Automatic Identification of English and Indonesian Parallel Documents

Jelita Asian Falk Scholer S.M.M. Tahaghoghi Justin Zobel
Email: {jelita,fscholer,saied,jz}@cs.rmit.edu.au
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia.

Aim: Parallel corpora are useful for cross-lingual information retrieval and other natural language processing tasks. However, current techniques for finding parallel documents rely on file names and file structures, and semantic and statistical information in documents. We aim to automate the identification of English and Indonesian parallel documents.

Methods: We have developed a new global alignment method, based on the methods used for applications such as matching of protein sequences, to align windows of words between documents and queries. The documents can be translated to the query language before the alignment, or – as is the case for Indonesian and English – they can remain untranslated if they share the same character set. For each pair of documents, we group words into windows of a certain size with a 50% overlap with the next window. We then count the number of unique words in common between the windows. We use a global alignment method for aligning protein sequences, based on the Needleman and Wunsch algorithm [1], to align these windows of words. The basic principle of the method is to find as many matches as possible between the two documents, and to punish when there is any insertion or deletion.

Results: We compare our alignment schemes, using different window sizes and penalty values, with results obtained by documents indexed using a search engine called Zettair. The following table shows that our alignment method can differentiate between parallel and non-parallel documents when compared to a search engine baseline. This differentiation is measured using a separation value, which is the difference between a highest false match and a lowest true match. The negative separation value indicates that the Zettair baseline ranks non-parallel documents higher than parallel documents. A smaller window size works well for untranslated English documents, while a larger window size works well for translated English documents.

		Untranslated	Translated
Zettair Baseline		-24.7097	-0.2471
	Window Size		
Optimum scheme	12	27.7194	—
	28	—	19.1492

Conclusions: Our global alignment method is successful in separating parallel documents between Indonesian documents with either translated or untranslated English documents.

References

[1] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, Volume 48, Number 3, pages 443–453, 1970.

Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.