Abstract 97

A Method to Sort Official Correspondence through Natural Language Processing

Tharindu Weerasooriya[1], Nandula Perera[2]

[1] Department of Statistics and Computer Science, University of Kelaniya, Sri Lanka
cyriltcw@gmail.com
[2] Department of English, University of Kelaniya, Sri Lanka
mcnandula@gmail.com

Natural language Processing (NLP) is a new branch of study in Computational Linguistics and the field has undergone rapid development over the past few decades. Keyword extraction is a popular application of NLP. The present study makes use of Stanford Core NLP, an NLP tool that enables Parts-of-Speech (POS) tagging in order to extract the keywords from official correspondence. POS tagging identifies all the parts of speech in a sentence and categorises them into the relevant grammatical categories. Capitalising on the grammatical uniformity of formal written English, the system is able to identify all the noun phrases and verb phrases of a sentence. Hence, the subject and the predicate of the sentence are isolated. Document sorting with regard to official correspondence is done through the system by analysing the 'object' line of an official letter or the 'subject' line of an e-mail, and listing the noun phrases and verb phrases. The document is then sorted to the relevant department. In order to prevent slips in the system, the remaining words of the 'object' / 'subject' lines are filtered through a keyword corpus. This increases the accuracy of the keyword extraction process. The present system proved to be more efficient that selecting keywords through a filter, as the POS tagging sorts and presents keywords in an order where the respondents are able to grasp the main idea of the sentence. The subsidiary list of words extracted through the key word corpus adds to the accuracy of the system. The present study is only limited to official correspondence in English. It could be modified to be adapted to other languages.

Key words: document sorting, keyword extraction, natural language processing, official correspondence, part-of-speech tagging