Arabic Query Expansion Using Interactive Word Sense Disambiguation

Riyad Al-Shalabi
Arab Academy for Banking and Financial Sciences Amman Jordan
rshalabi@aabfs.org

Ghassan Kanaan
Arab Academy for Banking and Financial Sciences Amman Jordan
ghkanaan@aabfs.org

Mustafa Yaseen
Al- Ahlyya Amman University Amman Jordan
myaseen@ammanu.edu.jo

Bashar Al-Sarayreh
Arab Academy for Banking and Financial Sciences Amman Jordan
bsarayreh @aabfs.org

Nada A. Naji
Arab Academy for Banking and Financial Sciences Amman Jordan
nadanaji@yahoo.com

Abstract

Word sense ambiguity is widely spread in all natural languages; a word may carry several distinct meanings. Human can figure out the suitable meaning according to the context in which the word occurs. The Arabic language is highly polysemous; in many situations we find it extremely necessary to disambiguate the word senses. This paper studies and compares the performance of a search engine before and after expanding the query through Interactive Word Sense Disambiguation (WSD). We found that expanding polysemous query terms by adding more specific synonyms will narrow the search into the specific targeted request and thus causes both precision and recall to increase; on the other hand, expanding the query with a more general (polysemous) synonym will broaden the search which would cause the precision to decrease.

Introduction

Information retrieval (IR) is a research area dealing with organization, storage and retrieval of information. The following components are present in an IR situation: user, information need, request (a spoken or written formulation of the information need), query (a search formulation in an IR system), information stored in computer(s), appropriate computer programs (S´andor et al. 2001). For the scope of this paper we will be concerned only with automatic information retrieval systems; automatic as opposed to manual and information as opposed to data or fact. An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request. This specifically excludes Question-Answering systems as typified by (Winograd 1972) and those described by (Minsky 1968). It also excludes data retrieval systems such as used by, say, the stock exchange for on-line quotations (Rijsbergen et al. 1979).

The fact that Arabic is a polysemous language may negatively affect the retrieval process; using broad and polysemous query terms may distract the retrieval process by having it dispersed into several categories and some of which may be irrelevant to the user actual request; as a result, the system performance will deteriorate and the user will be unsatisfied with the search results. Retrieving bad results is not always attributed to the search engine or IR mechanism, users might not always use the proper word or the right conjugation for a query term, some users may not even notice that the term they picked for the search is in itself is harmful (due to its being polysemous), moreover, ignoring the other synonyms of the query terms may cause many relevant documents not to be retrieved. Documents related to an IR query sometimes contain only the synonyms of the query words instead of the query words themselves. A simple IR system with no knowledge of synonyms fails to recognize the relevance of these documents to the query. So, we can improve recall of IR systems by considering the synonyms of the query words as a part of the IR query (Katz et al.). Collections may consist of different types of media and be of one or several languages; but this paper is based on an Arabic text-only collection with an IR system to which queries must be entered in Arabic. These may be single-term or multi-term queries.

Briefly, expanding the query by adding the term synonyms suitable for the user request, thus drawing the user's attention that a certain term has other senses will not only help retrieve more relevant documents holding the other synonyms but will also eliminate the drawback of broadening the query with too much irrelevant synonyms by allowing the user to add/remove the ones he/she finds suitable for his/her request. Studying how effective and fruitful is directing the search via an IR system using an Arabic document collection through word sense disambiguation is the objective of this paper.

Without detailed knowledge of the collection make-up and of the retrieval environment, most users find it difficult to formulate queries which are well designed for retrieval purposes. This difficulty suggests that the first query formulation should be treated as an initial attempt to retrieve relevant information (Alberer 2006). Then the documents initially retrieved could be examined for relevance and new improved query formulations could be constructed hoping that additional useful document

could be retrieved. Such query reformulation involves two basic steps: expanding the original query with new terms and reweighing the terms in the expanded query (Chen 2005). Query Expansion can be defined as the process of reformulating the query's bag-of-words to overcome the problem of mismatching potential documents and improving the performance of a search engine by including in the results the documents that are more relevant (of better quality), or at least equally relevant (Qiu et al. 1993) (Vectomova et al. 2006). Word sense disambiguation (WSD) involves the association of a given word in text meanings potentially attributable to that word (Ide et al. 1998). The synonyms of the query terms must be extracted from a dictionary or thesaurus that is integrated to the system.

Previous work

So far, there has been conflicting information on the effect of WSD on IR since there are many variations in implementing WSD-supported systems (Katz et al.); the language on which the system is based, its being a single-language-based or a cross-language IR (CLIR) system, automatic WSD systems employ some kind of a software called a disambiguater; this is based on one out of many disambiguation algorithms. Cruse (Cruse 1986), Pustejovsky and Levin, among others, investigated word meaning within the same language — monolingually — with the goal of quantifying meaning dimensions. An alternative approach is to use cross-linguistic correspondences for characterizing word meanings in natural language. This idea is explored by several researchers, Dyvik, Ide, Resnik & Yarowsky, and Chugur, Gonzalo & Verdejo but to date, it has not been given any practical demonstration (Diab 2003). Efthimiadis (Efthimiadis 2000) concluded that one-third of the terms presented to users in a list of candidate terms for query expansion were identified by the users as potentially useful for query expansion. Parapar, et al (Parapar et. al) reached to results showing that lexical expansion is not able to improve retrieval performance using the logical model they proposed. Sanderson (Sanderson 1994) concluded that word sense ambiguity is only problematic to an IR system when it is retrieving from very short queries. Al-Nobani (Al-Nobani 2008) studied the problems in IR systems when dealing with Arabic language and introduced new algorithms and solutions in several critical areas related to Arabic languages.

