Using Bilingual Chinese-English Word Alignments to Resolve PP-Attachment Ambiguity in English

Victoria Fossum
Dept. of Computer Science
University of Michigan
Ann Arbor, MI 48104
vfossum@umich.edu

Kevin Knight
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
knight@isi.edu

Abstract

Errors in English parse trees impact the quality of syntax-based MT systems trained using those parses. Frequent sources of error for English parsers include PP-attachment ambiguity, NP-bracketing ambiguity, and coordination ambiguity. Not all ambiguities are preserved across languages. We examine a common type of ambiguity in English that is not preserved in Chinese: given a sequence "VP NP PP", should the PP be attached to the main verb, or to the object noun phrase? We present a discriminative method for exploiting bilingual Chinese-English word alignments to resolve this ambiguity in English. On a held-out test set of Chinese-English parallel sentences, our method achieves 86.3% accuracy on this PP-attachment disambiguation task, an improvement of 4% over the accuracy of the baseline Collins parser (82.3%).

1 Introduction

1.1 Motivation

Errors in English parse trees negatively impact the quality of syntax-based MT systems trained using those parses. (Quirk and Corston-Oliver, 2006) show that, in treelet translation, BLEU scores on English-German and English-Japanese experiments degrade as the amount of training data used to train the source dependency parser decreases. In the string-to-tree syntax-based MT system described in (Galley et al., 2004) and (Galley et al., 2006), the quality of translation rules extracted from each English parse tree and bilingual Chinese-English word alignment deteriorates as the quality of the parses and word alignments decreases. To quantify the impact of parse and alignment quality upon rule quality in such a system, we extract a gold-standard set of rules by applying the minimal rule extraction algorithm described in (Galley et al., 2004) to a bilingual corpus with gold alignments and gold English parse trees. We then extract rules from the same corpus using two different sources of automatically produced alignments (GIZA++ *union* and GIZA++ *refined*) and an automatic parser (Collins, 1997), and compute the precision, recall, and f-measure of the extracted rules against the gold-standard rule set. Table 1 illustrates that errors in the automatic alignments have a somewhat greater impact upon extracted rules than errors in the automatic parses. Nonetheless, errors introduced by the automatic parses still impact rule f-measure, causing a decrease in rule f-measure from 100.00% using the gold parses to 74.62% using the automatic parses. Thus, improving parse quality is likely to impact the quality of rules extracted using such a method, and we hypothesize that improving parse quality will impact BLEU score as well.

Many parse errors in English are due to common types of syntactic ambiguity such as PP-attachment ambiguity, coordination ambiguity, and NP-bracketing ambiguity. An example of ambiguous PP-attachment in English is shown in Figure 1: the PP "from reporters" can modify the VP "answered" or the NP "questions".

As long as syntactic ambiguities are not preserved across languages, we can use bilingual word alignments to disambiguate the construction. For example, in Chinese, PP's generally appear directly be-

