Abstract

Plagiarism is entirely a paper copy of another person without including the source (original author) and later recognized as his own paper. Plagiarism will be very detrimental to the owner when it came to be published by people who are not eligible.

The automatic detection of plagiarism is a task that has acquired relevance in the Information Retrieval area and it becomes more complex when the plagiarism is made in a multilingual panorama, where the original and suspicious texts are written in different language. From a cross-lingual perspective, a text fragment in one language is considered a plagiarism of a text in another language if their contents are considered semantically similar no matter they are written in different languages and the corresponding citation or credit is not included.

In this project on cross-lingual plagiarism analysis are based on the exploitation of a statistical bilingual dictionary. This dictionary is created on the basis of a parallel corpus which contain original fragments written in one language and plagiarised versions of these fragment written in another language. Development of statistical bilingual dictionary by using IBM Model-1 alignment and Expectation Maximization Algorithm. Statistical bilingual dictionary is the core of this research.

Keyword: Plagiarism, Cross-Lingual Plagiarism, Corpus Parallel, Statistical Bilingual Dicionary, Fragment, IBM Model-1 alignment, Expectation Maximization Algorithm