Detection and Correction of Preposition and Determiner Errors in English: HOO 2012

Pinaki Bhaskar **Aniruddha Ghosh** **Santanu Pal** **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering, Jadavpur University
188, Raja S. C. Mallick Road
Kolkata – 700032, India
pinaki.bhaskar@gmail.com arghyaonline@gnail.com santanu.pal.ju@gmail.com sivaji_cse_ju@yahoo.com

Abstract

This paper reports on our work in the HOO 2012 shared task. The task is to automatically detect, recognize and correct the errors in the use of prepositions and determiners in a set of given test documents in English. For that, we have developed a hybrid system of an n-gram statistical model along with some rule-based techniques. The system has been trained on the HOO shared task's training datasets and run on the test set given. We have submitted one run, which has demonstrated an F-score of 7.1, 6.46 and 2.58 for detection, recognition and correction respectively before revision and F-score of 8.22, 7.59 and 3.16 for detection, recognition and correction respectively after revision.

1 Introduction

Writing research papers or theses in English is a very challenging task for those researchers and scientists whose first language or mother tongue is not English. Depicting their research works properly in English is a hard job for them. Generally their papers, which are submitted to conferences, may be rejected not because of their research works but because of the English writing, which makes the papers harder for the reviewer to understand the intentions of author. This kind of problem will be faced in any field where someone has to provide material in a language other than his/her first language.

The mentoring service of Association for Computational Linguistics (ACL) is one part of a response. This service can address a wider range of problems than those related purely to writing. The aim of this service is that a research paper should be judged only on its research content.

The organizer of "Help Our Own" (HOO) proposed and initiated a shared task in 2011 (Dale and Kilgarriff, 2010), which attempts to tackle the problem by developing tools or techniques for the non-native speaker of English, which will automatically correct the English prose of the papers so that they can be accepted. This tools and techniques may also help native English speakers. This task is simply expressed as text-to-text generation or Natural language Generation (NLG). In the 2011 shared task, all possible errors were covered which made the task enormously huge. In 2012, the task is more specific and only deals with determiners and prepositions as described in (Dale and Kilgarriff, 2011).

For this shared task, HOO, we have developed two models, one is rule-based model and the other is the statistical model for both determiners and prepositions. Then we have combined both these models and developed our system for HOO 2012.

2 Related Work

The English language belongs to the Germanic languages branch of the Indo-European language family, widely spoken on six continents. The HOO

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 201–207,
Montréal, Canada, June 3-8, 2012.

shared task is organized to help authors with writing tasks. Identifying grammatical and linguistic errors in text is an open challenge to researchers. In recent times, researchers (Heidorn, 2000) have provided quite a benchmark for spell checker and grammar checkers, which is commonly available. In this task it is aimed to correct errors beyond the scope of these commonly available checkers i.e. detection and correction of jarring errors at part-of-speech (POS) level, syntax level and semantic level. Earlier Heidorn (1975) developed augmented phrase structure grammar. (Tetreault et. al., 2008) has dealt with error pattern with preposition by non-native speakers. Meurers and Wunsch (2010) showed a surface based state-of-the-art machine learning technique, which deals with some frequently used prepositions. (Elghafari et al., 2010) worked on Data-Driven Prediction of Prepositions in English. Boyd et al. (2011) used an n-gram based machine-learning approach. Last year we have also participated in this shared task; our system report was reported in (Bhaskar et. al., 2011).

3 Corpus Statistics

There are two sets of data, training set and test set provided by the organizer. The training set has 1000 documents, which are collected from the FCE dataset. The publicly available dataset was in the native FCE format. So, the organizer first converted it to the HOO data format. Then CUP annotators found the errors and marked them up in the dataset. This year the task is only about the errors related to prepositions and determiners. So the organizer set only six types of errors, listed in table 1, which were dealt with this year. Hence, the other errors were discarded and replace with its corresponding standoff annotation in the training set. The training set consists of 1000 documents of total 374680 words, which means 375 words per document. All the standoff annotations of training set were provided and an example of the standoff annotation is shown in the figure 1. Table 2 gives the error statistics of training set as reported in (Dale et. al., 2012).