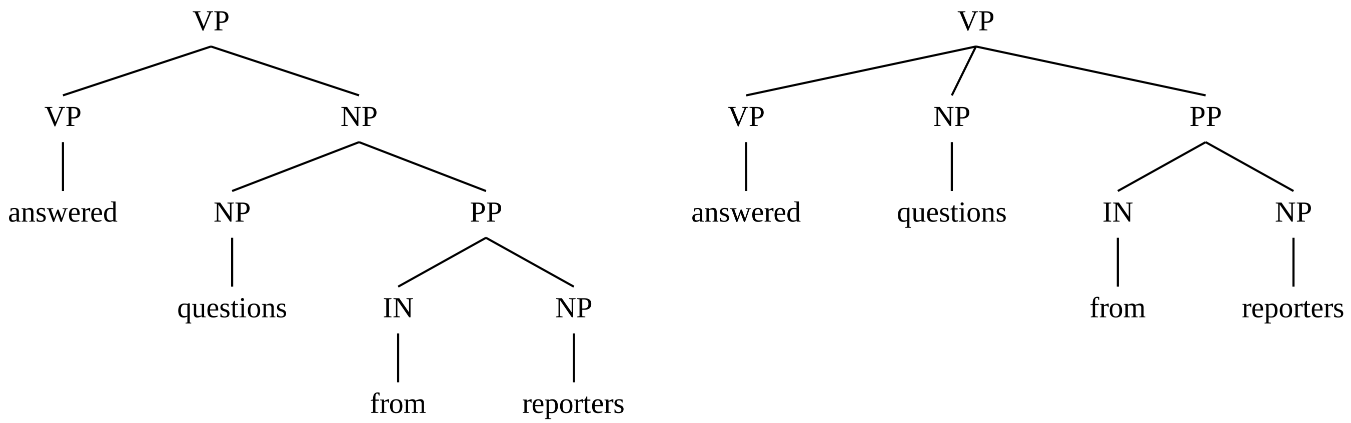

Figure 1: PP-attachment ambiguity in English

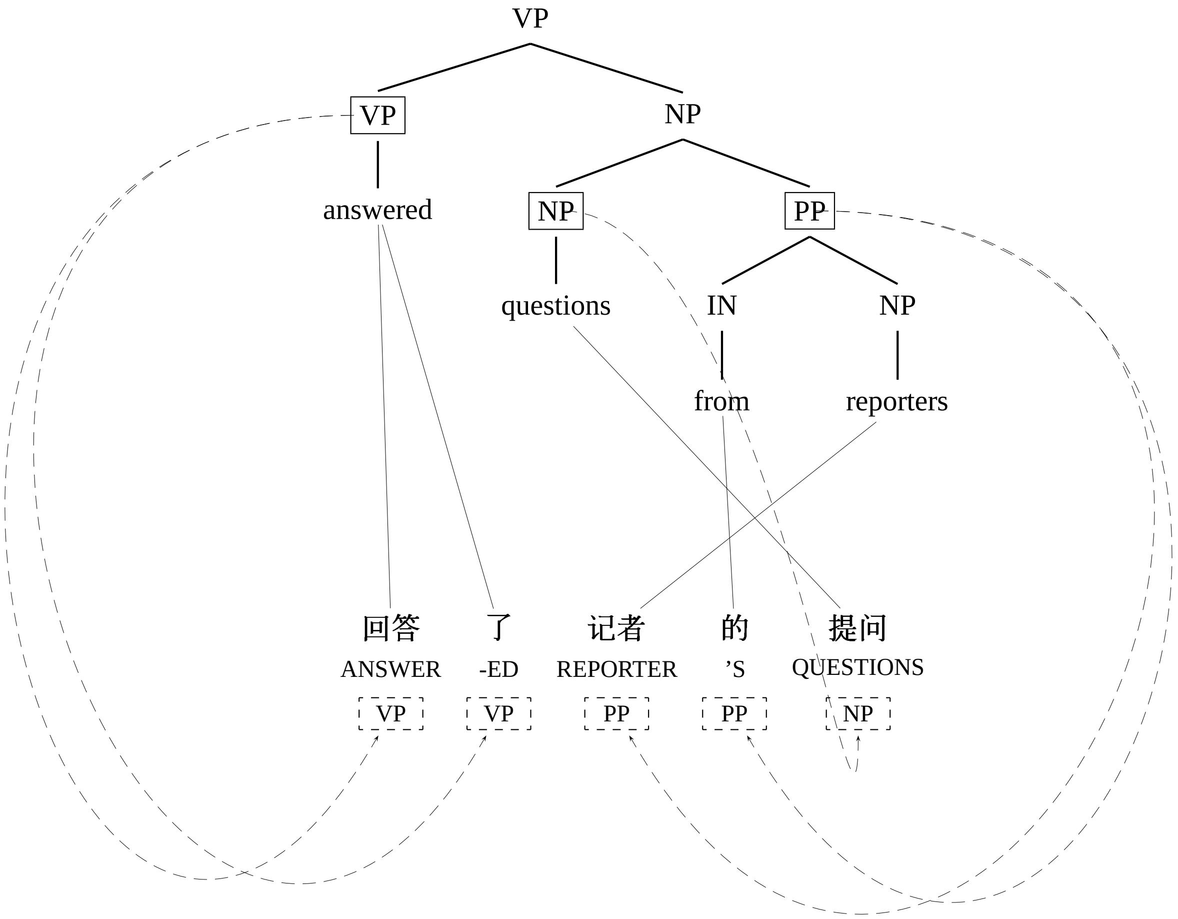

Figure 2: Resolving PP-attachment ambiguity using Chinese-English word alignments

