Enhanced Query Expansion in English-Arabic CLIR

Abdelghani Bellaachia and Ghita Amor-Tijani
Department of Computer Science
Washington DC 20052
{bell,gamor}@gwu.edu

Abstract

Arabic is a language with a particularly large vocabulary rich in words with synonymous shades of meaning. Modern Standard Arabic, which is used in formal writings, is the ancient Arabic language incorporated with loanwords derived from foreign languages. Different synonyms and loanwords tend to be used in different writings. Indeed, the Arabic composition style tends to vary throughout the Arab countries (Abdelali, 2004). Relevant documents could be overlooked when the query terms are synonyms or related to the ones used in the document collection. This could deteriorate the performance of a Cross Lingual Information Retrieval (CLIR) system. Query Expansion (QE) using the document collection is the usual approach taken to enrich translated queries with context related terms. In this study, QE is explored for an English-Arabic CLIR system in which English queries are used to search Arabic documents. A thesaurus-based disambiguation approach is applied to further optimize the effectiveness of that technique. Indeed, experimental results show that QE enhanced by disambiguation gives an improved effectiveness.

1 Introduction

Query Expansion (QE) proved to be effective in different Cross Lingual Information Retrieval (CLIR) systems. It enhances the concept of a query by adding to it context-related terms, using the top retrieved documents. Those documents are retrieved using the original query. Related terms are then extracted from the presumed relevant documents using co-occurrence analysis, which is based on the assumption that if two terms tend to co-occur then they tend to be related. Finally, those terms are added to the translated query to form the final expanded query.

Query Expansion was indeed used in an English-Arabic CLIR system (Darwish and Oard, 2003) and proved to be effective. However, some of the words retrieved using this technique might not be relevant to the query. Disambiguation could be applied to reach the optimum effectiveness that could be reached. Different disambiguation techniques exist and have been explored; some of them have been used together with query expansion. However, the effect of disambiguation on QE has not been analyzed in a CLIR system involving the Arabic language.

2 Related Work

Disambiguation has been applied in different steps of query processing to improve the performance of CLIR systems. It has been applied at the stage of translation to identify the most relevant translated terms. Assuming that terms in a query are related, their translations are analyzed so that those that relate to the same context are considered to be the correct translations. One of the approaches used for this analysis procedure is to check the top retrieved documents for the translations that co-occur within a certain window size. The analysis is based on the assumption that correct translations of query terms should co-occur in the target language documents. This approach, usually referred to as co-occurrence analysis, was shown to be effective in reducing the effect of ambiguity (Ballesteros and Croft, 1998; Hull, 1997). Kishida et al. (2004) demonstrated its effect on retrieval effectiveness in a German-English-French CLIR system. The most reasonable combination of the most frequently appearing terms in the top documents were selected to be the correct translations. This disambiguation procedure was followed by query expansion. It resulted in an increase in average precision from 26% (QE with no disambiguation) to 68% of the monolingual run.

Lesk (1986) analyzed lexical disambiguation using word overlap. Definitions of each sense of the ambiguous word were compared to the meaning of the words that co-occurred with it. The sense with the definition overlapping most with the definitions of the co-occurring words was considered to be the correct translation. Experiments on a limited set of text demonstrated 40-70% accuracy.

User assisted disambiguation was also used. In AIR project, which is an English-Arabic CLIR system, disambiguation was applied both on translated terms and after QE was completed (Liddy and Diekema, 2005). Results showed an improvement in retrieval.

Hasnah and Jaam (2002) explored a disambiguation technique that extends the concept of the two-phase match approach which was analyzed in a research conducted at the University of Maryland (Aljlayl and Frieder, 2001). They tested their approach on an Arabic-English CLIR system. It is based on the hypothesis that two terms are translations of each other if they have similar synonyms and related words. The set of synonyms of an Arabic query word were translated to English using a bilingual dictionary. For each set of translations, the set of related words to each single translation were retrieved using an English thesaurus. Those related words were translated to Arabic using the first match strategy. The best translations were selected based on a similarity measure between the set of related terms to the original Arabic query and to those of the retranslated query.

In our study, we analyze QE in an English-Arabic CLIR system and explore, for the first time, the benefit of applying disambiguation on Arabic expanded terms. In Section 3, our disambiguation technique which we refer to as "Disambiguated Query Expansion" (DQE), will be defined and illustrated. In Section 4, experimental results will be presented. Finally, a summary of the findings will be given in Section 5.

