treebank, and test it on the developing set of CTB 1.0. Figure 5 gives the performance curve, which reports the unlabeled F1 values of the projected parsers on all sentences of the developing set. We find that the best performance is achieved with λ between 1 and 2.5, with slight fluctuation in this range. It can be concluded that, the projected PCFG grammar and the candidate projected production set do represent two different kinds of constraints, and we can effectively coordinate them by tuning the weight coefficient. Since different λ values in this range result in slight performance fluctuation of the projected parser, we simply set it to 1 for the constituency projection on the whole FBIS corpus.

There are more than 200 thousand projected trees induced from the Chinese-English FBIS corpus. It is a heavy burden for a parser to train on so large a treebank. And on the other hand, the free translation and word alignment errors result in many projected trees of poor-quality. We design a criteria to approximate the quality of the projected tree y for the target sentence x:

$$\tilde{Q}(y) = \sqrt[|x|-1]{\prod_{d\in y}(p(d|\mathbf{G})\cdot e^{\boldsymbol{\lambda}\cdot\delta(d,\mathbf{D})})} \qquad (8)$$

and use an amount of best projected trees instead of the whole projected treebank to train the parser. Figure 6 shows the performance of the Berkeley Parser trained on different amounts of selected trees. The performance of the Berkeley Parser constantly improves along with the increment of selected trees. However, treebanks containing more than 40 thousand projected trees can not brings significant improvement. The parser trained on 160 thousand trees only achieves an F1 increment of 0.4 points over the one trained on 40 thousand trees. This indicates that the newly added trees do not give the parser more information due to their projection quality, and a larger parallel corpus may lead to better parsing performance.

The Berkeley Parser trained on 160 thousand best projected trees is used in the final test. Table 1 gives the experimental results and the comparison with related works. This is a sparse table since the experiments of previous researchers focused on different data sets. Our projected parser significantly

System	CTB-TEST-40	CTB1-ALL-10	CTB5-ALL-10	CTB5-ALL-40
(Klein and Manning, 2004)	—	46.7	—	—
(Bod, 2006)	—	47.2	—	—
(Seginer, 2007)	—	—	**54.6**	38.0
(Jiang et al., 2010)	40.4	—	—	—
our work	**52.1**	**54.4**	**54.5**	**49.2**

Table 1: The performance of the Berkeley Parser trained on 160 thousand best projected trees, compared with previous works on constituency projection and unsupervised parsing. CTB-TEST-40: sentences ≤ 40 words from CTB standard test set (chapter 271-300); CTB1-ALL-10/CTB5-ALL-10: sentences ≤ 10 words from CTB 1.0/CTB 5.0 after the removal of punctuation; CTB5-ALL-40: sentences ≤ 40 words from CTB 5.0 after the removal of punctuation.

outperforms the parser of Jiang et al. (2010), where they directly adapt the DCA assumption of (Hwa et al., 2005) from dependency projection to constituency projection and resort to a better word alignment and a more complicated tree projection algorithm. This indicates that the RCA assumption is more suitable for constituency projection than the DCA assumption, and can induce a better grammar that much more reflects the language-specific syntactic idiosyncrasy of the target language.

Our projected parser also obviously surpasses existing unsupervised parsers. The parser of Seginer (2007) performs slightly better on CTB 5.0 sentences no more than 10 words, but obviously falls behind on sentences no more than 40 words. Figure 7 shows the unlabeled F1 of our parser on a series of subsets of CTB 5.0 with different sentence length upper limits. We find that even on the whole treebank, our parser still gives a promising result. Compared with unsupervised parsing, constituency projection can make use of the syntactic information of another language, so that it probably induce a better grammar. Although comparing a syntax projection technique to supervised or semi-supervised techniques seems unfair, it still suggests that if a resource-poor language has a bilingual corpus parallel to a resource-rich language with a human-annotated treebank, the constituency projection based on RCA assumption is a promising substitute for unsupervised parsing.

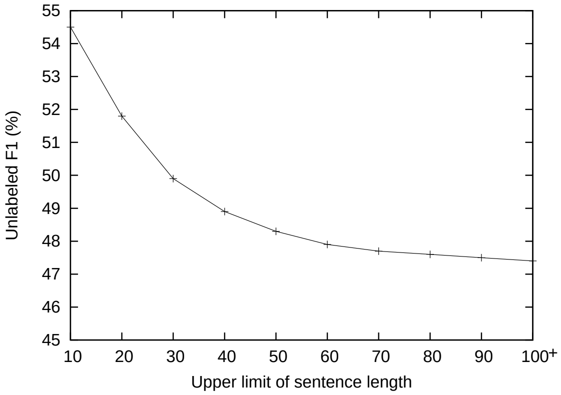

Figure 7: Performance of the Berkeley Parser on subsets of CTB 5.0 with different sentence length upper limits. $100+$ indicates the whole treebank.