Alignment	Parse	Rule Prec.	Rule Rec.	Rule F.
gold	gold	100.00	100.00	100.00
GIZA++ refined	gold	56.01	65.57	60.41
GIZA++ union	gold	63.57	52.31	57.39
gold	Collins	**73.25**	**76.04**	**74.62**
GIZA++ refined	Collins	44.04	54.46	48.70
GIZA++ union	Collins	50.70	44.18	47.22

Table 1: Impact of alignment and parse trees on precision, recall, and f-measure of extracted syntax-based translation rules

fore the head that they modify. Thus, PP-attachment ambiguity is not preserved from English to Chinese. Given the bilingual word alignments shown in Figure 2, we can deduce that the ordering of constituents on the Chinese side is "VP PP NP", indicating that the PP modifies the NP in Chinese, and presumably therefore in English as well.

1.2 Related Work

PP-Attachment Disambiguation Most previous work in PP-attachment disambiguation for English, whether unsupervised (Hindle and Rooth, 1993; Ratnaparkhi, 1998) or supervised (Brill and Resnik, 1994; Collins and Brooks, 1995), has focused on *monolingual* information such as relationships among the lexical heads of the VP ("answered"), NP ("questions"), and PP ("from") constituents, as well as the lexical head of the NP dominated by the PP ("reporters"). The statistical parsers of (Charniak, 1997) and (Collins, 1997) implement a variety of monolingual lexical and structural features to resolve syntactic ambiguities while constructing parse trees.

In contrast to these approaches, our approach uses *bilingual* word alignments to resolve PP-attachment ambiguity in English. In this respect, our approach is similar to that of (Schwartz et al., 2003), who leverage Japanese-English parallel bitext to improve the resolution of PP-attachment ambiguity on monolingual English text. In keeping with the approaches of (Hindle and Rooth, 1993) and (Ratnaparkhi, 1998), (Schwartz et al., 2003) estimate the probability of each possible attachment decision as follows: they first identify unambiguous instances of PP-attachment in English text, then compute the relative frequency of each attachment decision using these instances, conditioned on the verbs, nouns, and prepositions (or some subset thereof) appearing in the ambiguous construction. They subsequently use these statistics, computed over unambiguous instances, to estimate the probability of a PP attaching to an NP or VP in unseen (potentially ambiguous) English text.

(Schwartz et al., 2003) differs from the other unsupervised approaches in that the authors use *bilingual* information to identify unambiguous instances of PP-attachment. Specifically, they exploit the fact that PP-attachment is strictly unambiguous in Japanese by parsing both sides of a Japanese-English parallel bitext into LF, aligning nodes in the LF, and using the PP-attachment decision dictated by the Japanese side to infer the correct attachment decision on the English side. The authors evaluate their approach in two MT applications: English-Japanese and English-Spanish translation. They compare against a baseline method of PP-attachment ambiguity resolution that does *not* make use of the relative frequency statistics collected from the bilingual Japanese-English corpus. Their method improves Japanese-English translation quality but decreases Spanish-English translation quality, as measured by human evaluation.

Our work differs from that of (Schwartz et al., 2003) in several ways. First, because they evaluate their PP-attachment method only indirectly (by measuring its impact on English-to-Japanese and English-to-Spanish MT tasks), and not directly (by measuring the improvement in accuracy on the PP-attachment task), it is difficult to conclude from their results how effective their method is at improving PP-attachment accuracy (especially since their results in MT were mixed, with English-Japanese translation quality improving but English-Spanish translation quality worsening). In contrast, we evaluate our method directly on a PP-attachment task, and obtain a statistically significant gain of 4.0% in accuracy over the baseline Collins parser. Second, their method is unsupervised but requires a large parallel English-Japanese bitext in order to obtain reliable statistics of relative frequency for each set of lexical items; in contrast, our method is supervised but requires only a few hundred sentences of parallel English-Chinese bitext with manual parses on the English side during training. Finally, (Schwartz et al., 2003) implement hard cutoffs based on lexical associations, while we use a variety of features whose weights are learned discriminatively; thus our method appears to be more easily applicable to other problems in monolingual syntactic ambiguity resolution besides PP-attachment.

