Using Parse Features for Preposition Selection and Error Detection

Joel Tetreault
Educational Testing Service
Princeton
NJ, USA
JTetreault@ets.org

Jennifer Foster
NCLT
Dublin City University
Ireland
jfoster@computing.dcu.ie

Martin Chodorow
Hunter College of CUNY
New York, NY, USA
martin.chodorow
@hunter.cuny.edu

Abstract

We evaluate the effect of adding parse features to a leading model of preposition usage. Results show a significant improvement in the preposition selection task on native speaker text and a modest increment in precision and recall in an ESL error detection task. Analysis of the parser output indicates that it is robust enough in the face of noisy non-native writing to extract useful information.

1 Introduction

The task of preposition error detection has received a considerable amount of attention in recent years because selecting an appropriate preposition poses a particularly difficult challenge to learners of English as a second language (ESL). It is not only ESL learners that struggle with English preposition usage — automatically detecting preposition errors made by ESL speakers is a challenging task for NLP systems. Recent state-of-the-art systems have precision ranging from 50% to 80% and recall as low as 10% to 20%.

To date, the conventional wisdom in the error detection community has been to avoid the use of statistical parsers under the belief that a WSJ-trained parser's performance would degrade too much on noisy learner texts and that the traditionally hard problem of prepositional phrase attachment would be even harder when parsing ESL writing. However, there has been little substantial research to support or challenge this view. In this paper, we investigate the following research question: *Are parser output features helpful in modeling preposition usage in well-formed text and learner text?*

We recreate a state-of-the-art preposition usage system (Tetreault and Chodorow (2008), henceforth T&C08) originally trained with lexical features and augment it with parser output features. We employ the Stanford parser in our experiments because it consists of a competitive phrase structure parser *and* a constituent-to-dependency conversion tool (Klein and Manning, 2003a; Klein and Manning, 2003b; de Marneffe et al., 2006; de Marneffe and Manning, 2008). We compare the original model with the parser-augmented model on the tasks of preposition selection in well-formed text (fluent writers) and preposition error detection in learner texts (ESL writers).

This paper makes the following contributions:

- We demonstrate that parse features have a significant impact on preposition selection in well-formed text. We also show which features have the greatest effect on performance.
- We show that, despite the noisiness of learner text, parse features can actually make small, albeit non-significant, improvements to the performance of a state-of-the-art preposition error detection system.
- We evaluate the accuracy of parsing and especially preposition attachment in learner texts.

2 Related Work

T&C08, De Felice and Pulman (2008) and Gamon *et al.* (2008) describe very similar preposition error detection systems in which a model of correct prepositional usage is trained from well-formed text and a writer's preposition is compared with the predictions of this model. It is difficult to directly compare these systems since they are trained and tested on different data sets but they achieve accuracy in a similar range. Of these systems, only the DAPPER system (De Felice and Pulman, 2008; De Felice and Pulman, 2009; De Felice, 2009) uses a parser, the C&C parser (Clark and Curran, 2007)), to determine the head and complement of the preposition. De Felice and Pulman (2009) remark that the parser tends to be misled more by spelling errors than by grammatical errors. The parser is fundamental to their system and they do not carry out a comparison of the use of a parser to determine the preposition's attachments versus the use of shallower techniques. T&C08, on the other hand, reject the use of a parser because of the difficulties they foresee in applying one to learner data. Hermet *et al.* (2008) make only limited use of the Xerox Incremental Parser in their preposition error detection system. They split the input sentence into the chunks before and after the preposition, and parse both chunks separately. Only very shallow analyses are extracted from the parser output because they do not trust the full analyses.

Lee and Knutsson (2008) show that knowledge of the PP attachment site helps in the task of preposition selection by comparing a classifier trained on lexical features (the verb before the preposition, the noun between the verb and the preposition, if any, and the noun after the preposition) to a classifier trained on attachment features which explicitly state whether the preposition is attached to the preceding noun or verb. They also argue that a parser which is capable of distinguishing between arguments and adjuncts is useful for generating the correct preposition.

