Combining Dictionary and Web-mined Resources for Arabic-English Cross Language Information Retrieval

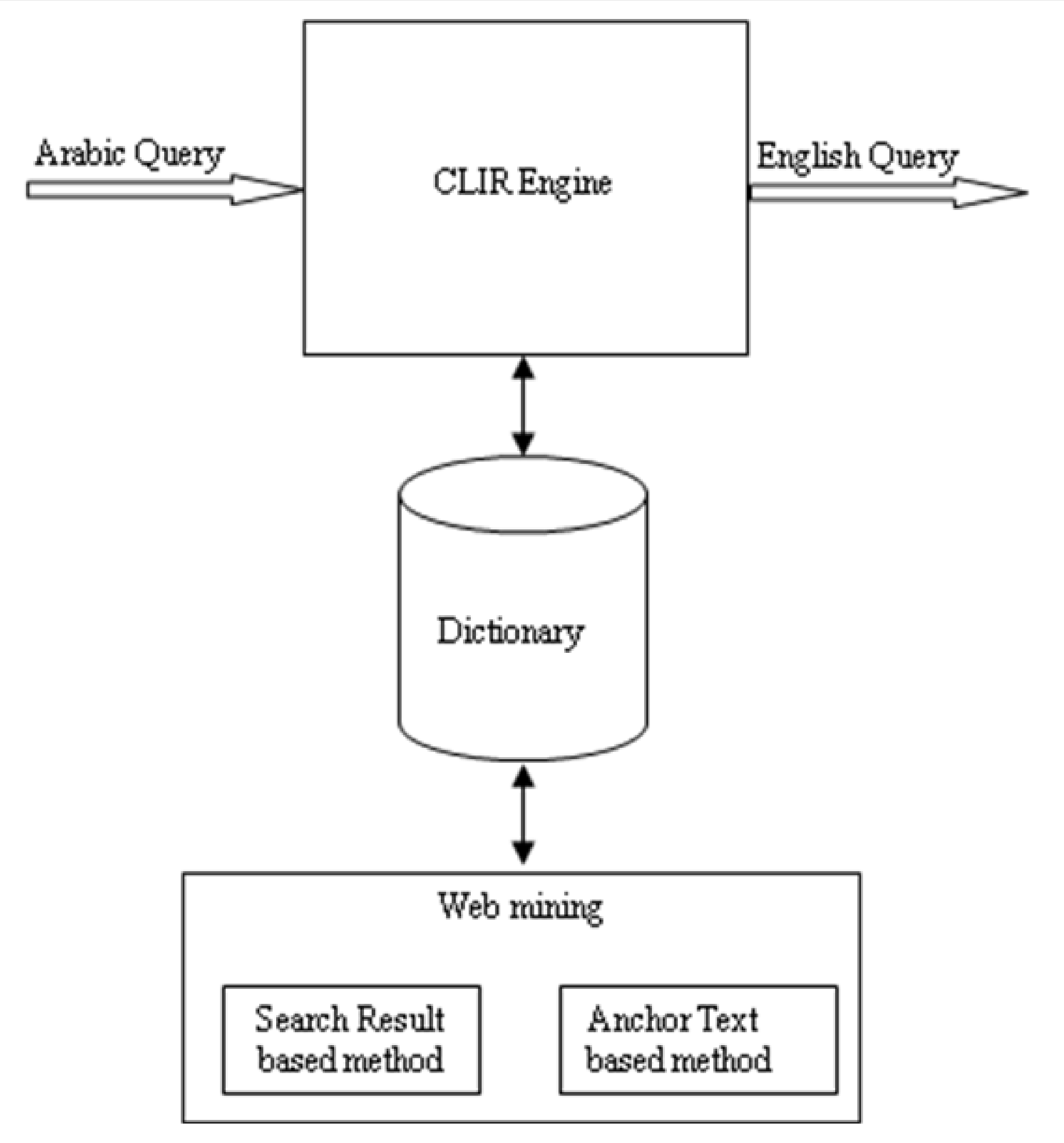
Introduction

•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?

