Abstract

The term parallel corpora are typically used in linguistic circles to refer to texts that are translations of each other. And the term comparable corpora refer to texts in two languages that are similar in content, but are not exact translations. In order to exploit a parallel text, some kind of text alignment, which identifies equivalent text segments (approximate sentences), is a prerequisite for analysis. Parallel corpora are very much essential in cross lingual or multilingual information retrieval. This paper presents an approach for automatic creation of English-Odia parallel corpus from comparable corpus. Generally Named entities, Proper nouns and common nouns play an important role in information retrieval. We tried to find the effectiveness of named entities, Proper nouns and common nouns in aligning English – Odia comparable document pair. We have taken the Odia parallel corpus (152 English-Odia documents) from TDIL, as well as we have crawled comparable Wikipedia pages for testing and the results are encouraging. We have used Stanford coreNLP tool and Google translator in our work.

References


Index Terms

Computer Science

Information Sciences

Keywords

Cross lingual information retrieval

Named entity

Comparable document

document similarity

key terms