3 Augmenting a Preposition Model with Parse Features

To test the effects of adding parse features to a model of preposition usage, we replicated the lexical and combination feature model used in T&C08, training on 2M events extracted from a corpus of news and high school level reading materials. Next, we added the parse features to this model to create a new model "+Parse". In 3.1 we describe the T&C08 system and features, and in 3.2 we describe the parser output features used to augment the model. We illustrate our features using the example phrase *many local groups around the country*. Fig. 1 shows the phrase structure tree and dependency triples returned by the Stanford parser for this phrase.

3.1 Baseline System

The work of Chodorow *et al.* (2007) and T&C08 treat the tasks of preposition selection and error detection as a classification problem. That is, given the context around a preposition and a model of correct usage, a classifier determines which of the 34 prepositions covered by the model is most appropriate for the context. A model of correct preposition usage is constructed by training a Maximum Entropy classifier (Ratnaparkhi, 1998) on millions of preposition contexts from well-formed text.

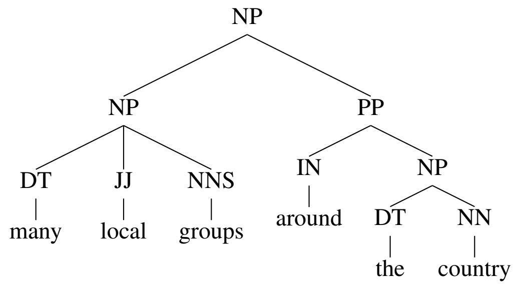
A context is represented by 25 lexical features and 4 combination features:

Lexical Token and POS n-grams in a 2 word window around the preposition, plus the head verb in the preceding verb phrase (PV), the head noun in the preceding noun phrase (PN) and the head noun in the following noun phrase (FN) when available (Chodorow et al., 2007). Note that these are determined not through full syntactic parsing but rather through the use of a heuristic chunker. So, for the phrase *many local groups around the country*, examples of lexical features for the preposition *around* include: *FN = country*, *PN = groups*, *left-2-word-sequence = local-groups*, and *left-2-POS-sequence = JJ-NNS*.

Combination T&C08 expand on the lexical feature set by combining the PV, PN and FN features, resulting in features such as PN-FN and PV-PN-FN. POS and token versions of these features are employed. The intuition behind creating combination features is that the Maximum Entropy classifier does not automatically model the interactions between individual features. An example of the *PN-FN* feature is *groups-country*.

3.2 Parse Features

To augment the above model we experimented with 14 features divided among five main classes. Table 1 shows the features and their values for our *around* example. The **Preposition Head and Complement** feature represents the two basic attachment relations of the preposition, i.e. its head (what it is attached to) and its complement (what is attached to it). **Relation** specifies the relation between the head and complement. The **Preposition Head and Complement Combined** features are similar to the T&C08 Combination features except that they are extracted from parser output.



```
amod(groups-3, many-1)
amod(groups-3, local-2)
prep(groups-3, around-4)
det(country-6, the-5)
pobj(around-4, country-6)
```

Figure 1: Phrase structure tree and dependency triples produced by the Stanford parser for the phrase *many local groups around the country*

Prep. Head & Complement
1. head of the preposition: *groups*
2. POS of the head: NNS
3. complement of the preposition: *country*
4. POS of the complement: NN
Prep. Head & Complement Relation
5. Prep-Head relation name: prep
6. Prep-Comp relation name: pobj
Prep. Head & Complement Combined
7. Head-Complement tokens: *groups-country*
8. Head-Complement tags: NNS-NN
Prep. Head & Complement Mixed
9. Head Tag and Comp Token: NNS-*country*
10. Head Token and Comp Tag: *groups*-NN
Phrase Structure
11. Preposition Parent: PP
12. Preposition Grandparent: NP
13. Left context of preposition parent: NP
14. Right context of preposition parent: -

Table 1: Parse Features

Model	**Accuracy**
combination only	35.2
parse only	60.6
combination+parse	61.9
lexical only	64.4
combination+lexical (T&C08)	65.2
lexical+parse	68.1
all features (+Parse)	68.5

Table 2: Accuracy on preposition selection task for various feature combinations

The **Preposition Head and Complement Mixed** features are created by taking the first feature in the previous set and backing-off either the head or the complement to its POS tag. This mix of tags and tokens in a word-word dependency has proven to be an effective feature in sentiment analysis (Joshi and Penstein-Rosé, 2009). All the features described so far are extracted from the set of dependency triples output by the Stanford parser. The final set of features (**Phrase Structure**), however, is extracted directly from the phrase structure trees themselves.

