May I check the English of your paper!!!

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

Abstract

This paper reports about our work in the HOO shared task 2011. The task is to automatically correct the English of a given document. For that, we have developed a hybrid system of a statistical CRF based model along with a rule-based technique has been used. The system has been trained on the HOO shared task training datasets and run on the test set given by the organizer of HOO. We have submitted one run, which has been demonstrated F-score of 0.204, 0.178 and 0.167 for detection, recognition and correction respectively.

1 Introduction

Writing the research papers or thesis in English is a very challenging task for those researcher and scientist whose first language or mother tongue is not English. Express their research works properly in English is a hard job for them. Generally their paper, which is submitted to a conference and may be rejected not because of their research works but because of the English writing, which makes the paper harder for the reviewer to understand intention 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 [1] 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, 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 it can be accepted. All though the native English speakers are also be helped by this tools and techniques. This task is simply expressed as a text-to-text generation or Natural language Generation (NLG).

For this shared task, HOO, we have developed two models, one is rule-based model and another is statistical model. Then we have combined both these models and developed our system for HOO, 2011.

2 Related Works

English Language belongs to the Germanic languages branch of the Indo-European language family, widely spoken on six continents. HOO shared task is organized to help authors with the writing tasks. Identifying grammatical and linguistic errors in a text of a language is an open challenge to the researchers. In recent times, researchers (Heidorn, 2000) have acquired 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.

[1] http://acl2010.org/mentoring.htm

al., 2008, has dealt with error pattern with preposition by non-native speakers.

3 System Description

At the beginning of the work, we found that generation of list of rules to detect and correct the probable linguistic errors is a non-exhaustive set. So we have decided to list out the errors from the training corpus documents. We have listed the errors document wise. After a close inspection of the document wise error list, the author is prone to make similar type of errors, which depicts the attributes of the author. The errors types are classified in to some coarse groups like wrong form, something missing, needs replacing etc. We decided to resolve the errors at different levels like POS level, syntax level and semantic level. Our system contains two models – a rule based model and a statistical model as described in the next sections.

3.1 Rule based model

The total corpus is first checked using conventional grammar tool and spell checkers. The data set is parsed using Stanford dependency parser[2]. While detecting and correcting errors, we have considered the coarse groups one by one.

Wrong Form Preposition (FT) & Needs replacing Preposition (RT): To detect and to correct the wrong forms of preposition we have used a list of devised manually appropriate preposition list. Certain cases are solved based only syntax though in many cases we have to check the semantics. To identify the semantics we have used output of Stanford dependency parser and part-of-speech(POS).E.g. after verb "create", "by" preposition is used if an object follows the verb.

Wrong Form verb (FV): To detect the wrong forms of the verb we have used a verb paradigm table, which will help also in suggesting appropriate verbal inflection.

Wrong Form determiner (FD): To detect the wrong forms of determiner we have used the conventional spell checker system.

Wrong Form Adverb(FY)&& Wrong Form Adjective (FJ): To detect the wrong form of adverbs and adjectives, we have used positional aspect. Adverbs appear around the verbs, in most cases after the verbs whereas adjective appears around nouns, in most cases before the noun. A dictionary-based approach is implemented to correct the wrong forms of adverbs and adjectives.

Needs replacing conjunction (RC) & Needs replacing punctuation (RP): In case of serial comma, the last comma is replaced with "and".

Unnecessary punctuation (UP): In case of serial comma, if last comma is followed by "and" then that punctuation is treated as an error. Though it is an optional correction due to debate over serial comma issue, it is one of most frequent errors in the corpus.

Countability of noun errors(CN)and wrong quantifier because of noun countability(CQ): Countability errors are detected by the conventional grammar tools. For both these type of errors, we have considered agreement of quantifier, noun countability and verb of the sentence. Among these three, if two of them agree then the other one is corrected. As example,

"multiple error is found in the text".

In the above example, as "is" and "error" have same agreement over countability "multiple" will be corrected to "single".

Verb agreement error (AGV): To detect verb agreement error we have identified the subject using dependency parsing. We have detected the error using verb paradigm table.

