multi Searcher: Can We Support People to Get Information from Text They Can't Read or Understand?

Farag Ahmed
Otto-von-Guericke University Magdeburg
Data & Knowledge Engineering Group
39106 Magdeburg, Germany
farag.ahmed@ovgu.de

Andreas Nürnberger
Otto-von-Guericke University Magdeburg
Data & Knowledge Engineering Group
39106 Magdeburg, Germany
andreas.nuernberger@ovgu.de

ABSTRACT

The goal of the proposed tool *multi Searcher* is to answer this research question: can we expect people to be able to get information from text in languages they can not read or understand? The proposed tool *multi Searcher* provides users with interactive contextual information that describes the translation in the user's own language so that the user has a certain degree of confidence about the translation. Therefore, the user is considered as an integral part of the retrieval process. The tool provides possibilities to interactively select relevant terms from contextual information in order to improve the translation and thus improve the cross lingual information retrieval (CLIR) process.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Search process*

General Terms

Algorithms, Performance, Experimentation, Human Factors

Keywords

cross lingual information retrieval, word sense disambiguation, Arabic

1. CLIR INTERACTION TOOLS

The increasing diversity of internet web sites has created millions of multilingual resources in the World Wide Web. Therefore, there is an urgent need to bridge barriers between languages in order to access this flood of multilingual information. In the past little attention was paid to develop multilingual interaction tools where users are really considered as an integral part of the retrieval process. However, the involvement of the user in CLIR systems by reviewing and amending the query had been studied, e.g., using Keizai [4], Mulinex [1] and recently MIRACLE [3]. These interfaces provide query translation from the source language into the target languages using bilingual dictionaries. Furthermore, they use a sort of "query assistant", which enables interactive disambiguation of the query translation process: supporting the user in selecting the correct translations out of a list of possible translations. In Mulinex the "query assistant" shows how the translated query term translates back into the source language in order to support those users who do not understand the target language. In Keizai, users have to select appropriate translations after examining source language definitions of each possible translation before the search is conducted. An important property of MIRACLE is the immediate feedback in response to any action (selecting/deselecting proposed translations), which gives the user an opportunity to refine the search. The MultiLexExplorer [2] follows a different strategy: it allows users to explore combinations of query term translations by visualizing EuroWordNet relations together with search results and search statistics obtained from web search engines.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. THE PROPOSED TOOL

multi Searcher deals with several CLIR issues. Firstly, there is translation ambiguity, i.e. one word in one language can have several meanings in another language. Secondly, the user's lack of knowledge in the target language. Here, the tool supports the user by providing interactive contextual information that describes the translation in the user's language. Due to the availability of the language resources needed for Arabic (dictionary and parallel corpora aligned at sentence level[1]) English was selected as test languages.

2.1 Translation Disambiguation

The translation process starts by translating the query terms; a set of possible translations of each of the query terms are obtained from the dictionary. Based on the translation sets of each term, sets of all possible combinations between terms in the translation sets are generated. Using co-occurrence data extracted from monolingual corpora[1], the translations are then ranked based on a cohesion score computed using *Mutual Information* (MI): Given a query $q = \{q_1, q_1, ..., q_n\}$, and its translation set $S_{qk} = \{q_k, t_i\}$, where $1 \leq k \leq n, 1 \leq i \leq m_k$ and m_k is the number of translations for query term k. The MI score of each translation combination can be computed as follows:

$$MI(q_{t_1}, q_{t_2}, ..., q_{t_n}) = log_2 \frac{P(q_{t_1}, q_{t_2}, ..., q_{t_n})}{p(q_{t_1})p(q_{t_2})...p(q_{t_n})} \quad (1)$$

with $P(q_{t_1}, q_{t_2}, ..., q_{t_n})$ being the joint probability of all translated query terms to occur together, which is estimated by counting how many times $q_{t_1}, q_{t_2}, ..., q_{t_n}$ occur together in the corpora. The probabilities $p(q_{t_1})p(q_{t_2})...p(q_{t_n})$ are esti-

[1]see www.nongnu.org/aramorph and www.ldc.upenn.edu

mated by counting the number of individual occurrences of each possible translated query term in the corpora.

