A genetic-based HAC technique for parallel clustering of bilingual Malay-English corpora

ABSTRACT

Multi Multilingual corpora, containing the same documents in a variety of languages, are becoming an essential resource for natural language processing. Clustering multilingual corpora provides us with an insight into the differences between languages when term frequency-based Information Retrieval (IR) tools are used. It also allows one to use the Natural Language Processing (NLP) and IR tools in one language to implement IR for another language. For instance, in this way, the most relevant articles to be translated from language Malay to language English can be selected after studying the clusters of abstracts in language English. In this paper, we report on our work on applying Hierarchical Agglomerative Clustering (HAC) to a large corpus of documents where each appears both in Malay and English. We cluster these documents for each language and compare the results both with respect to the content of clusters produced. On the data available, the results of clustering one language resemble the other, provided the number of clusters required is relatively small. Further, we study the effects of changing the method used to compute the inter-clusters distance that includes single link, complete link and average link distance between clusters. Finally, we describe an experiment employing a genetic algorithm to fine-tune the individual term weights in order to reproduce more closely a predefined set of clusters. In this way, clustering becomes a supervised learning technique that is trained to better reproduce known clusters in language Malay when applied to the corresponding documents in language English. Other possible applications include training the algorithm on a hand clustered set of documents, and subsequently applying it to a superset, including unseen documents, incorporating in this way expert knowledge about the domain in the clustering algorithm.