5 Conclusion and Future Works

This paper describes a relaxed correspondence assumption (RCA) for constituency projection. Under this assumption a supposed constituent in the target sentence can correspond to an unrestricted treelet in the parse of the source sentence. Different from the direct correspondence assumption (DCA) widely used in dependency projection, the RCA assumption is more suitable for constituency projection, since it fundamentally tolerates the syntactic non-isomorphism between the source and target languages. According to the RCA assumption we propose a novel constituency projection method. First, a projected PCFG grammar is induced from the word-aligned source-parsed parallel corpus. Then, the tree projection is conducted on each sentence pair by a PCFG parsing procedure, which integrates both the global knowledge in the projected PCFG grammar and the local knowledge in the set of candidate projected productions.

Experiments show that the parser trained on the projected treebank significantly outperforms the projected parsers based on the DCA assumption. This validates the effectiveness of the RCA assumption and the constituency projection method, and indicates that the RCA assumption is more suitable for constituency projection than the DCA assumption. The projected parser also obviously surpasses the unsupervised parsers. This suggests that if a resource-poor language has a bilingual corpus parallel to a resource-rich language with a human-annotated treebank, the constituency projection based on RCA assumption is an promising substitute for unsupervised methods.

Although achieving appealing results, our current work is quite coarse and has many aspects to be improved. First, the word alignment is the fundamental precondition for projected grammar induction and the following constituency projection, we can adopt the better word alignment strategies to improve the word alignment quality. Second, the PCFG grammar is too weak due to its context free assumption, we can adopt more complicated grammars such as TAG (Joshi et al., 1975), in order to provide a more powerful global syntactic constraints for the tree projection procedure. Third, the current tree projection algorithm is too simple, more bilingual constraints could lead to better projected trees. Last but not least, the constituency projection and the unsupervised parsing make use of different kinds of knowledge, therefore the unsupervised methods can be integrated into the constituency projection framework to achieve better projected grammars, treebanks, and parsers.

Acknowledgments

The authors were supported by National Natural Science Foundation of China Contract 90920004, 60736014 and 60873167. We are grateful to the anonymous reviewers for their thorough reviewing and valuable suggestions.

References

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. Bayesian synchronous grammar induction. In *Proceedings of the NIPS*.

Rens Bod. 2006. An all-subtrees approach to unsupervised parsing. In *Proceedings of the COLING-ACL*.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of the EMNLP*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine-grained n-best parsing and discriminative reranking. In *Proceedings of the ACL*.

Wenliang Chen, Jun’ichi Kazama, and Kentaro Torisawa. 2010. Bitext dependency parsing with bilingual subtree constraints. In *Proceedings of the ACL*.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of the NAACL-HLT*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the EMNLP*.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*.

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

Liang Huang, Wenbin Jiang, and Qun Liu. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In *Proceedings of the EMNLP*.

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

Wenbin Jiang, Yajuan Lü, Yang Liu, and Qun Liu. 2010. Effective constituent projection across languages. In *Proceedings of the COLING*.

A. K. Joshi, L. S. Levy, and M. Takahashi. 1975. Tree adjunct grammars. *Journal Computer Systems Science*.

Dan Klein and Christopher D. Manning. 2004. Corpusbased induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the ACL*.

Terry Koo and Michael Collins. 2010. Efficient third-order dependency parsers. In *Proceedings of the ACL*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of the ACL*.

Jonas Kuhn. 2004. Experiments in parallel-text based grammar induction. In *Proceedings of the ACL*.

André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of EMNLP*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the ACL*.

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

Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. 2006. Labeled pseudoprojective dependency parsing with support vector machines. In *Proceedings of CoNLL*, pages 221–225.

Franz J. Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the ACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the ACL*.

Anoop Sarkar. 2001. Applying co-training methods to statistical parsing. In *Proceedings of NAACL*.

Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the ACL*.

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

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using english to parse korean. In *Proceedings of the EMNLP*.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the ACL*.

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proceedings of the EACL*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*.

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*.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of the ACL-IJCNLP*.