World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:11, No:02, 2017

Extraction of Compound Words in Malay Sentences Using Linguistic and Statistical Approaches

Authors : Zamri Abu Bakar Zamri, Normaly Kamal Ismail Normaly, Mohd Izani Mohamed Rawi Izani

Abstract : Malay noun compound are phrases that consist of two or more nouns. The key characteristic behind noun compounds lies on its frequent occurrences within the text. Therefore, extracting these noun compounds is essential for several domains of research such as Information Retrieval, Sentiment Analysis and Question Answering. Many research efforts have been proposed in terms of extracting Malay noun compounds using linguistic and statistical approaches. Most of the existing methods have concentrated on the extraction of bi-gram noun+noun compound. However, extracting noun+verb, noun+adjective and noun+prepositional is challenging due to the difficulty of selecting an appropriate method with effective results. Thus, there is still room for improvement in terms of enhancing the effectiveness of compound word extraction. Therefore, this study proposed a combination of linguistic approach and statistical measures in order to enhance the extraction of compound words. Several preprocessing steps are involved including normalization, tokenization, and stemming. The linguistic approach that has been used in this study is Part-of-Speech (POS) tagging. In addition, a new linguistic pattern for named entities has been utilized using a list of Malays named entities in order to enhance the linguistic approach in terms of noun compound recognition. The proposed statistical measures consists of NC-value, NTC-value and NLC value.

Keywords : Compound Word, Noun Compound, Linguistic Approach, Statistical Approach
Conference Title : ICEMNLP 2017 : International Conference on Empirical Methods in Natural Language Processing
Conference Location : Kuala Lumpur, Malaysia
Conference Dates : February 12-13, 2017

Open Science Index, Cognitive and Language Sciences Vol:11, No:02, 2017 publications.waset.org/abstracts/66422.pdf