Bilingual Corpora for Monolingual Analysis (Yarowsky and Ngai, 2001) use bilingual word alignments to project part-of-speech taggers and NP-bracketers across languages; (Hwa et al., 2001; Hwa et al., 2005) extend this work to project syntactic dependency analyses across languages. Our work is similar to these approaches in that we use bilingual word alignments to induce a syntactic correspondence between languages; however, our focus is not on inducing analyses in the *target* language of the projection. Instead, we induce a syntactic correspondence from the source to the target language, then use that projection to resolve ambiguities in the syntactic analysis on the *source* side.

(Burkett and Klein, 2008) (to appear) parse both sides of a parallel English-Chinese bitext to generate a k-best list of English parses and a k-best list of Chinese parses, then rerank the $k \times k$-best list of English/Chinese parse tree pairs using the score assigned to each tree by the baseline parser; features of the word alignment; and features measuring structural correspondence between the English and Chinese trees in each pair. They obtain improvements in monolingual parse accuracy for both English and Chinese relative to state-of-the-art baseline English and Chinese parsers, and they obtain gains in translation quality when training a syntax-based MT system using the reranked trees. Our approach is different from that of (Burkett and Klein, 2008) in that we do not *rerank* a k-best list of parses; instead, we restrict ourselves to *repairing* common sources of attachment errors in English parses (specifically, PP-attachment).

1.3 Overview

The main contribution of this work is the use of Chinese-English bilingual word alignments to resolve PP-attachment ambiguity in English. Specifically, we address the following binary classification problem: given a “VP NP PP” sequence in English, should the PP be attached to the VP or the NP? To answer this question, we consider all instances of “VP NP PP” sequences in a bilingual corpus for which we have automatic Chinese-English word alignments, automatic English parses, and gold-standard English parses. In Section 2, we discuss two instances of PP-attachment ambiguity and illustrate how to use bilingual word alignments to resolve the ambiguity. Section 3.2 contains details about the data sets used. In Section 3.3, we present features of the automatic word alignment that can be used to determine whether the PP should be attached to the VP or the NP. In Section 3.4, we describe our procedure for training a perceptron for binary classification using these features. In Section 3.5 we describe our experimental setup, and in Section 4, we report the results of testing our classifier on a held-out portion of the data. We compare the accuracy of our classifier on this PP-attachment task against the accuracy of the PP-attachment decisions made by the baseline Collins parser. In Section 5, we analyze our results. Section 6 concludes and describes our future plans for extending the work presented here.

2 Bilingual Alignments and PP-attachment Ambiguity

Figure 1 illustrates a case where the correct attachment site of the PP (“from reporters”) is the NP (“questions”). To determine the correct attachment site for the English PP using the word alignments as a guide, we proceed as follows:

- **Project tags:** Wherever there is a VP NP PP sequence in the English parse tree, each node dominates a span of English words, and each English word is aligned with zero or more Chinese words. Label each of those aligned Chinese words with the category of the associated English node: “VP”, “NP”, or “PP”. Figure 2 gives an example; the resulting projected tag sequence on the Chinese side is “VP VP PP PP NP”.

- **Merge identical tags:** Merge adjacent labels in the Chinese tag sequence if they are identical, and ignore any words that have not received a projected tag. In Figure 2, the final merged tag sequence is “VP PP NP”. If a Chinese word receives more than one projected English tag, create a new hybrid tag combining the English tags for that word (for example, if the English VP and NP project onto the same Chinese word, that word receives a tag of “VP/NP”.)

- **Determine Chinese PP-attachment site:** As a general rule, PP’s in Chinese modify the head directly following them, so the Chinese tag sequence “VP PP NP” implies NP-attachment for the PP. We use a perceptron to learn such rules, predicting NP- versus VP- attachment.

- **Deduce English PP-attachment site:** Assuming that the PP-attachment dependency relation is preserved across languages, we can infer that the English PP should most likely be attached to the English NP.

