Greek-English Cross Language Retrieval of Medical Information

E. Kotsonis, T.Z. Kalamboukis, A. Gkanogiannis, and S. Eliakis

Department of Informatics
Athens University of Economics and Business
Athens, Greece
tzk@aueb.gr

Abstract. Health information systems on the web basically support the English language. To access high-quality online health information it is frequently a barrier for non-English speakers or speakers of English as a foreign language. In this work we present a cross-language retrieval system to support Greek users in the medical domain, overcome the language barrier. We have performed a case study on the impact of stemming in the cross lingual retrieval in association with dictionary based query translation techniques. Finally, we conclude with results from a preliminary evaluation of the Greek-English CLIR prototype.

1 Introduction

Cross-language information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the users query. Today search engines retrieve documents written in the same language as the query. Cross-language retrieval supports users of multilingual document collections by allowing them to submit queries in their own language, and retrieve documents in any of the languages covered by the retrieval system. CLIR systems can be used by people with good reading skills in a second language but poor skills in writing and therefore these users cannot compose a query that will fulfill their information need as they could do in their mother language.

Cross-language and monolingual retrieval functionality can certainly be provided by a single system. An effective monolingual retrieval is actually the core of a cross-lingual retrieval system [1]. Indeed when we search for documents written in a foreign language, we must choose between two primal approaches: either to translate the documents of the target language, or to translate the queries. In both cases the problem is reduced to the monolingual retrieval. However, both directions of translation have their weaknesses: translating large document collections could be, computationally, an impractical task and short queries on the other side introduce uncertainty in their translations. Furthermore query translation imposes a kind of cost, which must be paid at the most challenging time - when a search engine is trying to optimize response time for a large number of

X. Gao et al. (Eds.): MIMI 2007, LNCS 4987, pp. 109–117, 2008.

nearly simultaneous queries. Thus we are seeking for a simple and fast algorithm to translate the queries.

It is a well-known fact that information retrieval is not equally difficult for each language [2]. For example, the morphological analysis of the documents may be considered as minor for languages like English compared to languages like Greek with a rich inflectional and derivational system. The plural inflection of the English noun which, apart few exceptions, is very simple (add -s) while in Modern Greek there are 41 different inflectional suffixes. Also there are different forms of the written Greek language: such forms include classical Greek and Modern Greek. In this work we examine Modern Greek texts in the domain of medicine. This is actually a mixed language of modern with puristic Greek. It must be mentioned here that in any CLIR or monolingual retrieval system dealing in a language with a rich inflectional and derivational system stemming plays an important role on the performance.

The major approaches for CLIR include the use of bilingual dictionaries [3,4], parallel collections [5] and comparable collections [6] or some kind of combination of these. In this work we address the problem of disambiguation when dictionary-based techniques are used for the translation. In particular we present a case study on the impact of stemming in the complexity of a word-by-word translation algorithm with look-ups to bilingual dictionaries.

In the rest of this work we present the architecture of the CLIR system, the translation module and results from the OHSUMED database, a subset of the MEDLINE database. Finally we conclude on the performance of the algorithms used and extensions are proposed for future implementation.

2 System Architecture

The proposed system contains two subsystems: a multilingual subsystem, for retrieving bilingual documents (a collection of scientific articles in medicine available in the Greek web) and a cross language subsystem, which provides only the interface to the MEDLINE database using the PubMed search engine. The PubMed search engine[1] is maintained by the US National Center for Biotechnology Information and provides public access to the MEDLINE database over the web. The interface performs all the analysis of the query before its submission to the database, that is stop-words removal, stemming, automatic translation and disambiguation as well as procedures for query expansion. The system's architecture is presented in figure 1.

As far as the Greek database is concerned the Lucene search engine is used, an information retrieval system developed by Apache [2]. Lucene, supports many types of preprocessing, scoring, indexing, and retrieval models and supports several retrieval models, including the standard vector space model. To enhance the retrieval we have incorporated a Greek stemmer as well as a list of the most frequently used words (stopwords).