3 DQE Technique

The process of translating a query from one language to another often introduces ambiguity. Query terms usually have more than one translation, and those translations could have synonyms. Identifying the correct translations to use in the target query can be very complex. One of the techniques used to enhance query translation is query expansion. It is used to enhance the translated query with related terms extracted from the document collection. The basic approach of query expansion follows two steps: the identification of a presumed set of relevant documents; then, the selection of related terms used for query enrichment. This process of adding related terms to the translated query helps improve precision as more relevant documents could be identified and retrieved. However, not all expanded terms are necessarily directly related to the query. In this study, English WordNet is used to further disambiguate the expanded queries. It is used to analyze expanded query terms by comparing them to the definitions provided for the original query terms provided. We refer to this disambiguation process by the DQE technique.

3.1 WordNet

WordNet (Miller, 1990; Miller et al., 1990) is a large lexical database of English. It provides semantic knowledge of terms. It can be used in overcoming many problems related to the richness of natural languages. It was constructed by George Miller and his colleagues at the Cognitive Science Laboratory at Princeton University. The translations of a word are grouped into different "synsets". Each one of them includes synonyms with a distinct meaning and semantic pointers that define the relationship between the different synsets of a word (Richardson et al., 1994).

WordNet is used in information retrieval for two different purposes, mostly for indexing and word sense disambiguation. Researchers showed that using WordNet in indexing can improve retrieval (Gonzalo et al., 1998; Mihalcea and Moldovan, 2000). In their indexing technique, words in documents were disambiguated by first defining their Part Of Speech (POS) and the corresponding stems using the WordNet stemming algorithm. The documents were then processed and each word was replaced by its position in the text, its stem, its POS, and the offset which corresponded to the synset where the word occurs. The documents were then indexed using the stem and, separately, the offset and POS. This technique allowed the retrieval of a given sense of the word and its synonyms. Documents could be retrieved based on whether the indexed word is a keyword, a sense of a keyword, or a synonym of a keyword. This combined word-based and synset indexing improved recall by 16% and precision by 4% over basic word indexing (Mihalcea and Moldovan, 2000).

When used in word sense disambiguation, the definitions of the translations of the query terms are compared to each other. The translations with a common definition are considered to be the correct ones. This technique demonstrated that WordNet can indeed enhance retrieval effectiveness (Liu et al., 2004). An improvement ranging from 4% to 34% was achieved. The same research group tested using WordNet with pseudo relevance feedback. Ten words were extracted from the top retrieved documents. They were considered in the expanded query only if their definition from WordNet, including the synsets and hyponyms, contained one of the query terms. The ones that correlated formed the final expanded query. An additional improvement of 5% to 9% over word sense disambiguation was obtained.

3.2 DQE Algorithm

Using the DQE technique, our approach of QE enhanced by a thesaurus-based disambiguation, we try to benefit from the advantage of using query expansion taking into consideration that not all expanded terms are necessarily related to the query. Using QE, words co-occurring with the query terms in the set of highly ranked documents are added to the translated query. Using DQE, those terms are analyzed so that only those directly related to the original query are considered in the expanded query.

Few steps are followed in the DQE technique. The English query is first translated using an English-to-Arabic dictionary. QE is then applied to extract the expanded terms. Those terms are translated back to English, generating one translation. Also, related words to the original English query terms are obtained using the English WordNet. Finally, Arabic expanded terms are analyzed. Only those with a translation that occurs in the WordNet entry of one of the original query terms are considered in the expanded query. A more detailed explanation of this procedure is given later in Section 3.3. A formal description of the algorithm is given below:

```
For each query, qi, do
  Get set of expanded terms ETi;
  For each expanded term, eti, in ETi do
    get corresponding English translation teti;
    For each teti do
    if(teti ∩ {Wordnet definitions of query terms})=Φ)
      Eti = Eti -{eti}
    end for
  end for
end for
```

3.3 DQE Framework

Figure 1 represents the flow chart of the steps taken in the DQE technique. Each of the queries present in the query file is processed separately. The arrows indicate the flow of data, the oval blocks represent the main modules used in the processing of the queries, and finally the rectangular blocks represent the processed data.

