

THESAURUS-BASED QUERY DISAMBIGUATION METHOD FOR CROSS-LANGUAGE INFORMATION RETRIEVAL

Ahmad M. Hasnah Jihad M. Jaam

University of Qatar, Department of Computer Science
P.O. Box 2713, Doha, Qatar
{hasnah,jaam}@qu.edu.qa

Abstract: *Bilingual dictionaries are an important resource for query translation in cross language information retrieval. However, term translation is not an isomorphic process, dictionaries usually have several translations for a single word, therefore dictionary-based cross language retrieval systems must address the problem of ambiguity in language translation. In this paper we present and evaluate a conceptual disambiguation method using thesauruses. Two words are a translation of each other if they have the same conceptual representation (associated with the same set of words and synonyms). Our method is applied to the Arabic-English cross language retrieval problem. Experimental results show clearly that our method enhance significantly the retrieval performance.*

Keywords: *Disambiguation, Cross-language Information Retrieval, Bilingual dictionary.*

1. Introduction

With the ongoing increase of online multilingual documents and the growing number of the World Wide Web and Internet users from different countries (i.e. different native speaking languages), researchers are becoming increasingly interested in the problem of cross-language information retrieval (CLIR). Cross-language information retrieval is defined as taking a query in one language and retrieving relevant documents in other languages. The cross-language information retrieval problem is more complex than traditional information retrieval (where user queries and documents are in the same language) because some methods for query or document translation must be used to translate queries, documents or both before a traditional information retrieval system can be used. Several approaches for query translation have been proposed and tested such as bilingual dictionary, parallel corpora, and machine translation software etc (see [1, 2, 5, 8, 12, 14]). A brief description of the different translation method is given in the background section.

User queries are often ambiguous phrases. In cross-language information retrieval system the user query is translated into another language(s). Identifying the correct translation of the user query is a very complex task. In fact, one single term in the source language often have several translations in the target language, which could increase user query ambiguity. In this paper, we propose and evaluate a new method for query translation disambiguation for cross language information retrieval. The basis of our method is that two terms are a translation of each other if they have a highly similar conceptual representation, which means that the two words are mainly used in the same circumstances and are associated with the same set of words. The terms (word) conceptual representation is constructed using a language thesaurus to identify term synonyms and related words. Our method can be applied to any cross-language information retrieval systems that use bilingual dictionary as a method for query translation. However, in this paper we are interested in applying and evaluating our proposed method on the Arabic-English cross language retrieval.

The paper is organized as follows: In section 2, we give a brief overview of the different cross-language information retrieval and query translation methods. We also describe some bilingual dictionary disambiguation methods. In section 3, we discuss the characteristics of the Arabic language, while in Section 4, we present our disambiguation method for Arabic-English cross-language information retrieval. In Section 5, we discuss the different experiments that we carried out and finally, in Section 6 and 7, we give our results and conclusion respectively.

2. Background

In this section we present the main previous work in cross-language information retrieval (CLIR). Different methods of query translation are briefly explained. The previous work in the area of query translation disambiguation using bilingual dictionaries is also introduced.

2.1 Cross Language Information Retrieval (CLIR)

In cross language information retrieval (CLIR) system the user enters the query in one language and the system retrieves documents in another language or (languages). As a result, either documents or queries are translated. When storage space is limited and several languages must be accommodated, translating the query is more suitable than translating each document into every language [9,10].

There are three basic types of methods for query translation: bilingual dictionaries, multilingual corpora, and machine translation system (MT). Different cross-language information retrieval systems use one or combinations of these resources.

Corpus-based systems use parallel and or comparable corpora for query translation. In parallel corpus, the same documents exist in both the source and target language(s). In a comparable corpus, the relationship between documents is less clear, and several definitions have emerged in recent research works. Sheridan et al. [14, 5] uses a very strict definition: comparable documents are about the same event and written at the same time, each in different language. While Peters [11, 5] requires only that they are related by genre and style of diction. It is clear that as documents become less and less comparable, using them for CLIR becomes more and more difficult.