4 Evaluation

In Section 4.1, we compare the T&C08 and +Parse models on the task of preposition selection on well-formed texts written by native speakers. For every preposition in the test set, we compare the system's top preposition for that context to the writer's preposition, and report accuracy rates. In Section 4.2, we evaluate the two models on ESL data. The task here is slightly different - if the most likely preposition according to the model differs from the likelihood of the writer's preposition by a certain threshold amount, a preposition error is flagged.

4.1 Native Speaker Test Data

Our test set consists of 259K preposition events from the same source as the original training data. The T&C08 model performs at **65.2%** and when the parse features are added, the +Parse model improves performance by more than 3% to **68.5%**.[1] The improvement is statistically significant.

[1] Prior research has shown preposition selection performance accuracy ranging from 65% to nearly 80%. The differences are largely due to different test sets and also training sizes. Given the time required to train large models, we report here experiments with a relatively small model.

Model	**Accuracy**
T&C08	65.2
+Phrase Structure Only	67.1
+Dependency Only	68.2
+Parse	68.5
+head-tag+comp-tag	66.9
+left	66.8
+grandparent	66.6
+head-token+comp-tag	66.6
+head-tag	66.5
+head-token	66.4
+head-tag+comp-token	66.1

Table 3: Which parse features are important? Feature Addition Experiment

Table 2 shows the effect of various feature class combinations on prediction accuracy. The results are clear: a significant performance improvement is obtained on the preposition selection task when features from parser output are added. The two best models in Table 2 contain parse features. The table also shows that the non-parser-based feature classes are not entirely subsumed by the parse features but rather provide, to varying degrees, complementary information.

Having established the effectiveness of parse features, we investigate which parse feature classes contribute the most. To test each contribution, we perform a feature addition experiment, separately adding features to the T&C08 model (see Table 3). We make three observations. First, while there is overlapping information between the dependency features and the phrase structure features, the phrase structure features *are* making a contribution. This is interesting because it suggests that a pure dependency parser might be less useful than a parser which explicitly produces both constituent and dependency information. Second, using a parser to identify the preposition head seems to be more useful than using it to identify the preposition complement.[2] Finally, as was the case for the T&C08 features, the combination parse features are also important (particularly the tag-tag or tag/token pairs).

4.2 ESL Test Data

Our test data consists of 5,183 preposition events extracted from a set of essays written by non-native speakers for the Test of English as a Foreign Language (TOEFL®). The prepositions were judged by two trained annotators and checked by the authors using the preposition annotation scheme described in Tetreault and Chodorow (2008b). 4,881 of the prepositions were judged to be correct and the remaining 302 were judged to be incorrect.

Method	Precision	Recall
T&C08	0.461	0.215
+Parse	0.486	0.225

Table 4: ESL Error Detection Results

The writer's preposition is flagged as an error by the system if its likelihood according to the model satisfied a set of criteria (e.g., the difference between the probability of the system's choice and the writer's preposition is 0.8 or higher). Unlike the selection task where we use accuracy as the metric, we use precision and recall with respect to error detection. To date, performance figures that have been reported in the literature have been quite low, reflecting the difficulty of the task. Table 4 shows the performance figures for the T&C08 and +Parse models. Both precision and recall are higher for the +Parse model, however, given the low number of errors in our annotated test set, the difference is not statistically significant.

5 Parser Accuracy on ESL Data

To evaluate parser performance on ESL data, we manually inspected the phrase structure trees and dependency graphs produced by the Stanford parser for 210 ESL sentences, split into 3 groups: the sentences in the first group are fluent and contain no obvious grammatical errors, those in the second contain at least one preposition error and the sentences in the third are clearly ungrammatical with a variety of error types. For each preposition we note whether the parser was successful in determining its head and complement. The results for the three groups are shown in Table 5. The figures in the first row are for correct prepositions and those in the second are for incorrect ones.

The parser tends to do a better job of determining the preposition's complement than its head which is not surprising given the well-known problem of PP attachment ambiguity. Given the preposition, the preceding noun, the preceding

[2] De Felice (2009) observes the same for the DAPPER system.