The test dataset has another 100 documents, which contain total of 18013 words at an average of 180 words per document. The test data was processed as the training data was done, but the standoff annotation of the test documents was not provided before the task completion. The documents were provided in XML format as shown in the figure 2.

Error Type	Tag	Original	Correction
Replacement Preposition	RT	He was born **on** January	He was born **in** January
Missing Preposition	MT	Because it reminds me my childhood.	Because it reminds me **of** my childhood.
Unwanted Preposition	UT	Regarding **to** the accommodation	Regarding the accommodation
Replacement Determiner	RD	I used to going with my friends to **the** camp.	I used to going with my friends to **a** camp.
Missing Determiner	MD	That will be nice to go on 1st of July	That will be nice to go on **the** 1st of July
Unwanted Determiner	UD	The most suitable time for shopping is weekend when parents don't work and children haven't got **a** school.	The most suitable time for shopping is weekend when parents don't work and children haven't got school.

Table 1. Examples of the six types of error.

Error Type	**# Training**	**# Test**	
		# before Revised	**# after Revised**
UT	822	43	39
MT	1104	57	56
RT	2618	136	148
Prep	4545	236	243
UD	1048	53	62
MD	2230	125	131
RD	609	39	37
Det	3887	217	230
Total	**8432**	**453**	**473**
Words/ Error	44.18	39.77	38.08

Table 2. Error Statistics in the Training set.

```
<edit end="779" file="0004" in-
dex="0008" part="1" start="775"
type="UD">
  <original>the </original>
    <corrections>
      <correction>
        <empty/>
      </correction>
  </corrections>
</edit>

<edit end="1041" file="0004" in-
dex="0010" part="1" start="1039"
type="RT">
  <original>in</original>
    <corrections>
      <correction>at</correction>
  </corrections>
</edit>
```

Figure 1: An example of a standoff error annotation

```
<?xml version="1.0" encod-
ing="utf-8"?>
<HOO version="2.1">
  <HEAD sortkey="" source-
  type="FCE">
    <CANDIDATE>
      <AGE>20-30</AGE>
    </CANDIDATE>
  </HEAD>
  <BODY>
    <PART id="1">
      <P>Dear Chris</P>
      <P>I was great …</P>
      .
      .
      .
    </PART>
  </BODY>
</HOO>
```

Figure 2: An example of the XML format of documents

4 System Description

The task is consisted of two coarse parts – Preposition and Determiner detection, recognition and correction. In our previous year's hybrid model, to resolve preposition errors, a rule-based model was developed and for determiner errors, a linear statistical method was used. There was no linear statistical model for prepositions. So this year we have induced a statistical model to incorporate larger coverage of preposition error detection, which is not detected by the appropriate preposition list described in section 4.1.2.

To resolve preposition errors and determiner errors we have built a hybrid model for both of them and used a voting technique among the rule based and statistical model for determiners and rule based post processing for prepositions. The system architecture is shown in the figure 3.

4.1 Preposition Error Detection

4.1.1 Statistical Model for Preposition

An n-gram based linear statistical model is used. From the training corpus, it was trained with 3, 5 and 7-gram models. After testing, the 5-gram model is performing best as from 3-gram, the statistical model fails to classify since probability distance is too small among the probable set to distinguish proper one while in 7-gram it fails to score high as training data set is relatively small and there are no similar occurrences. For the statistical model, different linguistic information is taken as features. Initially, surface words are only considered which actually is similar to fingerprinting technique. Due to different inflected forms, the system fails to identify possible cases for a similar type of error with different inflected forms. Hence the root form of the word is included as a feature. Chunk information is included as a feature. The preposition with same word varies with if following word is animate or inanimate. As example,

```
collaborate with SB
collaborate in/on ST
```