2.2 Interactive Contextual Information (ICI)

When the user query is translated, it is looked up in the target language documents index in order to obtain the relevant documents (contextual information) for the translation. In order to get the equivalent documents in the source language the parallel corpora is queried. Since it is possible that some retrieved documents will be very similar – which would result in duplicate contextual information – the documents retrieved from the source language are automatically grouped and contextual information is selected only once from each cluster. As shown in Fig. 1, the finally selected contextual information is not provided to the user as raw text, but instead a classified representation of each contextual information term will be presented: each term of the contextual information is colored according to its related type and can be selected as disambiguating term (the user's query terms green, suggested terms by the tool based on high frequent co-occurrences in the context of the query bold blue and underlined, all remaining terms blue except stop words that are not selectable and black). For example, consider the following case where the user submitted the Arabic query "دين الحكومة *dyen alḥkwmh* ". The query term "الحكومة *alḥkwmh* " has two translations (the government or the administration), while the other term "دين *dyen* " has several possible translations .e.g. (Religion) or (Debt). Based on the MI score translation alternatives are displayed in ranked order together with their contextual information. Thus the user has the possibility to select the suitable translation. Here, the translations provided by the system (the government religion) and (the government debt) are correct even though they are used in a different context. This is due to the fact that (government) appears frequently in the context of "religion" or "debt". As shown in Fig. 1, the user is interested in the second ranked translation (debt government). Using the contextual information, the user can select one or more terms to improve the translation. To simplify the user's task, the tool automatically proposed relevant terms (highlighted in bold blue and underlined), .e.g., ("payment", "financial", "lending", "loan"). Once the user selects, for example, the interactive term "اقراض *āqrāḍ* " (loan or lending), the tool re-translates the modified query and displays the new translations ("debt government loan", "debt government lending" and "debt administration loan"), to the user. Using search engine integrated web services, the user can, with a simple mouse click, confirm the translation which will then be sent to his favorite search engine, retrieving the results and displaying them.

2.3 Evaluation

We selected randomly 20 Arabic queries from the corpora that included at least one ambiguous word having multiple translations. The number of senses per test word ranged from 1 to 14, and the average was 4.3. The number of query translation combinations ranged from 4 to 200 with the average being 29.1. In order to evaluate the performance of the tool, we used two measurements: applicability and precision. The applicability is the proportion of the ambiguous words that the algorithm could disambiguate. The precision is the proportion of the corrected disambiguated senses of

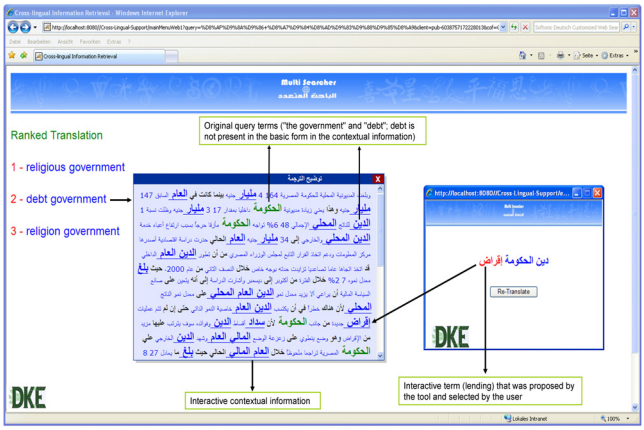

Figure 1: The translation alternatives with their contexual information in the source language.

the ambiguous word. The evaluation has been performed using monolingual corpora over the 20 test queries. The applicability and precision were 80% and 70%, respectively. The tool has been initially tested by 5 users who have (no knowledge or little knowledge) about the target language. The results were very encouraging, in that the tool could give the users a certain degree of confidence about the translation. Furthermore, the possibility to interactively select term/terms from the contextual information in order to improve the translation was praised.

3. CONCLUSIONS AND FUTURE WORK

We proposed a context-based CLIR tool, to support the user, in having a certain degree of confidence about the translation. It provides the user with interactive contextual information in order to involve her/him in the translation process. The translation ambiguity was taken into account by the use of a MI score based approach. Experiments about the accuracy of the tool proved that the tool has a certain degree of translation accuracy. In addition, a small pilot user study (5 participants) was conducted. A larger user study has already been designed and is underway.

4. REFERENCES

[1] J. Capstick, A. K. Diagne, G. Erbach, H. Uszkoreit, A. Leisenberg, and M. Leisenberg. A system for supporting cross-lingual information retrieval. *Inform. Proc. and Management*, 36(2):275–289, 2000.

[2] E. W. D. Luca, S. Hauke, A. Nürnberger, and S. Schlechtweg. MultiLexExplorer - combining multilingual web search with multilingual lexical resources. In *Combined Works. on Language-enhanced Educat. Techn. and Devel. and Eval. of Robust Spoken Dialogue Sys.*, pages 71–21, 2006.

[3] D. W. Oard, D. He, and J. Wang. User-assisted query translation for interactive cross-language information retrieval. *Inform. Proc. and Management*, 44(1):181–211, 2008.

[4] W. C. Ogden and M. W. Davis. Improving cross-language text retrieval with human interactions. In *33rd Hawaii Intl. Conf. on System Sciences*, volume 3, page 3044, 2000.