The missing coarse group is the one of the bigger challenge of this task. Deriving rules for this missing coarse group needs a lot of in depth study. Few rules have been devised though in certain cases those corrections are optional. Few syntactic rules can be generated

Missing preposition (MT): For missing preposition we have used the appropriate preposition list but it wasn't enough to detect. We have devised some handcrafted rules based on linguistic features.

i. After the occurrence of "all", it might be followed by "of" and sometimes an article after "of".

ii. If there is a connecting word pair like "not only" and "but" then if either of them is followed and preposition then other one will also be followed by same preposition.

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

iii. A pronoun can't be used following number. There should be a preposition among them, mostly "of".

3.2 3.2 Statistical model

Devising rules for the appropriate determiner before nouns is quite difficult. Hence we decided to use a sequence labeling based statistical tool named Conditional Random Field (CRF++). For training, we have marked determiner along with the two words following the determiner in the training corpus. For better accuracy of the statistical model, a large data set is required for learning. Hence we have used our published papers for the training of the statistical model. If a preposition precedes the determiner then the determiner is also marked. As features to the statistical system, we have used word, root form, POS tag, number marker (singular/plural/null) and word position. The statistical tool is trained using the training corpus and it used tri-gram model.

3.3 Post Correction

After intense analysis, depending on the nature of errors in the output of statistical system we developed a set of rules.

i. In certain cases where the words are marked, we search for a gerund or noun after the marked word. If words are occurring for the first time in the paragraph then those cases are ignored.

ii. If there is gerund or noun after marked words and that gerund or noun has appeared before in the paragraph then "the" determiner is inserted before the marked word.

iii. If there is gerund or noun after marked words and that gerund or noun has appeared before in the paragraph and "a" determiner is present before the marked word then it will replaced with "the".

3.4 Merging output

The rule-based model identifies various errors based on linguistic syntactic and semantic features. The statistical model identified the missing determiner errors and wrong determiner errors. The post correction corrects the missing determiner error and wrong determiner errors detected by statistical parser. The output of the rule based model and the statistical model are merged to produce the final output. The block diagram is shown in the figure 1.

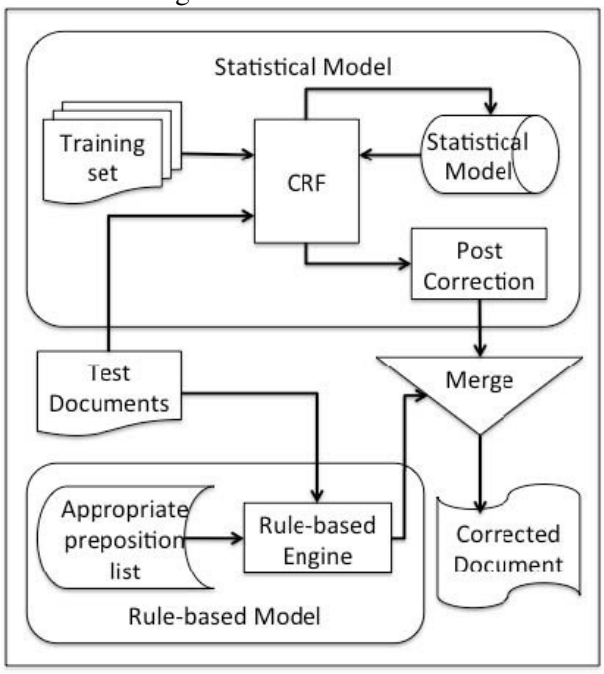

Figure 1:

4 Experimental Results

This paper reports about our research work as a part of HOO shared task. We have used a hybrid system consisting of a rule-based model and a statistical model followed by a post-processing. We have achieved F-score of 0.204, 0.178 and 0.167 in detection, recognition and correction respectively.

5 Conclusion

Our system has posed an accuracy of F-score 0.204, 0.178 and 0.167 in detection, recognition and correction respectively. Our system failed to detect and correct many syntactic and semantic errors like wrong "a" determiner. One error can be assigned with multiple tags. Hence deciding the appropriate tag is still an open debate.

Acknowledgments

The work has been carried out with support from Indo-French Centre for the Promotion of Advanced Research (IFCPAR) funded Project "An Advanced Platform for Question Answering System" (Project No. 4200-IT-1).

References

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.

Dale and Kilgarriff, 2010. Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task, International Natural Language Generation Conference 2010, Dublin, Ireland.

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.

http://en.wikipedia.org/wiki/Serial_comma

Strauss Jane, The Blue Book of Grammar and Punctuation, 10thEdition.