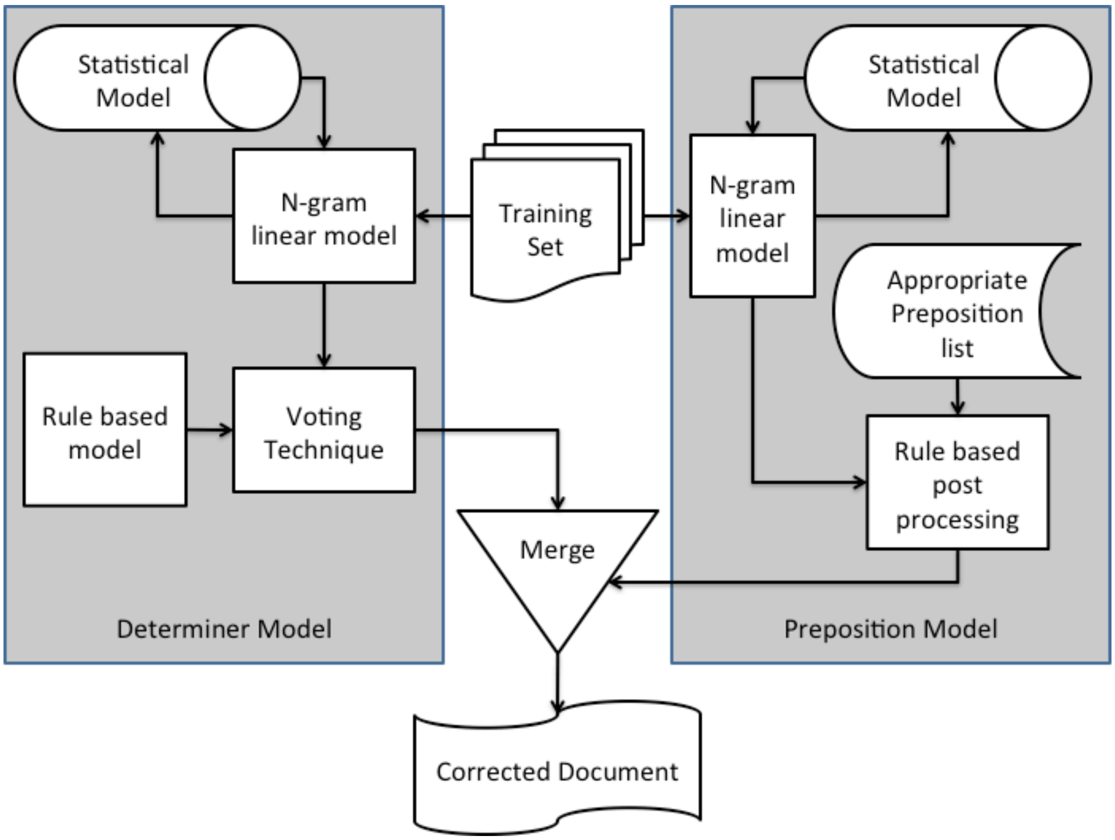

Figure 3. System Architecture

The text is parsed using the Stanford Dependency parser[1] to retrieve animate and inanimate information. After including animate and inanimate information the system didn't improve much as training data set is quite small and animate information is not correct for names. Hence, this feature is discarded from the statistical model.

4.1.2 Appropriate Preposition List

An appropriate preposition list consists of list of words along with preposition. The list is prepared in different corpus and training data. In the list, all possible formation with a word and preposition is stored. Let us take an example:

```
admit ST to SB
admit to
```

From corpus, two patterns for *admit* are found. Between admit and preposition something (ST) may come. Hence both of the entries are combined and formed in a regular expression format.

```
admit (ST)* to SB
```

4.1.3 Rule Based model for Preposition

Rule based post processing was applied on output of statistical model. For the rule based post processing, an appropriate preposition list was prepared manually. The list contains 1567 entries. The list is associated with animate and inanimate information. Hence, we aim to use dependency parser to identify subject object relation. Since the test data was in XML format, raw text was extracted from the XML document and the extracted sentences were parsed using Stanford dependency parser.