Dumais, Landauer, and Littman [8, 5] have used a matrix reduction technique called Latent Semantic Indexing (LSI) to extract language independent term and document representations from a parallel corpus. LSI applies singular value decomposition to the large, sparse term-document concurrence matrix (including terms from all parallel versions of the document) and extracts a subset of the singular vectors to form a new vector space.

Sheridan et al. [14,5] have created a reference corpus for cross-language information retrieval by aligning news stories in German and Italian by topic label and date and merging them to parallel documents. German queries are then translated into Italian by using these combined documents for query expansion in conjunction with a word similarity thesaurus. The similarity thesaurus technique is similar to LSI in that terms are related to one another by their distribution across documents, but it performs no dimensionality reduction of the term-document matrix.

Dictionary-based system perform query translation by looking up component terms and phrases in a bilingual dictionary and forming a target language query by concatenating some or all of the translations. Radwan and Fluhr [12] have constructed a cross-language text retrieval system known as SPIRIT which uses a ranked boolean retrieval system in conjunction with bilingual term, compound, and idiom dictionaries for query translation and document retrieval. They found that their dictionary-based cross-language system is more effective than machine translation based cross-language system.

Hull et al. [8] have constructed a dictionary-based cross-language system. They compared an automatic word-based translation model to one with manual correction of the dictionary and to one with a comprehensive multi-word terminology dictionary. The experiment result was promising.

The third method of query translation is by using machine translation software. Text translation is the process of mapping the query from the source language directly into one or more target language using a machine translation (MT) system. The quality of translation in current MT system is often low. Machine translation gives high quality translation when the applicable domain is limited. Usually, the users' queries are sequence of words without proper syntactic structure and they cover wide range of domains. MT needs more context than is in query for accurate translation, as a result, the performance of machine translation system is less than satisfactory [4,9,10].

2.2 Dictionary-Based Disambiguation Methods

Bilingual dictionaries often have several translations for a single query term. Query disambiguation is the process of retaining relevant translation and removing the translation noise. Several researches have been conducted on the problem of query disambiguation of the English query translation to other languages using a bilingual dictionary as a translation tool.

Yamabana et al. [8] have built an English-Japanese cross-language retrieval system which uses a bilingual dictionary, comparable corpora, and user interaction. For terms in the dictionary with more than one translation equivalent, the best equivalent for a particular domain is selected using a method based on statistical concurrence over comparable corpora. Each source language term is matched to the most similar term in the same language. The target language equivalents of this pair are then compared and the one with highest statistical concurrence frequency are selected as being appropriate in the selected domain. The system also has an interactive user interface, which allows the searcher to select the most appropriate translation equivalent using term definition in his/her own language.

Ballesteros and Croft [1,5] have built a cross-language information retrieval system that use query expansion to solve the query ambiguity problem. Their hypothesis is that additional terms, which are related to the primary concepts in the query are likely to be relevant and that by adding these terms to the query, the effect of incorrect equivalents generated during the translation process can be reduced. Query Expansion is performed using an automatic feedback technique to select new terms which occur frequently in conjunction with the query terms. This expansion can be conducted in the source or target language.

Davis [2,5] also starts with dictionary-based query translation and has developed two strategies that perform direct disambiguation to select the best translation equivalent. His system uses part of speech tagger to tag the query terms with part of speech information, which allows the system to select only those equivalents from the dictionary which have the same part of speech. It then measures the

similarity between each source language query term and the remaining equivalents to aligned sentences in a parallel corpus. Disambiguation is performed by selecting the equivalent whose sentence ranking is most similar to the source language term. Ranking can be compared across languages because the sentences are aligned.