The Search Result Based Method

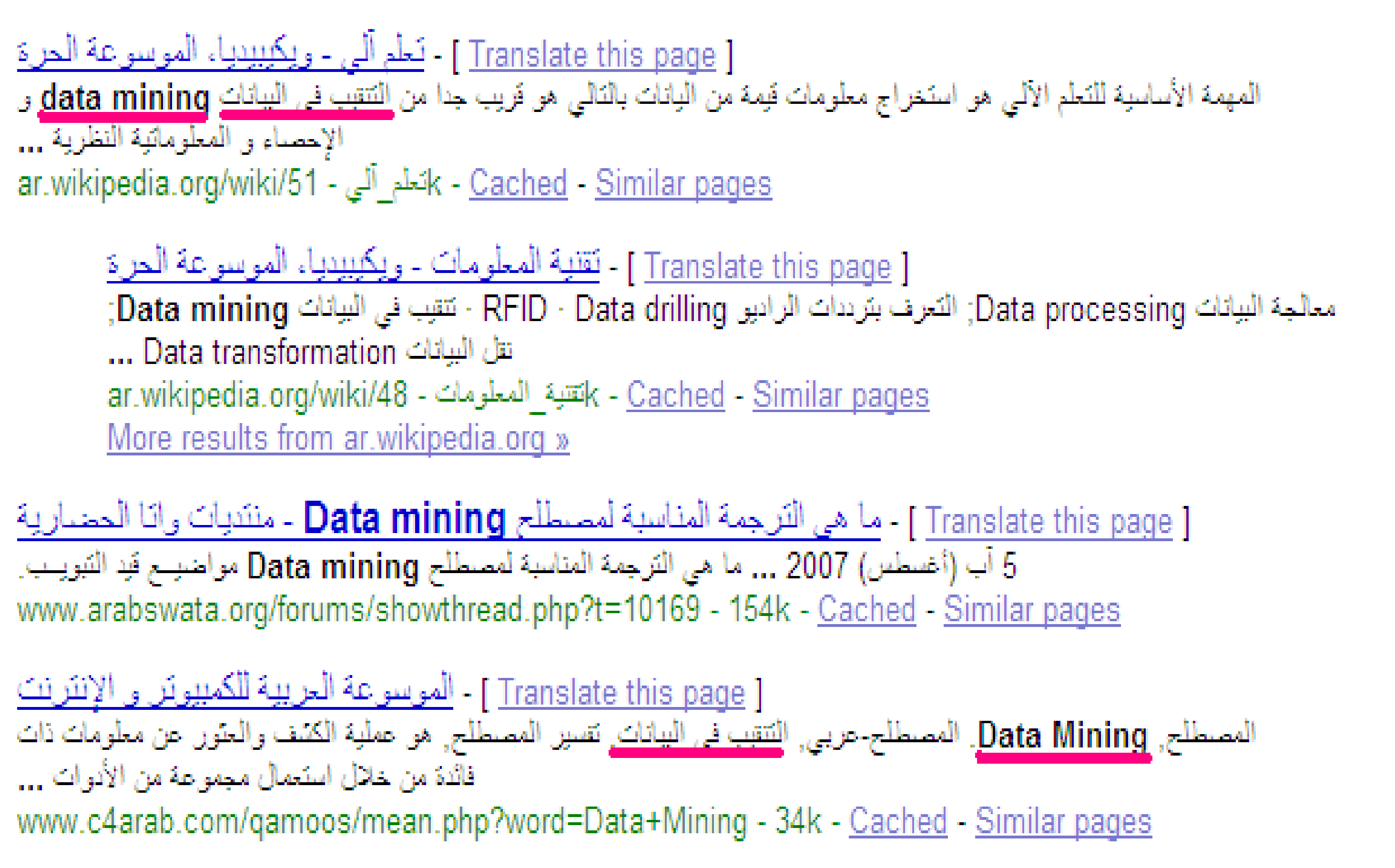
The Method

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 (التنقيب فى البيانات).



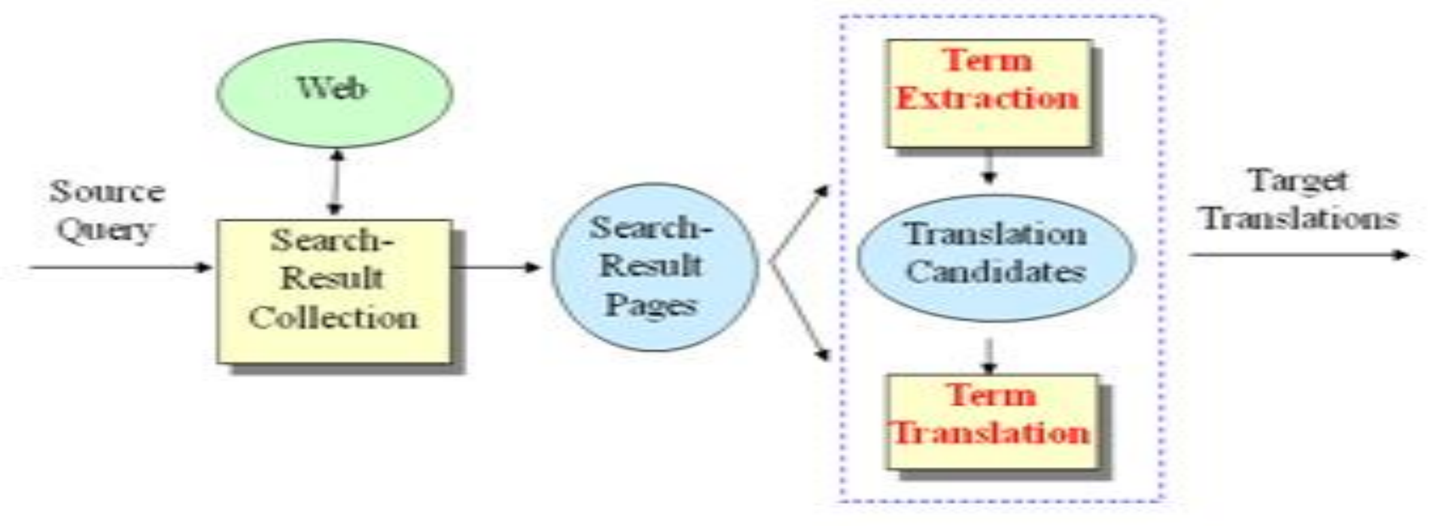
تعلم آلي - ويكيبيديا، الموسوعة الحرة - [Translate this page]
المهمة الأساسية للتعلم الآلي هو استخراج معلومات قيمة من البيانات بالتالي هو قريب جدا من التنقيب في البيانات data mining و
الإحصاء و المعلوماتية النظرية ...
ar.wikipedia.org/wiki/51 - تعلم_آليk - Cached - Similar pages

