Handling Outlandish Occurrences: Using Rules and Lexicons for Correcting NLP Articles

Elitza Ivanova Delphine Bernhard Cyril Grouin
LIMSI–CNRS, BP133, F-91403 Orsay Cedex, France
elitza.ivanova@knights.ucf.edu
dbernhard@unistra.fr, cyril.grouin@limsi.fr

Abstract

This article describes the experiments we performed during our participation in the HOO Challenge. We present the adaption we made on two systems, mainly designing new grammatical rules and completing a lexicon. We focused our work on some of the most common errors in the corpus: missing punctuation and inaccurate prepositions. Our best experiment achieved a 0.1097 detection score, a 0.0820 recognition score, and a 0.0557 correction score on the test corpus.

1 Introduction

The number of articles written by non-native English speakers makes it necessary to provide the community with tools that can be helpful in checking and improving the linguistic quality of those articles (Dale and Kilgarriff, 2010).

The correction of errors made by English as a Second Language (ESL) writers has been addressed in several recent studies. Different kinds of errors are targeted, both concerning closed classes of words such as articles, prepositions, modals or auxiliaries and open classes of words, such as nouns and verbs (Lee and Seneff, 2006; Felice and Pulman, 2008; Gamon et al., 2009; Rozovskaya and Roth, 2011). In the case of closed classes and commonly confused words, it is possible to cast the problem as an automatic classification task. The goal of the classifier is to predict the most likely candidate from a confusion set in the given context. This requires large training corpora of mostly error-free texts.

Another approach to error correction consists in using manually developed rules to identify and correct erroneous occurrences. This approach has, for instance, been adopted in the open-source LanguageTool proofreading tool[1] (Naber, 2003; Miłkowski, 2010).

In this paper, we describe our participation to the HOO2011 challenge. We present our systems and the configurations we used while participating in the test stage of the challenge.

2 Material and methods

2.1 Corpus

Over a total amount of 1,264 annotated errors in the training corpus, we noticed that the most common errors are of three types: a missing punctuation (16.6%), a missing determiner (12.7%), and a preposition to be replaced (8.6%). Each other type of errors accounts for less than 5% of all errors in the corpus.

2.2 Systems

As the training corpus is only composed of 19 annotated files, we decided not to use machine-learning based approaches. Moreover, as we are non-native English speakers, finding and annotating English errors in scientific papers would have been a hard task.

2.2.1 Language Tool

Our first system consists of an extension of the LanguageTool system, as it has not been developed

[1] http://www.languagetool.org/

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 254–256,
Nancy, France, September 2011.

```
<rule default="on" id="NEED_TO" name="need to">
    <pattern case_sensitive="no" mark_from="1">
        <token inflected="yes" postag="NN.*" postag_regexp="yes">need</token>
        <token postag="IN"><exception>to</exception></token>
        <token postag="VBG" postag_regexp="yes"/>
    </pattern>
    <message>Incorrect use of the preposition '\2' after '\1'. Normally, <suggestion>to <match no="3"
postag="VB"/></suggestion> is used.</message>
    <short>Wrong choice of preposition</short>
    <example correction="to seek" type="incorrect">I wish to stress the need <marker>of seeking</marker> a positive
outcome.</example>
    <example type="correct">I wish to stress the need to seek a positive outcome</example>
</rule>
```

Figure 1: Example LanguageTool XML rule.

specifically for text written by ESL writers. The system is based on linguistic resources and rules described in XML files that can be easily extended. We modified three resource files to deal with the HOO corpus: the grammar rules used to process the corrections, the compound words lexicon that lists the words that must be written with a dash, and the list of words that require "an" instead of "a" as a determiner, even though they do not begin with a vowel.

Figure 1 displays an example of an XML rule which deals with incorrect prepositions after the noun "need".

2.2.2 Commas module

In order to deal specifically with missing commas in figures larger than 1,000, we wrote an independent Python module.

2.2.3 CCAC

The second system[2] we used has been designed to perform both analyses of the quality, and spelling and grammatical correction of survey corpora and web content (Grouin, 2008). The final objective of this tool was to help indicate whether that noisy data could be used in an NLP chain of treatments to be applied further or not. This system is mainly based on unigrams of words and typographic rules. We adapted this system to English by producing a new lexicon of 19,000 unigrams of words from the *Financial Times* which we completed with 300 computational terms from the ACL corpus. This lexicon also includes the American version of British words.

[2]CCAC: Corpus Certification and Automatic Correction.

3 Experimental setup

We defined ten configurations based on several combinations of each system's parameters:

- Run 0: LanguageTool as it is from download;
- Run 1: LanguageTool with new rules;
- Run 2: As in run #1 plus commas module;
- Run 3: Run #0 plus new compounds lexicon;
- Run 4: Run #1 plus new compounds lexicon;
- Run 5: Run #4 plus commas module;
- Run 6: CCAC system;
- Run 7: CCAC system followed by run #5;
- Run 8: Run #5 followed by the CCAC system;
- Run 9: LanguageTool with punctuation correction only plus commas module.

