An evaluation of *Avalingua* based on learner corpora

Pablo Gamallo, Marcos García
CITIUS, University of Santiago de Compostela
Isaac González, Marta Muñoz, Iria del Río
Cilenis S.L
info@cilenis.com

Abstract

This article describes the use of two learner corpora to evaluate precision and recall of a linguistic tool, called *Avalingua*, based on Natural Language Processing techniques. Avalingua is a linguistic software aimed at automatically identifying and classifying spelling, lexical, and grammatical errors in written language. The objective of the tool is to analyse the linguistic errors of texts, assess the writing proficiency, and propose solutions with the aim of improving the linguistic skills of students. It makes use of natural language processing and knowledge-rich linguistic resources. It has been developed, so far, for Galician language, and can be applied to both L1 learners of Galician (i.e., language acquisition in children) and L2 language acquisition.

Avalingua for Galician consists of the following modules:

- Identification of spelling and lexical errors, which is constituted by two submodules: a state of the art spelling checker, and a classifier of usual lexical errors: e.g., Spanish and Portuguese expressions, archaisms, anglicisms, hyper-authentic terms, etc.
- Identification of integers, tokens with special characters, and proper names, which must not be considered as incorrect expressions even if they are not found in a reference (gold standard) vocabulary of Galician language.
- Identification of tokens constituted by productive affixes, and frequent Out Of Vocabulary words. Even if they are not contained in the reference vocabulary, they cannot be considered as lexical errors.
- Identification of fake friends Galician-Spanish, that is, words appearing in the reference vocabulary with very marginal senses, but which are also very common Spanish words.
- Identification of different types of grammatical errors by using a dependency based syntactic parser.
- Language identification to find citations or quotations in other languages within the text.

To evaluate the performance of *Avalingua*, we made use of two learner corpora. The first one is a collection of writing texts belonging to Galician children in the 3th year of secondary school (3th of ESO). The second one consists of texts written by adult Portuguese L2 learners of Galician language. Precision was computed by counting the number of

correct decision made by the system divided by the total number of decisions it made. Recall is the number of correct decisions made by the system divided by the total number of errors (spelling, lexical, and grammatical errors) found in the text. Avalingua achieved 93% precision and 66% recall when it was evaluated using the first corpus (Galician children), and 90% precision using the second corpus (Portuguese L2 learners). The high quality performance achieved in the two tests shows that the system can be viewed as an useful tool to help teachers to assess writing proficiency of students, not only for L1 acquisition, but also for L2 learning. The article will provide a more accurate analysis of the evaluation results, by taking into account issues such as frequent lexical and grammatical mistakes, texts with more error rate, and so on.

Given the modular structure of the system, it is possible to adapt Avalingua to other languages, such as English. In the last two years, there has been an increased interest in developing automatic tools for identifying and correcting both spelling and grammatical errors in texts written by English learners. This growi ng interest has led researchers to organize two international competitions, name ly HOO-20121[1] and CoNLL-20132[2] shared task, aimed at comparing the efficiency of different systems trained on a large collection of texts written by English students. Those learner corpora are completely annotated with error tags and correcti ons, and all annotations have been performed by professional English instructors . In future work, our objective is to apply Avalingua on English texts, by using as source of errors learner corpora available from HOO-2012 and CoNLL-2013. For this purpose, we will analyse the more frequent grammatical errors found in tho se copora and define appropriate correction rules with DepPattern.

References:

Dahlmeier, D. and Tou Ng, H. (2011) “Grammatical Error Correction with Alternating Structure Optimization ”, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, p. 915-923.

Dale, R. and Kilgarriff, A. (2010) “Helping Our Own: Text Massagin for Computational Linguistics as a New Shared Task”, *Proceedings of the 6th International Natural Language Generation Conference (NLG'10)*, p. 263-267.

Han, N., Tetreault, J.R., Lee, S., Ha, J. (2010) “Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System”, *Proceedings of the Language, Resources, and Evaluation Conference (LREC-10).*

Leacock, C., Chodorow, M., Gamon, M. and Tetreault J. (2010) “Automated Grammatical Error Detection for Language Learners ”, *Morgan & Claypool Publishers*, San Rafael, CA.

1 http://clt.mq.edu.au/research/projects/hoo/

2 http://www.comp.nus.edu.sg/~nlp/conll13st.html