Figure 3 illustrates a case where the correct attachment of the PP is to the main verb instead of to the object noun. In this case, tag projection proceeds as above; the resulting projected Chinese tag sequence is “NP PP VP”, thus indicating that the PP modifies the VP (Figure 4).

3 Methods

3.1 Problem Definition

There are two stages involved in disambiguating PP-attachment correctly. First, a parser must correctly label and bracket the main VP, object NP, and PP. Second, the parser must correctly choose an NP or VP attachment site for the PP. Since the latter problem is the focus of this work, we limit the scope of our classification task to those instances where the base VP, NP, and PP constituents have been labelled and bracketed correctly by the automatic parser. [1] In order to identify whether these constituentsa have been bracketed correctly, we refer to the gold standard parses. We then put the gold standard parses aside, and return to the following problem: given a sequence “VP NP PP” that the automatic parser has correctly labelled and bracketed, build a classifier that uses features of the automatic parse and the automatic bilingual word alignment to predict whether the PP should be attached to the VP or to the NP. To measure the accuracy of our classifier, we compute the percentage of correct attachment decisions, and compare this against the percentage of correct attachment decisions made by the baseline automatic parser.

3.2 Data Sets

Our training and test sets consist of bilingual Chinese-English sentence pairs that have been automatically parsed on the English side using the Collins parser (Collins, 1997), manually parsed on the English side to produce the gold-standard parses, and automatically word-aligned using GIZA++ with *refined* symmetrization (Och and Ney, 2003). GIZA++ is trained on 10M sentence pairs, but the total size of our PP-attachment data sets is 800 sentence pairs, from which we extract 300 instances of potentially ambiguous PP-attachment. Section 3.5 describes our experimental setup in further detail.

3.3 Features

In addition to the feature **collinsParserAttachment**, which is the attachment decision made by the baseline Collins parser, our feature set includes two types of features: lexical and alignment-based.

Lexical Features

- **englishPrepositionHead**: the lexical head of the English PP

- **projectedChinesePrepositionHead**: the Chinese word or words aligned to the lexical head of the English PP

Alignment-Based Features

- **projectedChineseTagSequence**: the sequence of part-of-speech tags after projection from English to Chinese

[1] Note that relaxing this restriction would affect our absolute performance numbers, but it would have no effect on our performance relative to the Collins parser: the Collins parser by definition fails on any case we have excluded.