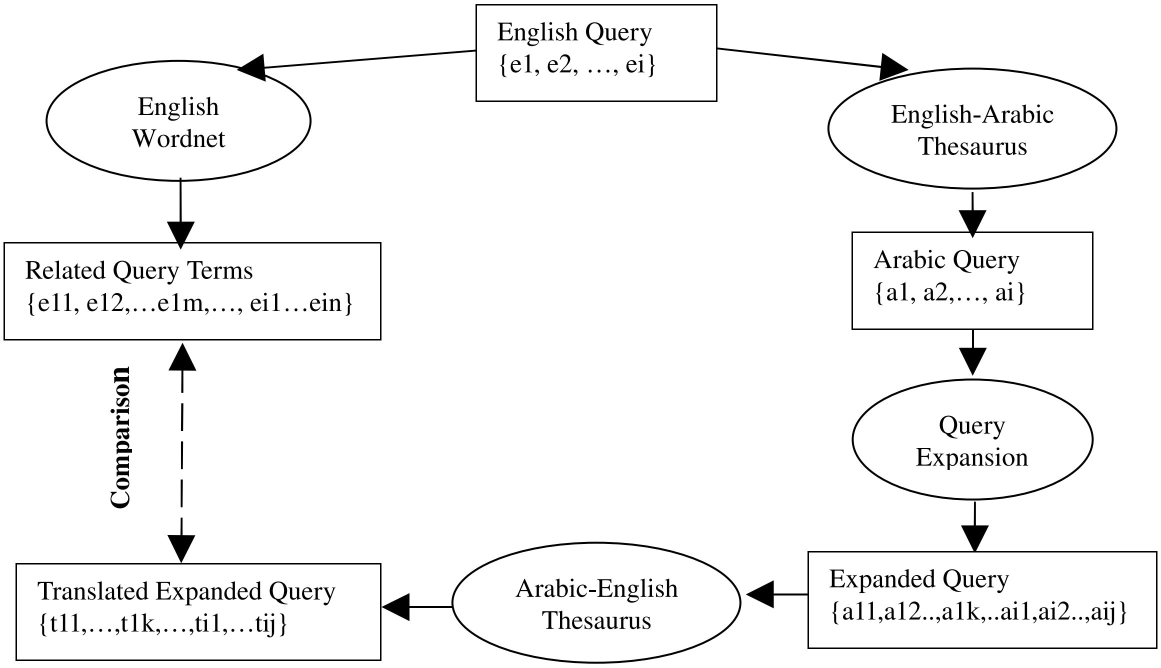

Figure 1: Query Processing Using the DQE technique

After the English query is fed into our system, it is translated using an English-to-Arabic dictionary. In case a translation is not found for one of the query words, transliteration is applied. Once the Arabic query is ready, query expansion is applied using Indri[1] tools. Expanded terms are translated back to English using an Arabic-to-English dictionary. One translation was considered for each term. The one with the highest probability was chosen; as it is clarified later in Section 4, the dictionary we are using provides probabilities for each set of translations. These translations are then compared with the definitions of the original query words given in WordNet. The synset with the correct meaning was selected manually to ensure a correct analysis. The set of synonyms together with their

[1] http://www.lemurproject.org/indri/

definitions were considered in the comparison. Only expanded terms found in the definitions of one of the query terms were kept in the final query. This procedure causes unrelated terms to be removed from the expanded query. For example, in the query “Bin Laden targets US”, general terms like: “infrastructure, movement, newspaper, unity” are removed from the expanded query as they can limit the retrieval effectiveness that could be gained using query expansion.

Figure 2 below illustrates the different steps of query processing in our system. After the query is tokenized and stemmed, query terms are translated. Untranslatable words are transliterated using either one of the following two approaches. In the first approach, the Transliteration N-Gram (TNG) technique is applied to extract spelling variants of the transliterated word (Bellaachia and Amor, 2008). Using this approach, a set of possible transliterations is obtained by using the n-gram approximate string matching technique on a transliteration automatically generated and the stemmed terms in the index of the document collection. In the other approach, referred to as the enhanced TNG (eTNG) (Bellaachia and Amor, 2008), POS disambiguation is applied on the set of transliterations extracted using TNG so they only include words with the same POS as the original query term. Once the translated Arabic query is formed, relevant documents are retrieved using the Indri search engine. Query expansion is then performed using context-related terms from the top n documents. To further disambiguate those expanded terms, WordNet disambiguation is applied as explained earlier in this section. A final retrieval is carried out using the disambiguated expanded Arabic query.

