Learning translation knowledge from created Persian-English Comparable Corpus for Cross-Language Information Retrieval

The fast growth of the World Wide Web and the availability of information in different languages have attracted much attention in research on cross-language information retrieval (CLIR). One of the main issues in CLIR is where to obtain the translation knowledge. Multilingual corpora, either parallel or comparable, are widely used for this purpose and are available in many language pairs. However, the Persian language, as a widely spoken language in the Middle East, does not have rich multilingual resources due to some of its special features and difficulties in constructing the corpora.

In this work, we create a Persian-English comparable corpus, which is, to the best of our knowledge, the first big comparable corpus for Persian and English. We construct the corpus from two independent news collections: Hamshahri news (www.hamshahrionline.ir) in Persian and BBC News (www.bbc.co.uk) in English and then align the documents. Intuitively, documents with similar content which are published on the same date are most probably talking about the same event. Based on these intuitions, we follow these steps to align the documents: Extract the keywords of the documents in the source language, translate the keywords to the target language and run the translated queries against the target collection. We then align the documents based on their similarity scores and publication dates. We tried several alternatives for constructing the comparable corpora, such as different translation methods and different retrieval models.

In the second step, we assessed the quality of the corpora using different criteria. As the first and most important criterion, we used a five-level relevance scale to manually evaluate the quality of alignments for one month. The evaluation results show that by properly translating the query words and using Okapi with pseudo relevance feedback as the retrieval model, we can come up with a high quality comparable corpus, for which 95% of the assessed matched articles are highly or fairly about related events.

We also used other criteria for further evaluation of the corpus, for example the ability to extract meaningful word associations from the documents. In order to extract word associations from the comparable corpus, we used a statistical method based on co-occurrence of words in the alignments. Our experiments show that, the confidence of matching decreases as we go down the list of associations but the word pairs are still related. This observation suggests that these results can be used in query expansion.

Furthermore, we did cross-language information retrieval using the cross-lingual word associations extracted from the comparable corpus. One challenge here is how to incorporate word associations into a CLIR mode. We study this issue in the language modeling framework. With the extracted translations, we construct the query language model in the target language corresponding to the given query. We then retrieve the documents in the target language using the generated query language models. Experiment results show promising results for extracting translation knowledge from the corpus, although it needs more exploration.