Data-Driven Correction of Function Words in Non-Native English

Adriane Boyd **Detmar Meurers**
Seminar für Sprachwissenschaft
Universität Tübingen
{adriane,dm}@sfs.uni-tuebingen.de

Abstract

We extend the n-gram-based data-driven prediction approach (Elghafari, Meurers and Wunsch, 2010) to identify function word errors in non-native academic texts as part of the Helping Our Own (HOO) Shared Task. We focus on substitution errors for four categories: prepositions, determiners, conjunctions, and quantifiers. These error types make up 12% of the errors annotated in the HOO training data.

In our best submission in terms of the error detection score, we detected 67% of preposition and determiner substitution errors, 40% of conjunction substitution errors, and 33% of quantifier substitution errors. For approximately half of the errors detected, we were also able to provide an appropriate correction.

1 Introduction

We take as a starting point the preposition prediction approach of Elghafari, Meurers and Wunsch (2010). They explore a surface-based approach for predicting prepositions in English which uses frequency information from web searches to choose the most likely preposition given the context. For each preposition found in the text, the prediction algorithm considers three words of context on each side, building a 7-gram with a preposition slot in the middle:

```
rather a question __ the scales falling
```

For each prediction task, a *cohort* of queries is constructed with each of the candidate prepositions in the slot to be predicted:

```
1. rather a question of the scales falling
2. rather a question in the scales falling
...
9. rather a question on the scales falling
```

The queries are submitted to the Yahoo search engine and the query with the largest number of hits provides the predicted preposition. If no hits are found for any of the 7-gram queries, shorter overlapping n-grams are used to approximate the 7-gram query. If there are still no hits, the overlap backoff will continue reducing the n-gram length until it reaches 3-grams. If no hits are found at the 3-gram level, the most frequent preposition (*of*) is predicted.

Elghafari, Meurers and Wunsch (2010) showed that this surface-based approach is competitive with published state-of-the-art machine learning approaches using complex feature sets (Gamon et al., 2008; De Felice, 2008; Tetreault and Chodorow, 2008; Bergsma et al., 2009). For a set of nine frequent prepositions (*of, to, in, for, on, with, at, by, from*), they accurately predicted 77%. For these nine prepositions, De Felice (2008) identified a baseline of 27% for the task of choosing a preposition in a slot (choose *of*). Humans performing the same task agree 89% of the time.

2 Our Approach

We extend the preposition prediction approach to four function word categories: conjunctions, determiners, prepositions, and quantifiers. Table 1 shows the sets of function words for each category and the associated HOO error codes. The function word lists are compiled from all single-word substitution errors of these types in the HOO training data.[1] The counts show the number of occurrences of the error types in the test data, along with the total number of occurrences of the function word candidates.

[1] We also removed the correction *using* from the preposition list since it is not a preposition.

Categ.	Codes	#	Candidates	#
Conj.	RC	2	but, if, whether, whereas, however, although	80
Det.	RD, FD, DD, AGD, CD, ID	17	a, whose, their, this, an, these, the, its, those	1572
Prep.	RT, DT	86	in, on, about, over, from, onto, for, among, of, into, within, to, as, at, under, between, with, by	2126
Quant.	RQ, FQ, CQ, DQ, IQ, AGQ	4	less, many, some, fewer, much, certain	78
Total		109		3856

Table 1: Function Words with Frequency in Test Data

To adapt the prediction approach for the HOO shared task, we replace the Yahoo search engine used by Elghafari et al. (2010) with the ACL Anthology Reference Corpus (ARC, Bird et al., 2008) and modify the prediction algorithm to keep the original token rather than predicting the most frequent candidate in cases where no hits for any n-grams are found. One drawback of ARC is that it contains native and non-native texts; we have not yet attempted to filter non-native texts.

Using ARC rather than web searches allows us to abstract away from the surface context by substituting POS tags and lemmas in the n-gram context. We use TreeTagger to tag and lemmatize ARC and create three different levels of context abstraction: a) surface context, b) POS context, and c) limited POS/lemma substitutions (POS for CD, SYM, LS; lemmas for comparative adjectives and most verbs). We use the same context throughout, though substitutions could be customized for each type, e.g., determiner selection depends on adjective and noun onsets (*a* vs. *an*), but preposition selection does not.

3 Results

We will discuss our results from two perspectives:

- **Global:** For each function word (correct or incorrect), was a correct prediction made?
- **Error detection:** For each function word substitution error, was the error detected/corrected?

For both perspectives, we can calculate precision and recall for the n-gram prediction approach:

$$\text{precision} = \frac{\text{correct predictions from n-gram approach}}{\text{\# predicted by n-gram approach}}$$

$$\text{recall} = \frac{\text{correct predictions from n-gram approach}}{\text{\# total prediction tasks}}$$

We here present the results for our run #2 in the HOO shared task, our best performing submission in terms of detection score. Run #2 uses the ARC reference corpus with limited POS/lemma substitutions, showing that an appropriate level of abstraction in the n-gram context can lead to improvement over purely surface-based contexts.