[1] `http://www.ncbi.nlm.gov/entrez/query.fcgi?db=pubmed`
[2] `http://lucene.apache.org`

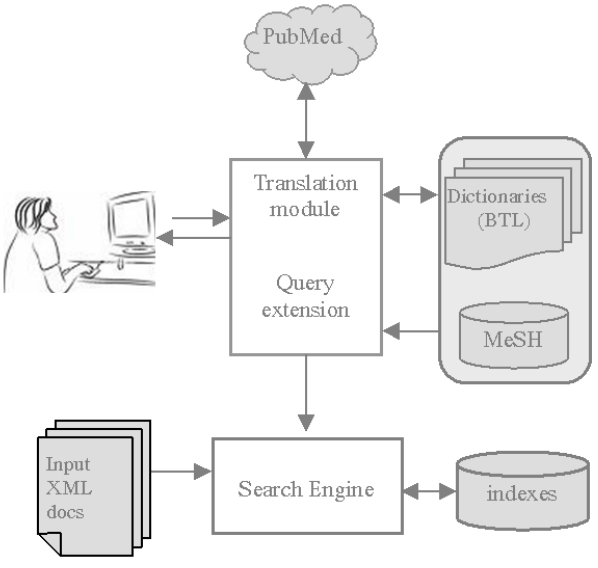

Fig. 1. CLIR, System's architecture

The Greek document collection contains bilingual articles in the medical domain, which are either entirely in English or in Greek, or they can use both languages with abstracts and references written in English. To ensure proper indexing of these documents using our standard architecture of Lucene, all documents are in UTF-8 format. The documents were indexed automatically using the TF*IDF weighting scheme [7]. Indexing includes stop-word removal, consulting a stopword list, stemming for the remaining tokens and weight estimation of terms, defined by $w(t,d)$

$$w(t,d) = TF(t,d) * log_2\left(\frac{N}{DF(t)}\right) \quad (1)$$

where $TF(t,d)$ is the number of occurrences of term t in the document d, $DF(t)$ is the number of documents in the collection that contain the term t and N is the total number of documents in the collection.

All the documents are in XML format. The structured data are used to filter the retrieved documents by the year of publication, and the thematic topic.

The user submits queries in natural language and has the choice to see the results from the Greek or the English database. Each document title in the ranked list of the results is followed by the best passage of the document containing the query words.

3 Dictionary Based Query Translation

In a CLIR system, users may be supported either to reformulate the query in order to choose the appropriate translations of the query terms, or to provide

a fully automated query translation unit. This last task introduces uncertainty due to the small size of the query. Although machine translation techniques are the state of the art in translation theyt are far from perfect and certainly not fast enough to be used in an online retrieval system. Dictionary-based approaches have been used in the literature for several languages in the past [4,8].

For the translation a Greek-English bilingual dictionary was used, consisting of about 40,000 fully inflected words or phrases. The dictionary was constructed by merging several Greek-English dictionaries (Bilingual Term Lists) and glossaries freely available in the web. Although we are not in a position to guarantee for the validity of the resulting dictionary we used these resources as the base for translation in our CLIR system.

Due to the morphological complexity of the Greek language, we expect the dictionary to have limited coverage. In order to improve on the coverage, a stem-based dictionary was derived from the original. However, although stemming improves the coverage of the dictionary introduces an additional level of uncertainty since more words with different meanings are conflated into the same stem. Thus in our case we face two levels of uncertainty: one introduced by the stemming process; and one due to the translation of words with more than one possible translation.