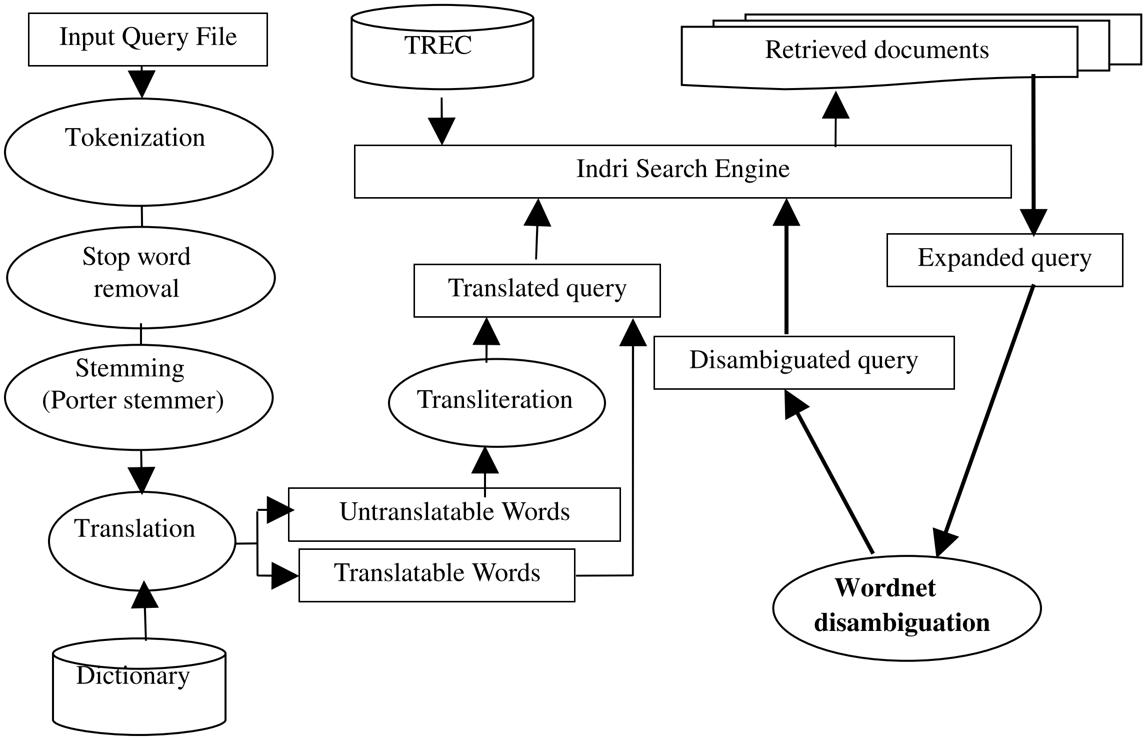

Figure 2: DQE Framework

4 Experimental Results

The document set used is known as the "Arabic Newswire Part 1". It contains approximately 383,872 news articles taken from the “Agence France Presse” Arabic newswire dated from May 13, 1994 through December 20, 2000. This collection consists of 76 Million tokens among which are 666,094 unique words. The 2002 Text Retrieval Conference (TREC 2002) topics in English were used to search the Arabic document set. The English topics were processed before they were fed into the search engine. After the queries were tokenized and stemmed, the query terms were translated using a dictionary[2] generated from the United Nations parallel English-Arabic corpus. This bilingual dictionary provides a set of translation probabilities for plausible translations. The English terms are processed with the Porter stemmer[3] and the Arabic translations with Al-Stem stemmer provided by Kareem Darwish (From the University of Maryland).

The following results are for 50 topic queries. Query expansion was used to enrich the query with 30 terms selected from the top 50 documents. This procedure is what is referred to as pseudo relevance feedback. The numbers were chosen after comparing the effectiveness of runs using QE with different numbers of terms and documents. Indri tools were used for that purpose. To further disambiguate the expanded query, we used WordNet to filter out unrelated expanded terms. The effect of our disambiguation approach is reflected in the results that follow. Performance improvement is analyzed for the TREC 2002 (English-Arabic) cross-language track. Mean Average Precision (MAP) is the metric used to compare the performance of different runs. Average precision is calculated as the mean of precision scores after each relevant document is retrieved. MAP is the mean of average precision values for a set of queries.

Different runs were carried and compared:

- *Mono*: The monolingual run was used as our baseline to evaluate the performance of the DQE technique. Al-Stem was used to stem both the Arabic queries and the Arabic document set.
- *QE*: Query expansion was applied on translated queries processed with the TNG technique also described in Section 3.3.
- *eQE*: Query expansion was applied on translated queries processed with the eTNG technique described in Section 3.3.
- *DQE*: Disambiguation using WordNet was applied on expanded queries processed with QE.
- *eDQE*: Disambiguation using WordNet was applied on expanded queries processed with eQE.