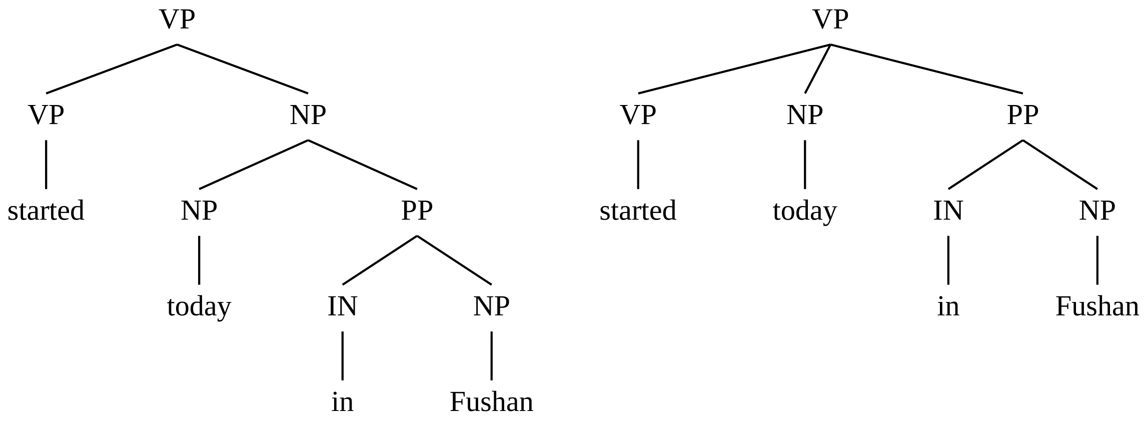

Figure 3: PP-attachment ambiguity in English

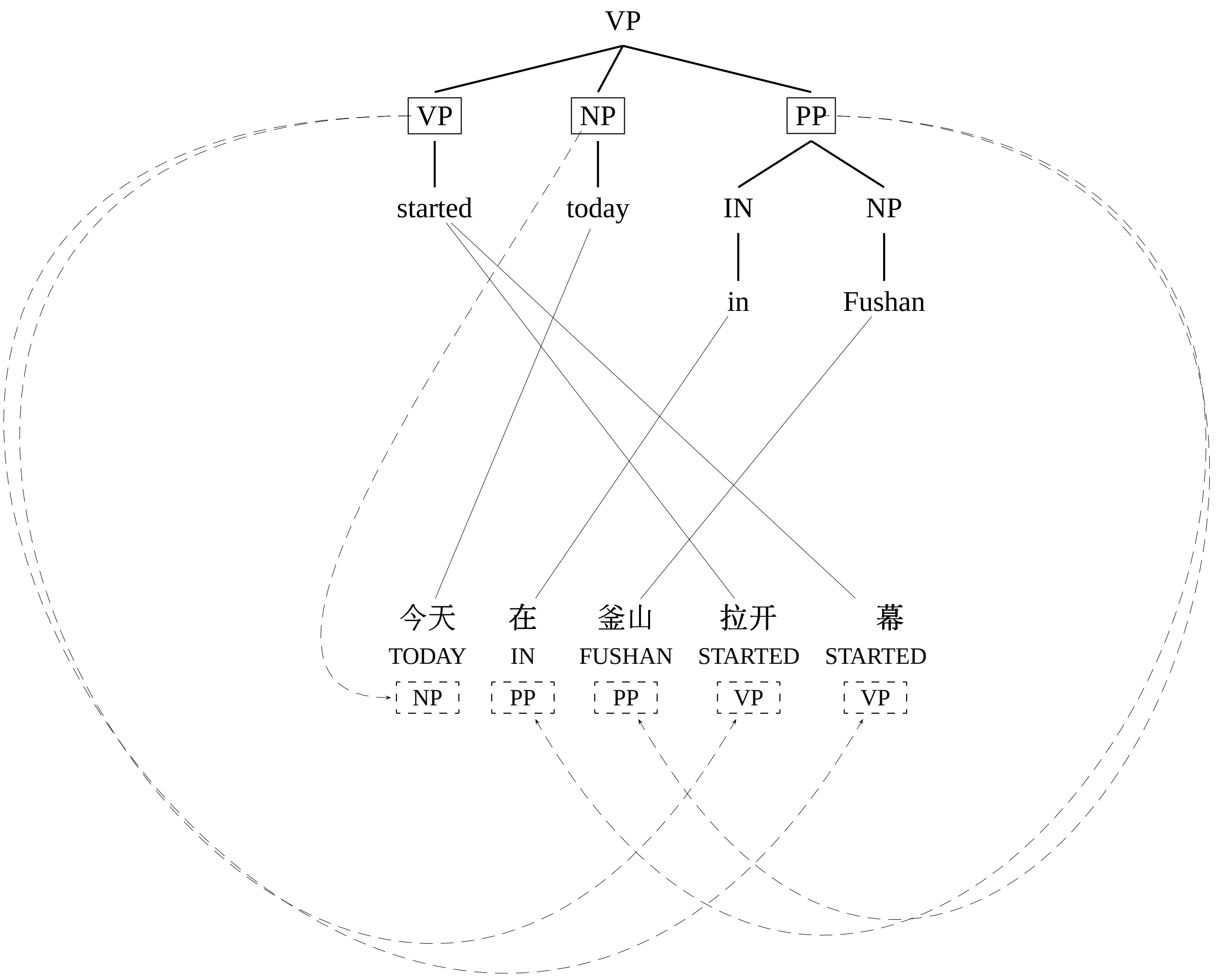

Figure 4: Resolving PP-attachment ambiguity using Chinese-English word alignments