3.1 Baseline

The counts in Table 1 show that there is a high global baseline accuracy (= keep original word) for this subtask in the HOO challenge. The baseline for all four categories is 97.2% and the individual function category baselines vary from 94.9% to 98.9%. Thus, predicting the original word would give a high global accuracy for the function word prediction task in the HOO data; however, it would obviously not detect or correct any errors.

3.2 Global Results

Figure 1 shows the global accuracy, precision, and recall as the minimum n-gram length is increased from 3 to 7. The global precision, recall, and accuracy are ∼70% for n-gram length 3. As the minimum n-gram length increases, the global accuracy and precision increase to 97% as recall drops to 1.5% since most 7-grams from the test data are not found in the reference corpus. Data sparsity issues are magnified by the fact that the n-gram context may contain additional errors.

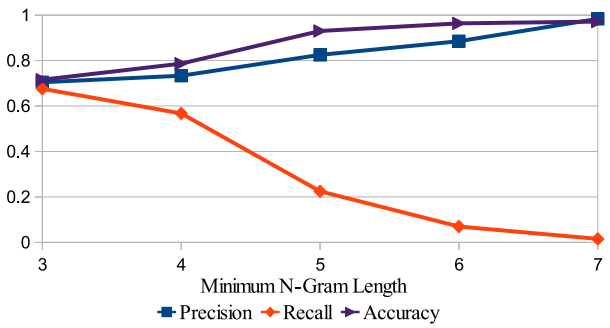

Figure 1: Global Accuracy, Precision, and Recall

3.3 Error Detection and Correction Results

Figure 2 shows the error detection/correction precision and recall as the minimum n-gram length increases from 3 to 6.

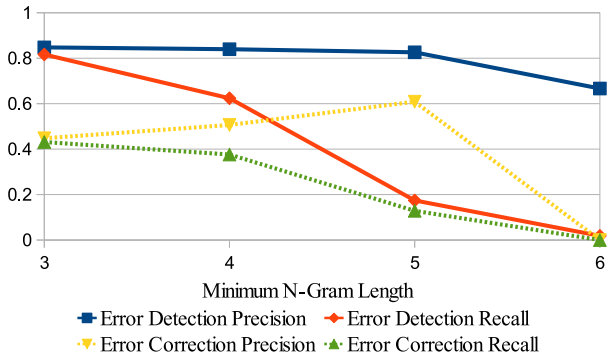

Figure 2: Error Detection and Correction F-Score

For 3-grams, the detection f-score is over 80% with a correction f-score of 44% (but keep in mind that the global accuracy is only 72% at this point). As the minimum n-gram length increases to 6, fewer errors are detected as longer n-grams are not found. From 3-grams to 5-grams, the detection precision stays relatively constant while the correction precision increases from 45% to 60%. Longer n-gram context thus leads to more accurate predictions.

4 Discussion and Conclusion

Extending the n-gram prediction approach (Elghafari, Meurers and Wunsch, 2010) with a genre-specific reference corpus and generalized contexts, we are able to detect 33%–67% of the targeted function word substitution errors in the HOO test corpus. We provide an appropriate correction for approximately half of the errors detected. However, our method currently miscorrects about ten function words for each one it detects as an error, which is reflected in the relatively low HOO detection precision score (14%) in the 'no bonus' condition.

As our approach was originally designed to predict rather than to correct function words, further customizations may improve the performance for correction tasks, which unlike prediction tasks have access to the word used in the original text. Instead of the raw counts we are currently using, one could weight the words in the candidate sets for each prediction task in order to account for global frequency (e.g., *the* is more frequent than *these* in contexts where both are correct) and in order to make it possible to add an explicit bias towards leaving the original word unmodified, since the HOO data shows that such a high percentage of function words in this genre are indeed correct.

The results we presented take into account only the four types of errors from the HOO error scheme of Table 1, however many errors involving function word substitutions in the HOO data are not actually annotated as such, but are part of other error types annotating multiple words. As a result, our system also detects some function word errors which were annotated as compound change, replace verb (e.g., phrasal verb error), wrong verb form, and replace adverb. The current HOO annotation scheme does not have the granularity to systematically identify all function word errors – a shortcoming worth addressing in order to support incremental, modular research on error detection. This is particularly relevant in light of the lack of inter-annotator agreement studies establishing which distinctions from the various error annotation schemes in the literature can reliably be annotated given the information present in the text (cf. Meurers, 2012, and references therein).

References

Shane Bergsma, Dekang Lin and Randy Goebel, 2009. Web-scale N-gram models for lexical disambiguation. In *IJCAI*.

Steven Bird, Robert Dale et al., 2008. The ACL Anthology Reference Corpus. In *LREC*.

Rachele De Felice, 2008. Automatic Error Detection in Non-native English. Ph.D. thesis, Oxford.

Anas Elghafari, Detmar Meurers and Holger Wunsch, 2010. Exploring the Data-Driven Prediction of Prepositions in English. In *COLING*.

Michael Gamon, Jianfeng Gao et al., 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *IJCNLP*.

Detmar Meurers, 2012. Natural Language Processing and Language Learning. In *Encyclopedia of Applied Linguistics*, Wiley-Blackwell, Oxford. http://purl.org/dm/papers/meurers-12.html.

Joel Tetreault and Martin Chodorow, 2008. Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection. In *COLING*.