

A Rule-Based Named-Entity Recognition for Malay Articles

ABSTRACT

A Named-Entity Recognition (NER) is part of the process in Text Mining used for information extraction. This NER tool can be used to assist user in identifying and detecting entities such as person, location or organization. Different languages may have different morphologies and thus require different NER processes. For instance, an English NER process cannot be applied in processing Malay articles due to the different morphology used in different languages. This paper proposes a Rule-Based Named-Entity Recognition algorithm for Malay articles. The proposed Malay NER is designed based on a Malay part-of-speech (POS) tagging features and contextual features that had been implemented to handle Malay articles. Based on the POS results, proper names will be identified or detected as the possible candidates for annotation. Besides that, there are some symbols and conjunctions that will also be considered in the process of identifying named-entity for Malay articles. Several manually constructed dictionaries will be used to handle three named-entities; Person, Location and Organizations. The experimental results show a reasonable output of 89.47% for the F-Measure value. The proposed Malay NER algorithm can be further improved by having more complete dictionaries and refined rules to be used in order to identify the correct Malay entities system.