ISSN: 1391-8796
Proceedings of
5th Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
February 15, 2018

Detecting and correcting real-word errors in Tamil sentences

Sakuntharaj R.[*1] and Mahesan S.[2]

[1]*Centre for Information and Communication Technology, Eastern University, Sri Lanka.*
[2]*Department of Computer Science, Faculty of Science, University of Jaffna, Sri Lanka*

The spell checker concerns the two types of errors namely *non-word errors* and *real-word errors*. Non-word errors can be of two categories. First one is that the word itself is invalid. The other is that the word is valid but not present in a valid lexicon. Real-word error means the word is valid but inappropriate in the context of the sentence. An approach to correcting real-word errors in Tamil language is proposed in this paper. A *bigram probability model* is constructed to determine appropriateness of the valid word in the context of the sentence using a 3GB volume of corpora of Tamil text. In case of lacking appropriateness, the word is marked as a real-word error and *minimum edit distance* technique is used to find lexically similar words, and the appropriateness of such words is measured by a *word-level bigram language probability model.* A hash table with word-length as the key is used to speed up the search for words to check for the lexical similarity. Words of lengths of *m*-1 to *m*+1 are considered with *m* being the length of the word found to be 'inappropriate'. Finally, top five words are selected as suggestion for correction. Test results show that the suggestions generated by the system are with 98% accuracy as approved by a Scholar in Tamil.

This technique may be used to check real word errors in other languages too with sufficient corpus to build the bigram probability model for the language.

Keywords: Tamil, Real-word error, Bigram, Minimum edit distance, Error correction

**Corresponding Author: mahesans@univ.jfn.ac.lk*