User tasks

The objective of this paper is to focus mostly on studying how effective and efficient expanding queries through disambiguating the word senses via user feedback on an Arabic IR system. The model will be based more on the user feedback for building queries, disambiguating them and most importantly, providing feedback of which documents are relevant to a certain query; this is a core task since the calculation of both precision and recall is totally based on this task. User tasks can be identified as follows:

1. Submit a query in Arabic.
2. Review the retrieved documents after having the query submitted (traditional search).
3. Deciding which documents are really relevant to the query
4. Resubmitting the search, at this step, the precision and recall can be actually calculated depending on the relevancy feedback performed in step (3)
5. Adding the expansion terms depending on the synonyms provided by the dictionary, the user needs to select the expansions synonyms carefully to match his/her request demands (enhanced search: interactive WSD). At this point the precision and recall will be recalculated depending on the new query (the original query + the hidden query: the expansion query terms).

IR system evaluation

In order to evaluate any system, we must apply the standard measurements to find the efficiency of its performance, and the relevancy of the search outcome, which is basically, measured using the Precision-Recall metric. In Information Retrieval, relevancy is not a binary evaluation (unlike Data Retrieval) but a continuous fraction, although relevancy judgment is "somewhat" binary, but it is still sometimes difficult to decide whether a document is relevant or not.

We need to compute the precision and recall for each of the selected queries before the expansion (traditional IR system) and after the best expansion (using the proposed interactive WSD technique); "the best expansion" is achieved by the user and how he picks the best candidate synonym for the expansion, it is supposed to be the best since the disambiguation is human-based and not machine-based, and the fact that the user actually knows what he's looking for (requesting) can be a guarantee of picking the best synonym. The evaluation process encompasses two distinct phases, these are:

- The traditional search phase, where the system will retrieve only the documents with the same stem as the query term.
- The interactive WSD phase (we will call it "the enhanced search phase"), where the system will retrieve the documents with the same stems as the original query as well as the ones with the same stems as the hidden query (the expanded query).

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

Results and conclusions

Query Expansion through interactive WSD is a very useful technique, but we should follow a "protocol" when employed to get the most and the best out of it and avoid its "side effects":

1. QE through Interactive WSD should be used when trying to narrow the search scope, i.e. when the query terms hold many distinct and different meanings; we tend to disambiguate the senses by expanding the query with the very specific ones.

2. QE through Interactive WSD should not be used when the original query terms are already clear, distinct and hold one specific meaning (i.e. not polysemous)
3. QE through Interactive WSD increases (or at least doesn't decrease) the recall when applied.
4. Precision decreases only when rule 2 above is not obeyed.

Results will be provided supporting our approach, we will give the values and plot them to produce a graph of precision and recall for a number of queries using the traditional version of the IR system, and compare the results with the same queries after being expanded, i.e. using the proposed technique.

Combining more than one QE technique by employing the best of each to create a hybrid method for expanding queries is one of the future "next-steps". Cross- language IR and QE are other difficult yet fruitful fields in the area of IR systems. Automated Intelligent WSD for Arabic IR would be a very crucial task, due to the complexity of the Arabic lexical structure, but hopefully, some time would come when we would say; we're just one step away from achieving it.

References

Dominich, S´andor Mathematical foundations of information retrieval. Dordrecht: Kluwer Academic (2001).

Winograd, T., Understanding Natural Language, Edinburgh University Press, Edinburgh (1972).

Minsky, M., Semantic Information Processing, MIT Press, Cambridge, Massachusetts (1968).

Van Rijsbergen, Keith, Information Retrieval. 2 ed. http://www.dcs.gla.ac.uk/Keith/Preface.html, (1979).

Kari Alberer, "Information Retrieval and Data Mining". Laboratories de systèmes d'informations répartis, (2006).

Berlin Chen, "Query Operations". Department of Computer Science & Information Engineering, National Taiwan Normal University, (2005).

Qiu Y., and Frei H., "Concept Based Query Expansion". In Procedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval, Pittsburg, 1993.

Vectomova O. and Wang Y., "A Study of the Efect of Term Proximity on Quey Expnsion", Journal of Information Science 23(4): 324-333, 2006.

Nancy Ide Vassar College & Jean Véronis Université de Provence, Word Sense Disambiguation: The State of the Art, Computational Linguistics, 1998.

Boris Katz, Ozlem Uzuner & Deniz Yuret, Word Sense Disambiguation for Information Retrieval, Massachusetts Institue Of Technology Cambridge, Massachusetts 02139.

D. Cruse. Lexical Semantics. Cambridge University Press, 1986.

Mona Talat Diab, Doctor of Philosophy, Word Sense Disambiguation Within A Multilingual Framework, University of Maryland, 2003.

Efthimiadis E. N., Interactive query expansion: A user-based evaluation in a relevance feedback environment, School of Library and Information Science, University of Washington, Seattle, ETATS-UNIS, 2000.

David Parapar, Álvaro Barreiro, David E. Losada, Query Expansion Using Wordnet With A Logical Model Of Information Retrieval, AILab, Department of Computer Science, University of A Coruña, Spain - Intelligent Systems Group, Department of Electronics and Computer Science, University of Santiago de Compostela, Spain.

M. Sanderson. Word-sense Disambiguation and Information Retrieval. In proceedings of ACM-SIGIR, 1994.

Alaa A. Al-Nobani, Improving Search Engines performance in query processing and indexing for Arabic Language, Faculty of Graduate Studies Al Balqa' Applied University, 2008.