تقنية المعلومات - ويكيبيديا، الموسوعة الحرة - [Translate this page]
معالجة البيانات Data processing; التعرف بترددات الراديو Data drilling · RFID · تنقيب في البيانات Data mining;
نقل البيانات Data transformation ...
ar.wikipedia.org/wiki/48 - تقنية_المعلوماتk - Cached - Similar pages
More results from ar.wikipedia.org »

ما هي الترجمة المناسبة لمصطلح Data mining - منتديات واتا الحضارية - [Translate this page]
5 آب (أغسطس) 2007 ... ما هي الترجمة المناسبة لمصطلح Data mining مواضيع قيد التبويب.
www.arabswata.org/forums/showthread.php?t=10169 - 154k - Cached - Similar pages

الموسوعة العربية للكمبيوتر و الإنترنت - [Translate this page]
المصطلح, Data Mining. المصطلح-عربي, التنقيب في البيانات تفسير المصطلح, هو عملية الكشف والعثور عن معلومات ذات
فائدة من خلال استعمال مجموعة من الأدوات ...
www.c4arab.com/qamoos/mean.php?word=Data+Mining - 34k - Cached - Similar pages

System model



The Anchor Text Based Method

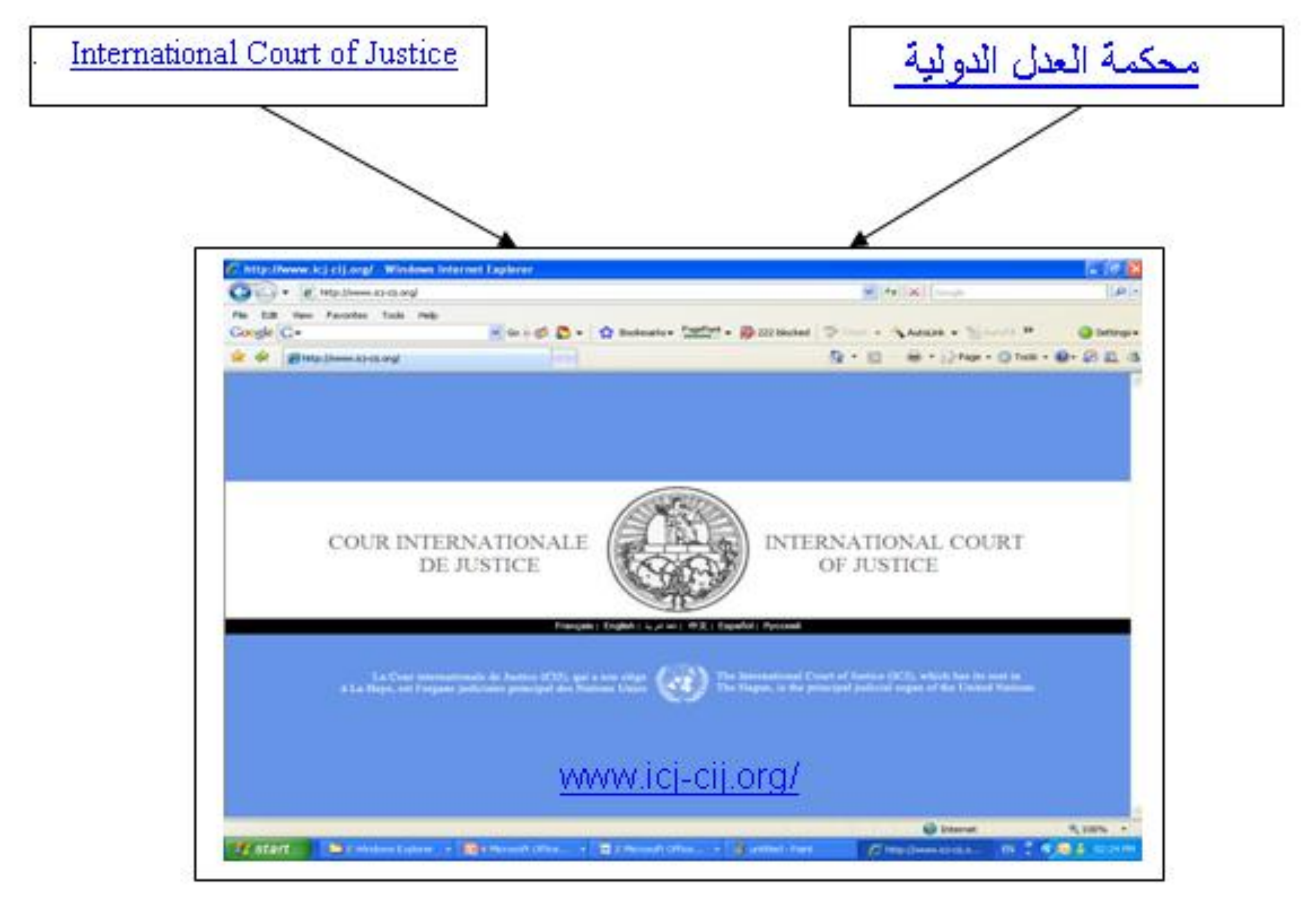
The Method

