Introduction

• Cross Language Information Retrieval (CLIR) is the retrieval of documents in a language different from the query language.

• Challenges for Arabic CLIR include: orthographic variations; morphology; broken plurals; ambiguity of the words; diacritization; and synonyms.

• The dictionary-based query translation is a traditional approach used by CLIR systems but translation ambiguity and words that are Out Of Vocabulary (OOV) are major problems.

• Web-mining has been demonstrated to be among the most efficient approaches to address these problems.

Research Questions

• How can translation candidates be extracted?

• How can correct translation candidates be selected?

• How can different resources be combined?

• Do these models achieve satisfactory effectiveness?

Introduction

• Cross Language Information Retrieval (CLIR) is the retrieval of documents in a language different from the query language.

• Challenges for Arabic CLIR include: orthographic variations; morphology; broken plurals; ambiguity of the words; diacritization; and synonyms.

• The dictionary-based query translation is a traditional approach used by CLIR systems but translation ambiguity and words that are Out Of Vocabulary (OOV) are major problems.

• Web-mining has been demonstrated to be among the most efficient approaches to address these problems.

The Search Result Based Method

The Method
Arabic search result pages from search engines contain rich snippets of summaries with a mixture of English query and Arabic aliases which might not be covered in Machine Readable Dictionaries.

The goal is to mine these summaries to determine potentially relevant translations.

Example

Bilingual search-result page. The example contains an English term (data mining) and the equivalent translation in Arabic (البيانات فى التحليل).