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

Elitza Ivanova, Delphine Bernhard and Cyril Grouin

LIMSI-CNRS, Orsay, France

1. Introduction

Approach: re-use and extend existing systems:

- Focus on grammatical errors and punctuation errors: rule-based system (LanguageTool)
- Focus on spelling errors: lexicon-based correction (CCAC)

2. Corpus

- 1,264 annotated errors in the training corpus
- Most common errors types:
  - a missing punctuation (16.6%),
  - a missing determiner (12.7%),
  - a preposition to be replaced (8.6%).

  $\rightarrow$ we only focused on these three kinds of errors.
- Each other type of errors accounts for less than 5% of all errors.

3. LanguageTool

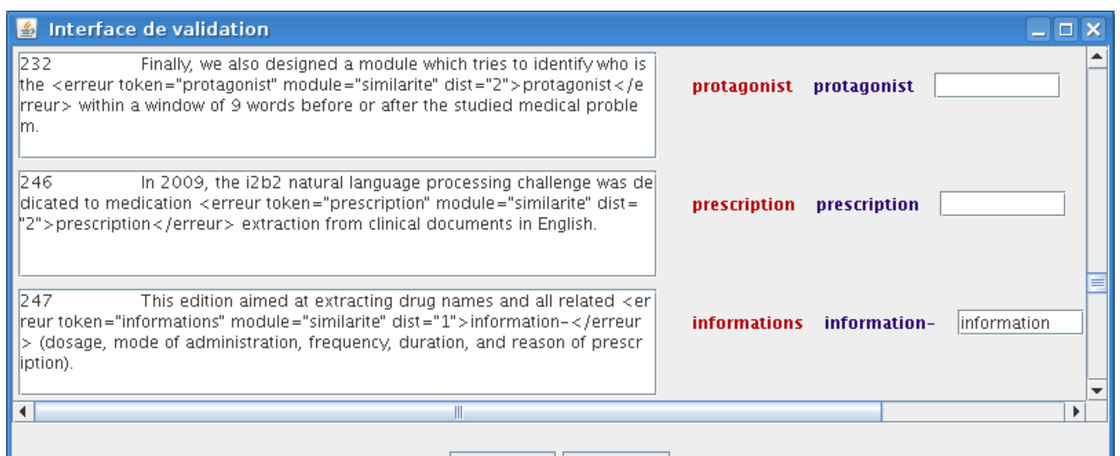
- Proofreading tool (`www.languagetool.org`) [Naber2003, Miłkowski2010].
- Modification of three resource files to deal with the HOO corpus:
  - grammar rules used to process the corrections;

```
<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>
```

  - compound words lexicon that lists the words written with a dash; graph-based, lexico-semantic, pair-wise, wide-coverage, etc.
  - list of words that require “an” instead of “a” as a determiner, even though they do not begin with a vowel. n-gram, ngram
- Module created to deal with missing commas in figures $>$ 1,000.

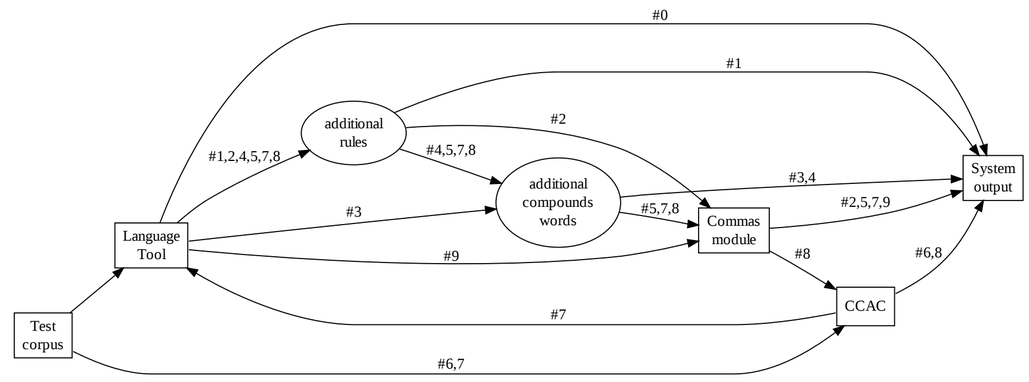
4. CCAC – Corpus Certification and Automatic Correction

- System designed to process survey corpora and web content [Grouin2008]: analyses of quality, spelling and grammatical correction.
- Adaptation to English:
  1. lexicon of 19,000 unigrams of words produced from the *Financial Times*;
  2. 300 additional computational terms from the ACL corpus (also includes the American version of British words).
- A graphic interface allows the user to check and eventually to correct erroneous corrections from CCAC (not used during the HOO test so as to avoid any human intervention):



5. Experimental Setup

- Ten configurations based on several combinations of each system’s parameters:



6. Evaluation

Table: 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	0.714	0.010	0.019	0.714	0.010	0.019	0.429	0.006	0.011
1	0.486	0.033	0.062	0.409	0.027	0.051	0.296	0.020	0.037
2	0.487	0.035	0.065	0.413	0.029	0.055	0.307	0.022	0.041
3	0.576	0.018	0.035	0.333	0.010	0.020	0.212	0.007	0.013
4	0.484	0.042	0.077	0.333	0.028	0.052	0.244	0.021	0.038
5	0.484	0.044	0.080	0.340	0.030	0.056	0.255	0.023	0.042
6	0.306	0.021	0.039	0.278	0.019	0.035	0.153	0.010	0.020
7	0.406	0.062	0.107	0.302	0.045	0.079	0.201	0.030	0.053
8	0.409	0.063	**0.110**	0.307	0.047	**0.082**	0.209	0.032	**0.056**
9	0.451	0.022	0.042	0.275	0.013	0.025	0.235	0.011	0.022

7. Discussion

- Best score on the training data using LanguageTool only;
- Best results on the test corpus using the combination of LanguageTool followed by CCAC (run #8). This demonstrates the complementarity of both tools when applied on a new corpus for which no specific rules had been designed.
- The CCAC system alone did not obtain good results (#6):
  - system designed to process very noisy data using basic correction modules;
  - the corrections to be made are finer in the HOO challenge than those of a web corpus.
- 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 needed to process all kinds of errors.
  - **Perspective:** automatically extract rules and missing words from the annotated corpus in order to reduce human intervention.

References

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.

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.

Groupe Information, Langue Ecrite et Signée (ILES) - Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur - Orsay, France
WWW: `http://www.limsi.fr`

Acknowledgements: This work has been partially financed by OSEO under the Quæro program.