G. Biswas et al. (Eds.) (2012). Proceedings of the 20th International Conference on Computers in Education. Singapore: Asia-Pacific Society for Computers in Education

Corpus-based online common error detector

John BLAKE
Institute of General Education,
Japan Advanced Institute of Science and Technology,
Japan
johnb@jaist.ac.jp

Abstract: In this paper the theoretical underpinnings and the practical construction of a corpus-based online common error detector are described. The selection of errors for inclusion was based on analysis of published accounts of common errors in the writing of Japanese learners of English, which were confirmed through statistical corpus analysis. Harnessing the concept of regular expressions and utilizing a MySQL database, the online error detector provides instant feedback on common errors in academic discourse that standard spell and grammar checker programs do not detect.

Keywords: error detection, Japanese learners, research articles, learner corpora

1. Introduction

The *raison d'être* for this common error detector is to provide automated feedback on drafts of research articles and abstracts written by Japanese researchers. The feedback is designed to provide easy-to-understand actionable advice that will improve the grammatical accuracy and formality of their drafts.

The inspiration for this project was the Common Error Detector, a JavaScript program created by Andy Morrall [1] at the Hong Kong Polytechnic University. His program searches for common errors made by Chinese speakers of English in academic essays using regular expressions (regexp).

The novelty of our error detector is twofold: firstly, the data set is tailored for Japanese researchers; and secondly, the error database is based on an extensive literature review and statistical analysis of large collections of texts (corpora). Our regexp data set includes phrases and words that are commonly misused, such as the noun *researches*. If this regexp is matched to the submitted text, the feedback message *research is an uncountable noun* is displayed. What is particularly useful about this detector is that it finds errors that standard spelling and grammar checkers cannot. This is first documented online error detector tailored for Japanese writers of English research documents.

2. Project overview

The project was divided into two key phases, namely: preparation and creation. The preparation phase involved reviewing the literature on error analysis and learner corpora, and designing the technical specifications. The creation phase comprised the creation of the web interface and a MySQL database, data input and analysis of the efficacy and accuracy of regexp using various corpora.

3. Preparation phase

3.1 Error analysis

Having analyzed a corpus of 2 million words, Izumi *et al* [2] noted the three most common errors in the spoken English of Japanese learners were related to articles (e.g. *the*), number (e.g. *–s*) and prepositions (e.g. *in*). These were closely followed by a variety of verb errors. Focusing on research articles drafted by Japanese academics, Orr and Yamazaki [3] proposed a set of twenty common problems. This was based on their analysis of a corpus of approximately 200,000 words of academic text written in English by Japanese researchers. The problems they identified are classified in Table 1. Regexp can be used to identify many of these errors at phrase level; but, with a few exceptions, cannot identify discourse-level errors.

Table 1: Twenty problems frequently found in research articles authored by Japanese

Noun phrase	**Verb phrase**	**Discourse**	
articles number quantification prepositions* titles and labels	copula - be modality* transitivity time, tense & aspect voice phrasal verbs	collocations summarizing discussion communication authenticity	authorship consistency citation relevance lexical richness density & complexity

* placed in these categories for pedagogical purposes

3.2 Learner corpora

A two-fold approach was adopted. First, having assessed the available corpora of Japanese learners of English, the Japanese sub-corpora of around 170,000 words from the 2010 beta Corpus of English Essays written by Asian University Students [4] was selected as being the most relevant. Second, the variables of the ideal corpus were established [5]. Following those parameters, a preliminary corpus of 250,000 words was created using texts from research articles in the fields of information, materials and knowledge science. All the texts were written in English and published in domestic conference proceedings or journals by Japanese researchers.

3.3 Technical details

A MySQL database, housing a live and a test version, is hosted on the university server. A student input interface was designed in which users submit text that is searched for regexp and feedback given for each error that is discovered within the error set. A subsidiary aim is also to harvest the submitted text and add it to a learner corpus which can be accessed when checking the efficacy of regexp. A teacher input interface was also needed to input the regexp and associated data as well as accessing both the live and test versions.

4. Creation phase

4.1 Creation of interface and database

The beta version of the database and web-interfaces was created. A website designer was consulted to identify ways to make the student interface more user-friendly.

4.2 Data input

Common errors made by Japanese learners of English were collated from four published sources [6] [7] [8] [9]. These were inputted into the database. These errors were further subdivided, and regular expressions were identified for each case.

After testing, feedback was worded for each error and trialed with focus groups of Japanese researchers. This gave rise to numerous issues. For example, the initial example of *researches* may actually be correct when *research* is used as a verb in the present simple tense with a third-person subject, but in research documents the noun is much more frequently used. Feedback messages were assigned a priority, namely: warning or advice. Warnings include those regexp which would be incorrect in all situations, such as *these researches.* Advice includes regexp that would often be incorrect such as *lots of*, which is rather informal for research documents, but in some situations could be appropriate.

4.3 Statistical analysis

The regular expressions were systematically tested on the Japanese learner corpora using Antconc3.2.4w [10], a concordance program. Items that were not able to identify errors accurately were discarded from the database.

5. Further development

To enable more refined searches of errors, we plan to assigned parts of speech to both the submitted text and the preliminary corpus using a part-of-speech tag set and an automatic tagger, which will enable word-category disambiguation.

References

[1] Morral, A. (2010). Common Error Detector. <http://www2.elc.polyu.edu.hk/cill/errordetector.htm>

[2] Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T. and Isahara, H. (2003). Automatic Error Detection in the Japanese Learners' English Spoken Data, *Procedings of ACL,* Sapporo, Japan.

[3] Orr, T., and Yamazaki, A. K. (2004, Sept 24 – Oct 1). Twenty problems frequently found in English research papers authored by Japanese researchers. *(pp. 23-35), Professional Communication Conference Proceedings, International, 2004.* doi. 10.1109/IPCC.2004.1375270

[4] Ishikawa, S. (2008). *Eigo koupasu to eigo kyouiku: deeta toshiteno tekusuto. [Engish corpus and language education: Text as data]*. Tokyo: Taishukanshoten.

[5] Atkins, S., Clear, J., and Ostler, N. (1992). Corpus design criteria, *Literary and Linguistic Computing, 7:* 1-16.

[6] Thompson, I. (2001). Japanese speakers. In M. Swan and B. Smith (Eds.) *Learner English: A teacher`s guide to interference and other problems.* Cambridge: Cambridge University.

[7] Barker, D. (2003). *Eigo to Nakanaori Dekiru Hon* [The book for becoming friends again with English]. Tokyo: Aruku.

[8] Webb, J. H. M. (2006). *151 common mistakes of Japanese students of English*. Tokyo: The Japan Times.

[9] Barker, D. (2008). *An A - Z of Common English Errors for Japanese Learners (Japanese version)*. Tokyo: BTB Press.

[10] Anthony, L. (2012). *AntConc (Version 3.2.4) [Computer Software]*. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/