4 Evaluation and discussion

The evaluation of our pipeline on the test corpus is given in Table 1. We achieved our best results using the combination of LanguageTool followed by CCAC (run #8); we obtained a 0.1097 detection score, a 0.0833 recognition score, and a 0.0589 correction score, without any bonus (Dale and Kilgarriff, 2011).

The CCAC system used independently did not obtain good results (#6). This system has been designed to process very noisy data using basic correction modules (to add or to remove diacritics, to process geminates, and at last to propose corrections based on the Levenshtein distance). Within the framework of the HOO challenge, the corrections to be made are finer than those of a web corpus.

While on the training data we achieved our best

Table 1: Official evaluation on the test corpus (no bonus scores)

Run	Det P	Det R	Det S	Rec P	Rec R	Rec S	Cor P	Cor R	Cor S
0	0.7143	0.0095	0.0187	0.7143	0.0095	0.0187	0.4286	0.0057	0.0112
1	0.4861	0.0331	0.0620	0.4085	0.0274	0.0514	0.2958	0.0199	0.0372
2	0.4868	0.0350	0.0653	0.4133	0.0293	0.0548	0.3067	0.0218	0.0406
3	0.5758	0.0180	0.0349	0.3333	0.0104	0.0202	0.2121	0.0066	0.0128
4	0.4835	0.0416	0.0767	0.3333	0.0284	0.0523	0.2444	0.0208	0.0384
5	0.4842	0.0435	0.0797	0.3404	0.0303	0.0556	0.2553	0.0227	0.0417
6	0.3056	0.0208	0.0390	0.2778	0.0189	0.0354	0.1528	0.0104	0.0195
7	0.4063	0.0615	0.1068	0.3019	0.0454	0.0789	0.2013	0.0303	0.0526
8	0.4085	0.0634	**0.1097**	0.3067	0.0473	**0.0820**	0.2086	0.0322	**0.0557**
9	0.4510	0.0218	0.0415	0.2745	0.0132	0.0253	0.2353	0.0114	0.0217

score using LanguageTool only,[3] on the test corpus, the combination of both LanguageTool and CCAC performed best. This demonstrates the complementarity of both tools when applied on a new corpus for which no specific rules had been designed.

For the time being, our systems only deal with some types of errors (especially punctuation and prepositions), due to time constraints for developing new resources and tools. Further work is thus needed to process all other kinds of errors. When improving the LanguageTool resources, we manually designed new rules and added new items in the lexicons. In order to improve this process, it would be interesting to automatically extract rules and missing words from the annotated corpus in order to reduce human intervention.

Acknowledgments

This work was partly realized as part of the Quaero Programme, funded by Oseo, French State agency for innovation. We thank the organizers for their work and reactivity during this challenge. We enjoyed being involved in this very fun task, even more so because such a correction tool would be useful for us non-native English speakers!

References

Robert Dale and Adam Kilgarriff. 2010. Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task. In *International Natural Language Generation Conference Proceedings*, pages 261–266, Dublin, Ireland.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proc. of ENLG*, Nancy, France.

Rachele De Felice and Stephen G. Pulman. 2008. A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proc. of Coling*, pages 169–176, Manchester, UK, August.

Michael Gamon, Claudia Leacock, Chris Brockett, William B. Dolan, Jianfeng Gao, Dmitriy Belenko, and Alexandre Klementiev. 2009. Using Statistical Techniques and Web Search to Correct ESL Errors. *Calico Journal*, 26(3).

Cyril Grouin. 2008. Certification and Cleaning-up of a Text Corpus: towards an Evaluation of the "grammatical" Quality of a Corpus. In *Proc. of LREC*, pages 1083–1090, Marrakech, Morocco.

John Lee and Stephanie Seneff. 2006. Automatic Grammar Correction for Second-Language Learners. In *Proc. of InterSpeech*, pages 1978–1981.

Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software – Practice and Experience*, 40:543–566.

Daniel Naber. 2003. A Rule-Based Style and Grammar Checker. Master's thesis, Technische Fakultät, Universität Bielefeld.

Alla Rozovskaya and Dan Roth. 2011. Algorithm Selection and Model Adaptation for ESL Correction Tasks. In *Proc. of the 49th Annual Meeting of ACL*, pages 924–933, Portland, Oregon.

[3] We obtained a 0.3487 detection, 0.2995 recongition and 0.2947 correction scores (with bonus) on run 5 and a 0.3485 detection, 0.2969 recognition and 0.2925 correction scores (with bonus) on run 8 on training corpus.