3. Arabic Language Characteristics

The Arabic language has more complex morphology than the English language. Arabic language is highly rich in synonyms, a simple word such as *طعام* which means food has the following synonyms (*غذاء*, *مأكل*, *مأكل*, *مأكل*, *مأكل*, *مأكل*). This fact causes a problem of mismatch between the user query terms and the dictionary entries, as a result, part of the user query is not correctly translated, thus leading to unsuitable or incomplete target query.

Vowels are used in the Arabic language to identify and change the meaning of a word, part of speech, verb tense, etc. Machine-readable Arabic-English and English-Arabic dictionaries, online documents, and user queries are not vowelized. Thus, two words that have exactly the same letters in both, the dictionary and the user query may have extremely different meanings. For example, the word *علم* could mean (science, flag, teach, and knew) based on vowelization. As we see from the example there is a big potential for wrong or ambiguous translation. The system often unable to identify exactly which of the different meanings the user intended to use in his query. The system is unable to correctly identify the query words in the source language in order to correctly translate them to the target languages.

Users from different Arab countries use different words to refer to the same thing. Some of these words have no meaning or are being used to reference other thing in different countries. Thus, the problem of missing word (no translation) in the dictionary has a big potential to exist especially that the user query is usually not standardized while most dictionaries are based on the standard Arabic language. For example the Arabic word *مدرسة* means course in Jordan while it does not mean anything in Egypt or Lebanon and it is not part of any standard Arabic-English dictionary.

Finally, some Arabic characters could be written in several ways that are different from the standard Arabic writing (this is usually the case in the user query and online Arabic document). As a result, a mismatch between the user query vocabulary and dictionary entries could occur and lead to wrong or no translation of some of the source query. For example, in modern writing style, the word *بلا* is usually written as *بلا* without the *ا*.

As we have seen there are several problems in the Arabic language that contribute to the complexity of the Arabic-foreign cross-language retrieval. In addition, the different characteristics of the Arabic language compar to English require an adjustment and reinvestigation of approaches proposed for English cross-language retrieval before they can be used or adapted for the Arabic Language. In the next section we present our methodology to disambiguate query translation and we apply it to Arabic-English query translation as our study case.

4. Methodology

In dictionary-based cross language retrieval, it is common for a single word in the source language to have several translations, where some of them are with totally different meanings. Removing the noise terms will increase dramatically the retrieval performance. The basis of our method is that two terms are a translation to each other if they are used in the same circumstances or associated with similar set of words. In other words, have relatively similar sets of synonyms and related words. Our methodology is based on representing each term in the source (i.e., the Arabic Language) query, and every possible English translation (of each Arabic term) by list of synonyms and related words using a machine readable thesauruses. The identification of the set of most relevant translation in the target language (English) is based on the calculation of similarity between the Arabic term representation vector and the representation vectors of all possible English translation. Our method can be described as follows:

Let $Q_A = (a_1, a_2, \dots, a_n)$ be an Arabic query where a_1, \dots, a_n are the query terms. For every term $t = a_i$ in Q_A , repeat the following steps:

- 1- Represent the Arabic term t by a list of its synonyms and related words $A_t = (s_1, \dots, s_m)$ using a machine readable Arabic thesaurus.
- 2- Translate The term t into English by using a machine readable Arabic-English bilingual dictionary, using every match strategy, i.e. for every Arabic term t all matched English terms $E_t = (e_1, \dots, e_k)$ are considered as possible translation.
- 3- Replace every English translation e_j in E_t with a list of synonyms and related words in English $ES_t = (p_1, \dots, p_r)$ using a machine readable English thesaurus.
- 4- Translate back into Arabic every translation e_j and its synonym and related words list ES_t using a machine readable English-Arabic bilingual dictionary with first match strategy
- 5- Use Cosin formula [13] to calculate the similarity between the Arabic term t vector A_t (synonym and related words list) and retranslated vector (ES_t) of each e_j in E_t
- 6- The set of translated terms e_j that score a similarity greater than a threshold with t based on the two vectors (A_t & ES_t) similarity calculation is considered as the most appropriate translation of t .
- 7- If non of retranslated vectors ES_t yield back the Arabic term t or present a high similarity to the Arabic term t representation vector A_t , a first match translation approach is applied.

