Multiword phrases indexing for Malay-English cross-language information retrieval

ABSTRACT

Cross-Language Information Retrieval (CLIR) is the process of providing queries in one language and returning documents relevant to that query which is written in a different language. A popular approach to CLIR is to translate the query into the language of the documents being retrieved. One of the simplest and most effective methods for query translation is to perform dictionary look-up based on a bilingual dictionary. However, lack of dictionary coverage prune two problems: proper names and compound words handling. Relevance concept words consist of proper names and compound words, were applied in document and query indexing and query translation processes. We believed by using concept-based indexing and translations makes proper names and compound words translation possible. A series of experiments conducted to test the compound words and proper names translation methods in CLIR system. The best retrieval performance obtained from the combination of query translation approach-select all translations listed in the dictionary, alternative weighting scheme and proper names identification and translation. For both Malay and English document collection, these approaches outperformed query translation approach, select all translations listed in the dictionary, by 1.0 and 9%. The results show that proper names and compound words translations were important in query translation for Malay-English CLIR.

Keyword: Concept-based IR; Cross-language information retrieval; Query translation; Bilingual dictionary; Proper names identification and translation