4.1 Performance Evaluation

The results below illustrate an 11-point non-interpolated average precision, averaged over the 50 TREC 2002 queries used in our experiments.

Table 1: Average Precision of the Mono, QE, DQE, eQE, and eDQE Runs

Recall	*Mono*	*QE*	*DQE*	*eQE*	*eDQE*
0	0.5968	0.4466	0.4643	0.4685	0.5053
0.1	0.3845	0.3289	0.3489	0.3439	0.3725
0.2	0.3192	0.2857	0.3094	0.309	0.3305
0.3	0.2831	0.2545	0.2759	0.2688	0.3014
0.4	0.2494	0.2338	0.2501	0.2471	0.2792
0.5	0.2306	0.1965	0.2185	0.2093	0.2416
0.6	0.1839	0.1742	0.2031	0.1894	0.2277
0.7	0.1567	0.1552	0.169	0.1636	0.2011
0.8	0.1298	0.1229	0.1322	0.1234	0.1648
0.9	0.0651	0.0798	0.0803	0.088	0.1145
1	0.0134	0.0133	0.0136	0.0152	0.0181

The table above presents precision values of different runs: Mono, QE, DQE, eQE, and eDQE. The performance of each one of them is illustrated in the graph that follows. The graph shows how close the CLIR system's effectiveness is to the monolingual retrieval run, in which the effect of translation ambiguity is not present.

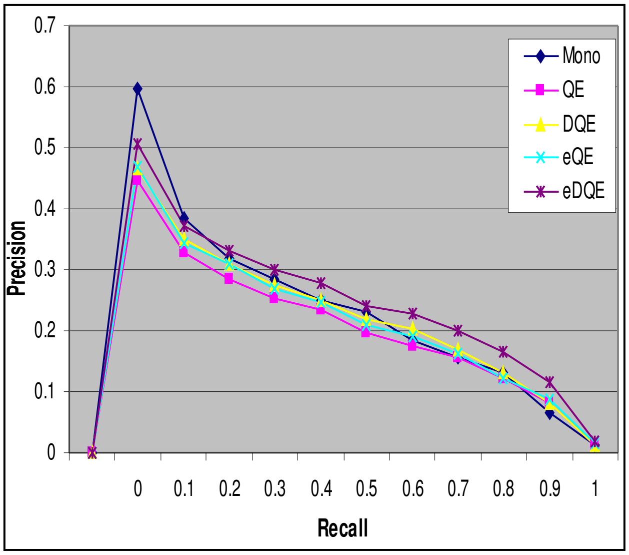

Figure 3: Precision-Recall Graph of the Mono, QE, DQE, eQE, and eDQE Runs

As we observe from the graph, some improvement is gained when disambiguation is applied on expanded queries. This improvement reflects the effect of filtering the expanded terms to include only those significant to the query. Query expansion enhanced with WordNet disambiguation surely enhances the effectiveness of the system.

[2] http://www.glue.umd.edu/~dlrg/clir/trec2002/bilingual_dictionary2.tar.gz

[3] http://tartarus.org/~martin/PorterStemmer

4.2 Summary of Results

Table 2 shows the MAP values of different runs. It reflects the percent improvement of both the DQE and eDQE approaches over QE (%QE).

Table 2: Mean Average Precision: percent improvement of DQE over QE

	MAP	*%QE*
QE	0.1982	---
DQE	0.2147	**8%**
eQE	0.209	5.4%
eDQE	0.2415	**22%**

Table 3 depicts the efficiency of the same techniques in improving our CLIR system as a percent effectiveness of the monolingual run (%Mono). This represents how much each technique adds to the effectiveness as the monolingual run reflects a CLIR system without the effect of translation ambiguity.

Table 3: Mean Average Precision: effectiveness of DQE compared to monolingual

	MAP	*%Mono*
Mono	0.2186	---
TNG	0.1553	71%
QE	0.1982	88%
DQE	0.2147	**98%**
eQE	0.209	95.6%
eDQE	0.2415	**110.5%**

The tables above show the percent improvement of using QE with and without WordNet disambiguation. When queries are processed using QE, an 88% effectiveness of the monolingual run is achieved using query expansion alone; whereas, an additional 8% improvement is gained when using WordNet to further disambiguate the expanded terms. As a result, effectiveness reaches 98% of the monolingual run. Also, query expansion combined with WordNet disambiguation gave an improvement of 38% over the TNG technique described in more detail in (Bellaachia and Amor, 2008). On the other hand, when queries are processed using eTNG, effectiveness reaches 110.5% of the monolingual run. Our DQE approach does indeed improve the MAP and performance overall.