The following figure (Figure1) is a graphical representation of our methodology.

3- Each of the possible English translations of the Arabic word is represented by its synonyms and related words using an English thesaurus. Table 3 gives the vectors (set of synonyms and related words) representing each of the English translations

Table 3: English vectors of all possible translation

Word	Synonyms and related words
Dulcimer	Music instrument, Music, Soundbox, Musician, band, concert
Zither	Music instrument, Soundbox, Music, Musician
Law	Rule, regulation, principle, axiom, jury, constitution, court, judge, guideline, authority, police
Code	Rules, regulations, law, ethic, morality
Principle	Precept, doctrine, dogma, tenet, ideal, belief
Norm	Usual, standard, average, custom, mark

4- The vector of each possible English translation is translated back to Arabic using English-Arabic bilingual dictionary with first match strategy. Table 4 gives the retranslated English vectors into Arabic

Table 4: Retranslated English vectors

English word	Re-Translated vector
Dulcimer	, , , , , ,
Zither	, , ,
Law	, , , , , , , , , , , ,
Code	, , , , ,
Principle	, , , , ,
Norm	, , , ,

5- The similarity between the retranslated vector of the English words Table 4 and the vector of the Arabic word Table 2 is calculated using cosine formula [13]. Table 5 gives the similarity result.

Table 5: Similarity results.

Word	Similarity
Dulcimer	0.0
Zither	0.0
Law	0.55
Code	0.47
Principle	0.15
Norm	0.0

As we see from Table 5, the English words that may be used as a possible translations of the Arabic term are Law and Code. As a result, the translations related to music that is far away from the query

meaning are dropped which dramatically enhance the retrieval precision and reduce irrelevant retrieved documents.

5. Experiments

Several experiments are conducted to evaluate the efficiency of our approach in enhancing Arabic-English cross-language retrieval. Our query translation and disambiguation approach performance is compared to the performance of English monolingual information retrieval. It is also compared to performance of Arabic-English cross-language retrieval using first-match and every-match translation and disambiguation methods. Our English monolingual IR system and the Arabic-English CLIR system are based on the vector space model. In vector space model a vector in the n-dimensional space represents each document in the collection, where each dimension represents an index term. Relevant documents are identified by calculating the similarity between document vectors and query vector [13]. The following two different types of experiments are conducted:

- 1- English monolingual information retrieval experiment
- 2- Arabic-English cross-language retrieval using disambiguation mechanisms

5.1 Experimental Environment

Our experiments are conducted on a full year of the *Gulf Times* English newspaper articles. Our document collection consists of 10700 English articles in different subjects such as politics, economy, religion, etc. occupying a space of around 90.5 MB of disk storage. A set of queries in Arabic language is also required to conduct the different experiments. Some students (who are regular newspaper reader) in the senior year and from different majors at the University of Qatar supplied us with sixty Arabic queries. These queries vary in length, subject and writing style and vocabulary. Relevance judgment to identify the relevant articles associated with every query of the test set is carried out by a set of the University of Qatar Students. Each student is given the query set in addition to part of the document collection and was asked to identify the relevant set of documents to each query. For each query we identified the set of documents agreed upon by the majority of students.

Al-Qamous Arabic-English and English-Arabic machine readable dictionary form **SAKHR Software** is used in the translation process. **Al-Qamous** is a general purpose bilingual dictionary with the ability to translate a word from Arabic into English and visa versa. An online Arabic thesaurus form www.Ajeeb.com is used in the process of Arabic term synonyms and related word identification process. The online *Merriam-Webster* dictionary and thesaurus is used in the identification of the synonyms and related words for an English term.