	OK	
	Head	Comp
Prep Correct	86.7% (104/120)	95.0% (114/120)
Prep Incorrect	-	-
	Preposition Error	
	Head	Comp
Prep Correct	89.0% (65/73)	97.3% (71/73)
Prep Incorrect	87.1% (54/62)	96.8% (60/62)
	Ungrammatical	
	Head	Comp
Prep Correct	87.8% (115/131)	89.3% (117/131)
Prep Incorrect	70.8% (17/24)	87.5% (21/24)

Table 5: Parser Accuracy on Prepositions in a Sample of ESL Sentences

verb and the following noun, Collins (1999) reports an accuracy rate of 84.5% for a PP attachment classifier. When confronted with the same information, the accuracy of three trained annotators is 88.2%. Assuming 88.2% as an approximate PP-attachment upper bound, the Stanford parser appears to be doing a good job. Comparing the results over the three sentence groups, its ability to identify the preposition's head is quite robust to grammatical noise.

Preposition errors in isolation do not tend to mislead the parser: in the second group which contains sentences which are largely fluent apart from preposition errors, there is little difference between the parser's accuracy on the correctly used prepositions and the incorrectly used ones. Examples are

```
(S (NP I)
   (VP had
       (NP (NP a trip)
           (PP for (NP Italy))
       )
   )
)
```

in which the erroneous preposition *for* is correctly attached to the noun *trip*, and

```
(S (NP A scientist)
   (VP devotes
       (NP (NP his prime part)
           (PP of (NP his life))
       )
       (PP in (NP research))
   )
)
```

in which the erroneous preposition *in* is correctly attached to the verb *devotes*.

6 Conclusion

We have shown that the use of a parser can boost the accuracy of a preposition selection model tested on well-formed text. In the error detection task, the improvement is less marked. Nevertheless, examination of parser output shows that parse features can be extracted reliably from ESL data.

For our immediate future work, we plan to carry out the ESL evaluation on a larger test set to better gauge the usefulness of a parser in this context, to carry out a detailed error analysis to understand why certain parse features are effective and to explore a larger set of features.

In the longer term, we hope to compare different types of parsers in both the preposition selection and error detection tasks, i.e. a task-based parser evaluation in the spirit of that carried out by Miyao *et al.* (2008) on the task of protein pair interaction extraction. We would like to further investigate the role of parsing in error detection by looking at other error types and other text types, e.g. machine translation output.

Acknowledgments

We would like to thank Rachele De Felice and the reviewers for their very helpful comments.

References

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, Prague, Czech Republic, June.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Michael Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Rachele De Felice and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 english. In *Proceedings of the 22nd COLING*, Manchester, United Kingdom.

Rachele De Felice and Stephen Pulman. 2009. Automatic detection of preposition errors in learning writing. *CALICO Journal*, 26(3):512–528.

Rachele De Felice. 2009. *Automatic Error Detection in Non-native English*. Ph.D. thesis, Oxford University.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Proceedings of the COLING08 Workshop on Cross-framework and Cross-domain Parser Evaluation*, Manchester, United Kingdom.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, Genoa, Italy.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modelling for ESL error correction. In *Proceedings of the International Joint Conference on Natural Language Processing*, Hyderabad, India.

Matthieu Hermet, Alain Désilets, and Stan Szpakowicz. 2008. Using the web as a linguistic resource to automatically correct lexico-syntactic errors. In *Proceedings of LREC*, Marrekech, Morocco.

Mahesh Joshi and Carolyn Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316, Singapore.

Dan Klein and Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 423–430, Sapporo, Japan.

Dan Klein and Christopher D. Manning. 2003b. Fast exact inference with a factored model for exact parsing. In *Advances in Neural Information Processing Systems*, pages 3–10. MIT Press, Cambridge, MA.

John Lee and Ola Knutsson. 2008. The role of PP attachment in preposition generation. In *Proceedings of CICling*. Springer-Verlag Berlin Heidelberg.

Yusuke Miyao, Rune Saetre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of the 46th Annual Meeting of the ACL*, pages 46–54, Columbus, Ohio.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.

Joel Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd COLING*, Manchester, United Kingdom.

Joel Tetreault and Martin Chodorow. 2008b. Native Judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*, Manchester, United Kingdom.