5 Conclusion

Different disambiguation techniques have been used together with Query expansion to further disambiguate translated queries. Different researchers have shown that this combined disambiguation approach improves retrieval efficiency in different CLIR systems. In this study, we have shown that applying disambiguation on expanded terms in an English-Arabic CLIR indeed enhances performance. We chose to use a corpus-based disambiguation to keep the most relevant expanded terms in the expanded query. Words returned using QE were analyzed and compared to the definitions of the original query terms given in WordNet. Only those used in the definitions of one of the query terms were considered in the final expanded query. This enhancement in query expansion improved performance as an additional 22% in MAP was obtained after applying disambiguation.

References

Abdelghani Bellaachia and Ghita Amor-Tijani. 2008. Proper Nouns in English-Arabic Cross Language Information Retrieval. To appear in IEEE Symposium on Computers and Communications (ISCC'08).

Abdelghani Bellaachia and Ghita Amor-Tijani. 2008. Enhanced Transliteration in an English-Arabic CLIR system. Submitted to the International ACM SIGIR Conference on Research and Development in Information Retrieval.

Ahmed Abdelali. 2004. Localization in Modern Standard Arabic. Journal of the American Society for Information Science and Technology (JASIST), Vol. 55, N. 1, 2004. pp. 23-28.

Ahmed Abdelali, Jim Cowie, Hamdi S. Soliman. 2004. Arabic Information Retrieval Perspectives. Proceedings of JEP-TALN 2004 Arabic Language Processing, Fez 19-22 April 2004.

Ahmad M. Hasnah and Jihad M. Jaam. 2002. Thesaurus-based Query Disambiguation Method for Cross-Language Information Retrieval. International Journal of Intelligent Computing & Information Sciences, Vol. 2, N.2, July 2002.

David A. Hull. 1997. Using structured queries for disambiguation in cross-language Information Retrieval. In: American Association for Artificial Intelligence (AAAI) Spring Symposium on Cross-Language Text and Speech Retrieval, Palo Alto, CA. 84-98.

Elizabeth D. Liddy and Anne R. Diekema. 2005. Cross Language Information Exploitation of Arabic. Power point presentation, April 2005.

Julio Gonzalo, Felisa Verdejo, Irina Chugur and Juan Cigarran. 1998. Indexing with Wordnet synsets can improve text retrieval. Proceedings of the COLING/ACL'98 Workshop on Usage of Wordnet for NLP, Montreal, Canada.

Kareem Darwish, Douglas W. Oard. 2003. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. Proceedings of the Text Retrieval and Evaluation Conference (TREC 2003).

Kazuaki Kishida, Noriko Kando and Kuang-Hua Chen. 2004. Two-stage Refinement of Transitive Query translation with English Disambiguation for Cross-Language Information Retrieval: A Trial at CLEF 2004. Proceeding of the 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Vol. 3491-2005, pp.135-142.

Lisa Ballesteros and W. Bruce Croft. 1998. Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 64-71.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. Proceedings of SIGDOC-86: 5th International Conference on Systems Documentation, Toronto, Canada, 24-26.

Mohammed Aljlayl, Ophir Frieder. 2001. Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation. Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, LA, 295-302.

Rada Mihalcea and Dan Moldovan. 2000. Semantic Indexing using Wordnet senses. Proceedings of ACL Workshop on IR & NLP, Hong Kong.

Ray Richardson, Alan F. Smeaton, and J. Murphy. 1994. Using Wordnet as a Knowledge Base for Measuring Semantic Similarity between Words. Proceedings of Artificial Intelligence and Cognitive Science Conference, Trinity College, Dublin.

Shuang Liu, Fang Liu, Clement Yu, Weiyi Meng. 2004. An Effective Approach to Document Retrieval via Utilizing Wordnet and Recognizing Phrases. Proceedings of the 27th annual international ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR ’04), Sheffield, Yorkshire, UK, 266-272.

Miller G. A. (1990). Nouns in WordNet: A Lexical Inheritance System. International Journal of Lexicography, Vol.3, N.4, pp. 245 - 264.

Miller G. A., Beckwith R., Felbaum C., Gross D., and Miller K., (1990). Introduction to WordNet : An Online Lexical Database. International Journal of Lexicography, Vol.3, N.4, pp. 235 - 244.