5.2 English monolingual information retrieval experiment

In this experiment we evaluate the performance of English (query)-English (documents) retrieval. First the set of Arabic queries are translated to English manually by human expert in Arabic-English translation. The set of resulted English queries is supplied to our system and the monolingual retrieval process is conducted. Recall and Precision measures are used to evaluate the monolingual retrieval performance. Table 6 gives the average precision of the sixty queries run in the monolingual

mode. This experiment results is used as a base line to evaluate the three cross language retrieval experiments.

5.3 Arabic-English cross-language retrieval using disambiguation mechanisms

We have investigated the use of first-match, every-match and our proposed conceptual disambiguation mechanisms for dictionary-based Arabic-English cross-language retrieval. In every-match experiment each word in the source language query is replaced by all possible translations available in the bilingual dictionary. In the first-match experiment, one translation per query term is considered instead of all possible translation.

Every-match method introduce extraneous terms in the query, some of them are irrelevant to the user query, as a result, several irrelevant documents is retrieved and the performance of the cross-language retrieval system decreases. For example the word *qanun* which means law have the following translation using every match approach (dulcimer, zither, law, code, principle, norm). The first two translations are related to music, thus documents related to music are retrieved even though the query is about law. First-match could run into same problem as it is clear from the example. In addition, first-match approach could result in losing some of the relevant translation. Therefore we devised our method that retain the most relevant terms to the original query by trying to remove all incorrect and ambiguous terms from the translated query. The hypothesis of our method is that two terms are a translation to each other if they have similar synonyms and related words. In this experiment our proposed disambiguation and filtering approach described before is used to translated the Arabic queries into English before the retrieval process is carried out.

6. Result and Discussion

Precision and recall are the most commonly used measures to evaluate monolingual and cross-language information retrieval systems. Precision is the ratio of relevant retrieved document number to the total number of documents retrieved. Recall is the ratio of relevant retrieved document number to the total number of relevant documents in the collection. The first experiment we carried out is an English monolingual IR. English queries that have been manually translated from Arabic are supplied to the system and the retrieval process is carried out on the English collection of documents. This experiment is used as a base line of our disambiguation experiment. Table 6 gives the average precision of English monolingual and three different disambiguation experiments. As we can see from Table 6, all Arabic-English cross retrieval resulted in low retrieval accuracy as compared to the English monolingual retrieval, the Every-Match disambiguation method performed the poorest while our conceptual translation method gave the best performance compared to all disambiguation and translation methods.

In Table 7, We summarize the statistical significant test interpretation of our experiments. The evaluation is conducted using the paired *t-test* [15]. The obtained values demonstrate that the α performance differences of our disambiguation and First-Match (FM) methods over the Every-Match (EM) method are significant at a 99% confidence interval for our dataset. The performance differences of our disambiguation method over the First-Match (FM) is significant at 89% confidence interval for our data set.

Table 6: Average precision value of sixty queries in different experiments

	Average Precision	% Monolingual
English monolingual experiment	0.48	
Every-Match disambiguation experiment	0.26	54%
First-Match disambiguation experiment	0.33	69%
Thesaurus Disambiguation method	0.36	75%

Table 7: Statistical Significance Test.

Thesaurus-based vs EM	FM vs EM	Thesaurus-based vs FM
0.01	0.01	0.101

We have investigated the performance of first-match and every-match disambiguation approaches. First-match approach gave an average precision equal to 69% of the monolingual information retrieval precision. Every-match disambiguation approach gave an average precision equals to 54% of the monolingual information retrieval precision. As we can see both first-match and every-match translation methods lead to 31% and 46% drop in performance respectively compared to monolingual information retrieval. Our disambiguation method gives an average precision of 0.36, which is equal to 75% of the English monolingual retrieval performance. It is clear that our approach out perform first-match and every-match disambiguation approach. Our method solves the problem of first-match approach by not limiting the translation of the query terms to the first translation (which cause to lose some of the relevant translation) and filters other translations to reduce the effect of incorrect translation. The problem of several incorrect translations that lead to the large drop in the performance of every-match approach is solved by our method by retaining the most relevant translations. This is accomplished by using all possible translation in the first stage of translation then our method identify the most relevant translations by checking the translation that have similar synonyms and related word to the Arabic term.