After parsing the document with the dependency parser, subject and object information was extracted. From all the sentences, proposition are detected and cross-validated with the appropriate preposition list. The preposition is dependent of the local association of the word around it. For the baseline

[1] http://nlp.stanford.edu/software/lex-parser.shtml

model, we have found that due to the list being small, few errors are being detected. Hence from the training corpus, the appropriate proposition list is enriched. The list is prepared in regular expression format. Here is an example:

```
ask * out +   invite on a date
```

In the above example, + means the two phrases have a similar meaning and * means one or more words can appear between the two words. Hence, when a match is found from the appropriate proposition list with the first word or the preposition, the words local to it are validated. Since the task is about correcting preposition errors, only words are matched with the list.

```
grateful to SB for ST
```

In the above example, *ST* means something or an object and *SB* means somebody or a subject, this information being retrieved from the dependency parser.

4.2 Determiner Error Detection

At the beginning of the determiner error detection task, we found that generation of list of rules to detect and correct the probable linguistic errors is a non-exhaustive set. Hence, we have decided to use a statistical model. After the statistical model, a rule based system is implemented with a few rules for *the* determiner devised from grammar books as for certain patterns statistical model fails to identify.

4.2.1 Statistical Model for Determiner

Similarly to preposition error detection, here a 5-gram linear statistical model is used. As same authors are prone to repeat same types of mistakes, we have decided to list out the errors from the training corpus documents. We have listed the errors document wise. In the training corpus, age information of author is mentioned. Hence documents are grouped according to age. After a close inspection of the document wise error list, the age group is prone to make similar type of errors, which depicts the attributes of the age group. Our statistical model is trained with every set of training data grouped by age separately. Hence different statistical models are prepared for different age groups. Now statistical model are applied according to the age group. It is found that age wise training incurred better result than single statistical model over whole data.

4.2.2 Rule Based Model for Determiner

It is found that statistical models works best for detecting the *a* and *an* determiner whereas performance drops for *the* determiner. Hence, rules for *the* are crafted manually from grammar books. A few rules for *a* and *an* are defined based on the first letter of the following word.
Among the determiners, usage of *the* is the most complicated one. For the rule based system different lists like nation, nationalities, unique objects, etc are produced. A few of the rules, which have been developed for the *the* determiner are mentioned below.

1. In most cases, if a sentence starts with a proper noun or common noun *the* is dropped.
2. Before a country name, *the* is dropped except if starts with *kingdom* or *republic*.
3. They system checks whether a common noun is appeared in a previous line of the document, i.e. it has already been referred to, in which case *the* is added.
4. If subject and object belong to same class i.e. they share the same hyponym class, *the* is added to the subject.
5. In case of superlatives like *best*, *worst* etc. *the* is added.
6. Before numerals, *the* is added.
7. Before unique things, *the* is added. Uniqueness is defined if a thing has single embodiment like *moon* etc.
8. It is found that if some geographical location is mentioned at a position other than start of sentence, *the* is added.

For different rules word lists are prepared such as a unique things list, superlatives, common nouns, country names, citizenships etc.

For *a* and *an* determiner correction, a list of different phonemes is prepared. Rule based system trims the first two characters and maps them into a phoneme to decide between *a* and *an*.

4.2.3 Voting Technique

The voting technique is used on the output of the rule based model and the statistical model. For *a* and *an* determiners, statistical model works best, especially in missing determiner and unnecessary determiner but for wrong determiner the rule based model performs better. For *the* determiner, the statistical model identified missing determiner and unnecessary determiner cases to some extent whereas list based rule-based system elevates the accuracy.

5 Evaluation

The system was evaluated for its performance in detecting, recognizing and correcting preposition and determiner errors in English documents. Separate scores were calculated for detection, recognition and correction for both the errors of preposition and determiner separately and then combined scores were also calculated. For all results, the organizer has provided three measures: Precision, Recall and F-Score. The precise definitions of these measures as implemented in the evaluation tool, and further details on the evaluation process are provided in (Dale and Narroway, 2012) and elaborated on at the HOO website.[4]

Each team was allowed to submit up to 10 separate runs over the test data, thus allowing them to have different configurations of their systems evaluated. Teams were asked to indicate whether they had used only publicly available data to train their systems, or whether they had made use of privately held data. We have submitted only one run (JU_run1) which has demonstrated F-scores of 7.1, 6.46 and 2.58 for detection, recognition and correction respectively before revision. And after revision it has demonstrated F-scores of 8.22, 7.59 and 3.16 for detection, recognition and correction respectively. Table 3 shows all the results of our run. We had used only publicly available data to train our systems, which are provided by the organizer as training set; we didn't use any privately held data.