- **projectedChineseTagSeqLength**: the number of tokens in the sequence of part-of-speech tags after projection from English to Chinese
- **initialChineseTag**: the initial tag in the projected Chinese sequence
- **projectedChineseTagAfterFirstPP**: the tag immediately following the first occurrence of a PP in the projected Chinese sequence
- **projectedFinalChineseTag**: the final tag in the projected Chinese sequence
- **splitNP**: whether or not the English NP tag was split into discontinuous tags on the Chinese side during projection
- **splitPP**: whether or not the English PP tag was split into discontinuous tags on the Chinese side during projection
- **splitVP**: whether or not the English VP tag was split into discontinuous tags on the Chinese side during projection

3.4 Perceptron Training

We train a perceptron for binary classification (Rosenblatt, 1958) to solve the PP-attachment problem using the features described in Section 3.3. We initialize the weights w of all features h to 0, and the bias b to 0. We make multiple passes over the training data. For each sentence pair in the training data, we represent the sentence pair as a vector x, where x_i is the value of feature h_i for the sentence pair. Our predicted attachment y_{hyp} is NP-attachment if $w \cdot x + b > 0$ and VP-attachment otherwise. If our predicted attachment y_{hyp} matches the gold attachment decision y_{gold}, then the example is correctly classified and we proceed to the next example. Otherwise, the example is incorrectly classified and we update the weights so that $w' = w + y_{gold} * x$ and the bias so that $b' = b + y_{gold}$. We stop training when the number of incorrect classifications no longer decreases on the training set. After training, we return the average weight vector over all iterations of training, following (Collins, 2002).

Method	% Correct PP-Attachment
Collins parser	82.3%
Perceptron classifier	**86.3%**

Table 2: PP-attachment accuracy of perceptron classifier vs. baseline Collins parser

3.5 Experiments

After training a perceptron classifier, we apply our classifier during testing to instances where the automatic parser has correctly identified "VP NP PP" sequences, and we predict the attachment site of the PP using the features described in Section 3.3 and the learned weights.

Due to the limited size of our data set (we use 800 sentence pairs of parallel Chinese-English text), we train and test our classifier using 10-fold cross-validation. We extract 300 instances of "VP NP PP" sequences from 800 sentences of parallel data, and divide this set of 300 instances into 10 sets of 30 instances each. We train on 9 of the sets, and measure accuracy on the held-out set. We then average the test set accuracy over all 10 iterations of cross-validation.

4 Results

We measure the accuracy of our method by classifying each instance of "VP NP PP" appearing in the test set as either attachment to the NP or attachment to the VP, and compare the accuracy of our method against the Collins parser baseline (Table 2). Our method achieves an average accuracy of 86.3% on held-out test sets in 10-fold cross validation, compared to 82.3% for the baseline Collins parser. This improvement in accuracy of 4% is statistically significant under a paired t-test (p=0.015).

5 Discussion

Our results (Table 2) show a statistically significant improvement of 4.0% in accuracy over the baseline Collins parser. To determine the relative contribution of each type of feature to the accuracy of the classifier, we perform feature ablation: we remove each type of feature from consideration in turn, and measure the impact upon classifier accuracy relative to the accuracy achieved using *all* feature types. Ta-

Features	**% Accuracy**
All Features	86.3
-projectedChinesePrepositionHead	84.7
-splitVP	85.3
-splitNP	85.3
-splitPP	85.3
-englishPrepositionHead	85.7
-finalChineseTag	85.7
-projectedChineseTagAfterFirstPP	86.0
-initialChineseTag	86.0
-projectedChineseTagSequence	86.3
-projectedChineseTagSeqLength	86.3

Table 3: Feature Ablation: Accuracy of PP-Attachment Classifier with Individual Features Removed

ble 3 illustrates the impact of removing each feature type upon the classifier accuracy; features are listed in the order of greatest impact upon classifier accuracy.

5.1 Avenues for Improvement

One problem with our current approach is that we use a relatively small data set of 300 instances; obtaining a larger set of bilingual sentence pairs with gold-standard English parses would likely improve performance of our classifier. A second problem is that our current method of tag projection from English to Chinese is not very robust to noise in the alignments; frequently, multiple English tags are projected onto the same Chinese word, and some Chinese words lack a projected tag altogether. By taking into account the presence of unaligned words on the Chinese side, and by using a more principled approach to handle the case of multiple English tags projecting onto a single Chinese word, we expect to improve the predictive power of the **chineseTagSequence** feature. A third problem is that we do not yet make use of any syntactic analysis on the Chinese side; we expect that incorporating the output of a Chinese part-of-speech tagger will improve the accuracy of tag projection.