7. Conclusions and Perspectives

Our results show that the performance of Arabic-English cross-language retrieval using bilingual dictionary can be enhanced and improved to a precision of 75% of the English monolingual retrieval. First-match translation approach suffers from the problem of ignoring some of the relevant translations that results in missing some of the relevant documents. Every-match approach uses all different translations (some of the are irrelevant to the user query) in the retrieval process, as a result, several irrelevant documents are retrieved causing the precision of the system to have a major drop. Our proposed disambiguation and filtering approach overcomes some of those translation problems. It starts with considering all different translation so there is no lose of relevant translation, it then filters those different translation through a similarity comparison between the source language synonyms and related

words and the different translation synonyms and related words. Translations that presents a high similarity will only be considered in the retrieval process.

However, if we consider the conceptual relations between the words contained in the query, we can improve translation and filtering quality. We are actually trying to define another version of query translation algorithm. In the near future, we will assess the quality of the translation, and how well it will improve document filtering.

8. References

- [1] Ballesteros, L. Croft W. B., 1996. Dictionary-based method for cross-lingual information retrieval. In proceeding of the 7th International DEXA conference on Database and Expert Systems Applications, pp. 791-801.
- [2] Davis M., 1997. New experiments in cross-language text retrieval at NMSU's computing research lab. In the 5th Text Retrieval Conference (TREC-5), pp. 447-454.
- [3] Dumais, S.T., Landauer, T. K. and Littman, M. L. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In Grefenstette, G., Smeaton, A. and Sheridan, P. (Editors). Working notes on the workshop on Cross-linguistic information retrieval. ACM SIGIR, Zurich, Switzerland, pp 16-23.
- [4] Gachot, D., Lange, E., and Yang, J. 1998. The SYSTRAN Browser: An application of machine translation technology in multilingual information retrieval. In Cross-Language Information Retrieval. G. Grefenstette, editor, pp. 105-118.
- [5] Grefenstette G., 1996 editor. Workshop on Cross-Linguistic Information Retrieval - SIGIR'96.
- [6] Hasnah, A., and Jaam, J. 2002. Query disambiguation for Arabic-English cross language information retrieval. In proceeding of the first International Conference on Intelligent Computing and Information Systems ICICIS. Egypt. pp. 297-301.
- [8] Hull D., 1997. Using structure queries for disambiguation in cross-language information retrieval. In Working notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, pp. 73-81.
- [9] Hull, D., and Grefenstette, G. 1996. Querying across languages. A dictionary-based approach to multilingual information retrieval. In Proceeding of the 19th Annual international ACM SIGIR Conference on Research and Development in Information retrieval, pp. 49-57.
- [10] Hull, D., and Grefenstette, G. 1996. Experiments in multilingual information retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- [11] Peters, C. and Picchi, E. 1995. Capturing the comparable: a system for querying comparable text corpora. In proceeding of Analisi Statistica dei Dati Testuali (JADT), pp. 247-254.
- [12] Radwan Khaled. 1994. Vers l'Acces Multilingue en Langage Naturel aux Bases de Donnees Textuelles. Phd Thesis, Unversite de Paris-Sud, Center D'Orsay.
- [13] Salton, G. and McGill, M. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.
- [14] Sheridan, P. and Ballerini, J. P. 1996. Experiment in multilingual information retrieval using the SPIDER system. In proceeding of the 19th ACM/SIGIR Conference, pp. 58-65.
- [15] Wonnacott, R. Wonnacott, T. 1990. Introductory Statistics, John Wiley & Sons, Fourth Edition.