Anchor text is a brief description of an out-link in a Web page. A Web anchor-text set is a set of text anchors pointing to the same page (URL). This set may contain similar description texts in multiple languages.

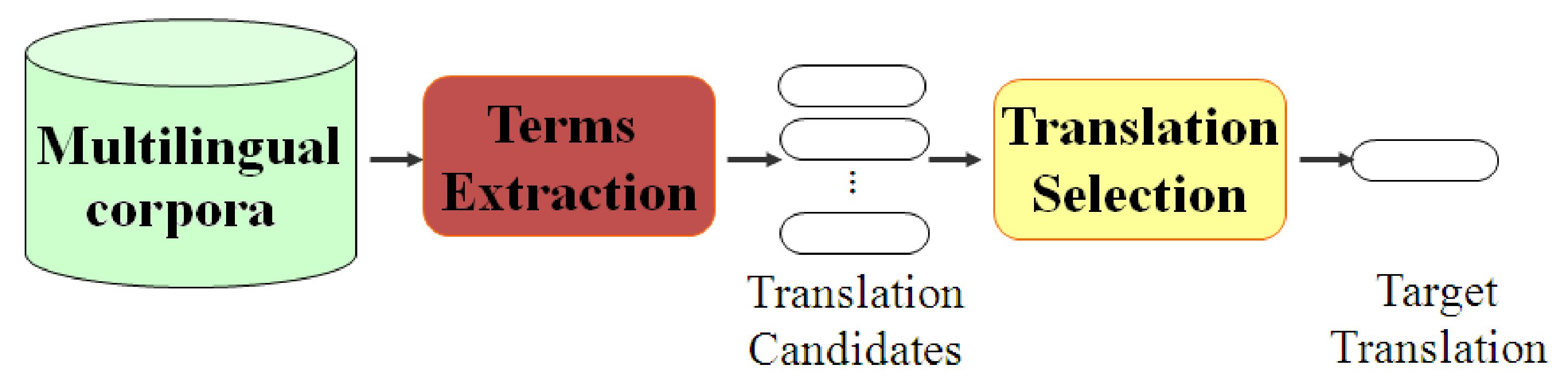
The goal is to build a small multilingual corpus then extract translation for OOV terms from it.

Example

Two text anchors in different languages point to official website of the judicial body of the United Nations: "International Court of Justice" and "محكمة العدل الدولية".


International Court of Justice
محكمة العدل الدولية
COUR INTERNATIONALE DE JUSTICE
INTERNATIONAL COURT OF JUSTICE
www.icj-cij.org/

System model



Expected Outcomes

•Addressing the Out Of Vocabulary (OOV) and ambiguity problems for dictionary-based CLIR approaches; and resolving ambiguity issues in Arabic IR.

•Contributing to the improvement of techniques for Arabic Machine Translation (MT).

Hisham Abdalla - hisham@cs.uct.ac.za
Hussein Suleman - hussein@cs.uct.ac.za

Department of Computer Science
Digital Libraries Laboratory
http://www.cs.uct.ac.za/research/dll