A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles

Abstrak

The Malay language is an Austronesian language spoken in most countries in the South East Asia region that includes Malaysia, Indonesia, Singapore, Brunei and Thailand. Traditional linguistics is well developed for Malay but there are very limited resources and tools that are available or made accessible for computer linguistic analysis of Malay language. Assigning part of speech (POS) to running words in a sentence for Malay language is one of the pipeline processes in Natural Language Processing (NLP) tasks and it is not well investigated. This paper outlines an approach to perform the Part of Speech (POS) tagging for Malay text articles. We apply a simple Rule-based Part of Speech (RPOS) tagger to perform the tagging operation on Malay text articles. POS tagging can be described as a task of performing automatic annotation of syntactic categories for each word in a text document. A rule-based POS tagger generally involves a POS tag dictionary and a set of rules in order to identify the words that are considered parts of speech. In this paper, we propose a framework that applies Malay affixing rules to identify the Malay POS tag and the relation between words in order to select the best POS tag for words that have two or more valid POS tags. The results show that the performance accuracy of the ruled-based POS tagger is higher compared to a statistical POS tagger. This indicates that the proposed RPOS tagger is able to predict any unknown word's POS at some promising accuracy.