6 Conclusion and Future Work

We have presented a method for English PP-attachment disambiguation using automatic bilingual Chinese-English word alignments. Our results confirm our hypothesis that bilingual information can help to resolve monolingual ambiguities as long as the ambiguities are not preserved across languages. In future work, we plan to extend the work presented here in the following ways:

- Expand our feature set for PP-attachment disambiguation to improve performance
- Improve the accuracy of the projected English-to-Chinese tag sequence by incorporating syntactic resources on the Chinese side, such as a Chinese part-of-speech tagger
- Address other types of syntactic ambiguity, such as NP-bracketing and coordination ambiguity
- Measure the impact of our method upon Chinese-English translation using a syntax-based MT system trained on the improved English parse trees

We expect that improving the accuracy of English parse trees by addressing common sources of syntactic ambiguity will improve not only parse accuracy, but also translation accuracy of a syntax-based MT system trained using these parses.

Acknowledgments

This research was supported by DARPA Contract No. HR0011-06-C-0022, and by the AT&T Labs Fellowship Program. We thank Steven Abney, David Chiang, and Liang Huang for helpful discussions, and Radu Soricut for providing his implementation of the Collins parser.

References

Eric Brill and Philip Resnik. *A rule-based approach to prepositional phrase attachment disambiguation*. Proceedings of COLING, 1994.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, Vol. 19, No. 2, 1993.

David Burkett and Dan Klein. *Two Languages are Better than One (for Syntactic Parsing)*. Proceedings of EMNLP, 2008.

Eugene Charniak. *Statistical parsing with a context-free grammar and word statistics*. Proceedings of AAAI, 1997.

Michael Collins. *Three Generative, Lexicalised Models for Statistical Parsing*. Proceedings of ACL, 1997.

Michael Collins. *Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms*. Proceedings of EMNLP, 2002.

Michael Collins and James Brooks. *Prepositional Phrase Attachment through a Backed-off Model*. Proceedings of the Workshop on Very Large Corpora, 1995.

Ido Dagan, Alon Itai, and Ulrike Schwall. *Two Languages are more Informative than One*. Proceedings of ACL, 1991.

Ido Dagan and Alon Itai. *Word Sense Disambiguation Using a Second Language Monolingual Corpus*. Computational Linguistics, Vol. 20, No. 4, 1994.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. *What's in a Translation Rule?* Proceedings of HLT/NAACL-04, 2004.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. *Scalable Inference and Training of Context-Rich Syntactic Translation Models*. Proceedings of ACL, 2006.

William Gale, Kenneth Church, and David Yarowsky. *A Method for Disambiguating Word Senses in a Large Corpus*. Computers and the Humanities, Vol. 26, No. 5-6, 1992.

Donald Hindle and Mats Rooth. *Structural Ambiguity and Lexical Relations*. Proceedings of ACL, 1993.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. *Evaluating translational correspondence using annotation projection*. Proceedings of ACL, 2001.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. *Bootstrapping parsers via syntactic projection across parallel texts*. Natural Language Engineering, Vol. 11, Issue 3, 2005.

Franz Josef Och and Hermann Ney. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, Vol. 29, No. 1, 2003.

Chris Quirk and Simon Corston-Oliver. *The Impact of Parse Quality on Syntactically-Informed Statistical Machine Translation*. Proceedings of EMNLP, 2006.

Adwait Ratnaparkhi. *Statistical models for unsupervised prepositional phrase attachment*. Proceedings of ACL, 1998.

Frank Rosenblatt. *The perceptron: A probabilistic model for information storage and organization in the brain*. Psychological Review, Vol. 65, No. 386, 1958.

Lee Schwartz, Takako Aikawa, and Chris Quirk. *Disambiguation of English PP Attachment Using Multilingual Data*. Proceedings of MT Summit IX, 2003.

David Yarowsky and Grace Ngai. *Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora*. Proceedings of NAACL, 2001.