6 Conclusion and Future Works

Our system has achieved F-scores of 8.22, 7.59 and 3.16 in detection, recognition and correction respectively. Our system failed to detect and correct many syntactic and semantic errors like wrong *a* determiner. Since the data consists of mostly mail conversation, it retains huge number of spelling mistakes, which misdirected the statistical, and rule based model to detect probable errors. For *the* determiner, if the size of the produced lists increases, better accuracy can be achieved with the rule-based system. Co-reference is another issue to identify, as *the* determiner is used mostly subsequent references. Anaphora resolution might therefore be of some help.

Element	**Task**	**Before Revision**			**After Revision**		
		Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Preposition	Detection	6.10	7.63	6.78	7.12	8.61	7.79
	Recognition	5.42	6.78	6.03	6.44	7.79	7.05
	Correction	3.05	3.81	3.39	3.73	4.51	4.08
Determiner	Detection	7.73	6.45	7.04	9.39	7.42	8.29
	Recognition	7.73	6.45	7.04	9.39	7.42	8.29
	Correction	1.66	1.38	1.51	2.21	1.75	1.95
Combined	Detection	6.93	7.28	**7.10**	8.19	8.25	**8.22**
	Recognition	6.30	6.62	**6.46**	7.56	7.61	**7.59**
	Correction	2.52	2.65	**2.58**	3.15	3.17	**3.16**

Table 3. Results for Preposition, Determiner and Combined (preposition and determiner) errors.

[4] See www.correcttext.org/hoo2012.

Acknowledgments

We acknowledge the support of the IFCPAR funded Indo-French project "An Advanced Platform for Question Answering Systems" and the DIT, Government of India funded project "Development of English to Indian Language Machine Translation (EILMT) System Phase II".

References

Adriane Boyd and Detmar Meurers. Data-Driven Correction of FunctionWords in Non-Native English. In 2011 Generation Challenges, HOO: Helping Our Own in the Proceedings of *the 13th European Workshop on Natural Language Generation* (ENLG), 28th – 30th September, 2011, Nancy, France.

Anas Elghafari, Detmar Meurers and Holger Wunsch, 2010. Exploring the Data-Driven Prediction of Prepositions in English. In the Proceedings of *the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010.

George Heidorn. 2000. Intelligent writing assistance. In R Dale, H Moisl, and H Somers, editors, *Handbook of Natural Language Processing*, pages 181–207. Marcel Dekker Inc.

GE Heidorn. 1975. Augmented phrase structure grammars. In: BL Webber, RC Schank, eds. *Theoretical Issues in Natural Language Processing*. Assoc. for Computational Linguistics, pp.1-5.

J R Tetreault and M S Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In Proceedings of *the 22nd International Conference on Computational Linguistics*, pp-865-872, Manchester,2008.

Pinaki Bhaskar, Aniruddha Ghosh, Santanu Pal and Sivaji Bandyopadhyay. May I correct the English of your paper!!!. In 2011 Generation Challenges, HOO: Helping Our Own in the Proceedings of *the 13th European Workshop on Natural Language Generation* (ENLG), pp 250-253, 28th – 30th September, 2011, Nancy, France.

Robert Dale and A Kilgarriff. 2010. Helping Our Own: Text massaging for computational linguistics as a new shared task. In Proceedings of *the 6th International Natural Language Generation Conference*, Dublin, Ireland, pages 261–266, 7th-9th July 2010.

Robert Dale and A Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In Proceedings of *the 13th European Workshop on Natural Language Generation* (ENLG), 28th – 30th September, 2011, Nancy, France.

Robert Dale and George Narroway. 2012. A framework for evaluating text correction. In Proceedings of *the Eighth International Conference on Language Resources and Evaluation* (LREC2012), 21–27 May 2012.

Robert Dale, Ilya Anisimoff and George Narroway (2012) HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In Proceedings of *the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada, 7th June 2012.