In what follows we have experimented with three algorithms for automatic query translation based on dictionary look-ups:

1. **Word by word translation:** In the word-based scenario, all the possible translations of a word remain in the translated query. Words of the target language that are present in the original query remain unchanged. For the experiments the stemmer described in [2] was used. We shall refer to this as stemmer-1. At this stage an investigation of the impact of the morphological normalization (stemming) on retrieval effectiveness was carried out, by testing a more conservative, based strictly on grammar rules, stemmer [9]. The experimental results presented in the next section show a significant improvement of the new stemmer (stemmer-2) in the case of the simple word-by-word translation algorithm.
2. **Word by word translation and disambiguation:** To reduce ambiguity due to stemming a filtering step is applied that selects the most appropriate translation. A given Greek word, g, first is stemmed to g' and then translated using the dictionary, see figure 2. Suppose

$$T(g') = \{(e_1, g_1), (e_2, g_2), \ldots, (e_k, g_k)\} \tag{2}$$

is the set of all possible translations where by the pair (e_i, g_i) we denote a couple of an English word or phrase with its corresponding translation into Greek. Our filter function selects as the most appropriate translation, e_i, the one with minimum Levenstein distance ($\min ||g - g_i||$) between the original word g and g_i. Other distance measures based on n-grams have been tested but not reported in the present work. In table 1 we present two examples of translating the queries No 8 and No 73 of the OHSUMED database.

3. **Phrase based translation:** Accurate translation demands larger units. Other studies [3] have shown that phrases are a natural way of refining queries. In the phrase based translation an approach is proposed that uses phrases from dictionary as fundamental units for translation of the query. All the phrases in the dictionary that are present in the query are sorted by their size in ascending order and then substituted by their translation counterparts starting from the largest one. The size of a dictionary entry equals to the number of words it contains. To achieve this goal the dictionary was first indexed using the Lucene search engine. For a query Q, the dictionary was searched to find an entry, say Pr, that best matches with Q using the similarity metric

$$sim\,(Pr, Q) = \frac{|Pr \cap Q|}{|Pr|} \tag{3}$$

By $|Pr|$ we denote the number of words of a dictionary entry. From the answers we keep only those with $sim\,(Pr, Q) = 1$. In that case it holds that $Pr \subseteq Q$. However, the similarity function does not ensure that the terms reserve their order inside the phrase and the query. To ensure that words inside Pr and Q have the same order an additional parsing of the query is needed.

From the experimental results it is evident that translation by phrases outperforms all other dictionary-based techniques. Indeed many of the medical terms in the dictionary are compound terms.

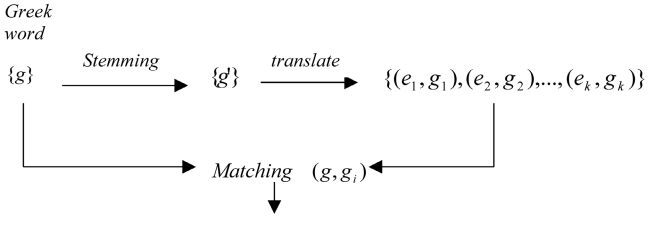

Fig. 2. Filtering of the most appropriate translation

4 Experimental Results

To test the performance of the algorithms proposed we utilized the **OHSUMED** test database[3], a subset of the MEDLINE database, extracted for the monolingual retrieval research [10]. This database is accompanied by a collection of 106 English language queries. We have used the corrected versions of these queries. For all but 5 queries, relevant document subsets are known. We use the 233,445 documents subset that contains abstracts and MeSH phrases for each document.

[3] `ftp://medir.ohsu.edu/pub/ohsumed`

Table 1. Translation of the OHSUMED No.8 and No.73 queries

	Query No.8	**Query No.73**
Original English	*work-up of hypertension in patient with horseshoe kidney*	*portal hypertension and varices, management with TIPS procedure*
Query translated by an expert	Διαγνωστικές εξετάσεις για υπέρταση σε ασθενή με πεταλοειδή νεφρό.	Πυλαία υπέρταση και κιρσοί οισοφάγου, αντιμετώπιση με διασφαγιτιδική ενδοηπατική πυλαιοσυστημική παράκαμψη.
Word by World translation	diagnosis diagnostics diagnostic examination survey examination interrogation test hypertension supreme superlative disease weak patient illness sickness complain asthenia patients inpatient slim kidneys nephron kidney kidneys kidney renal nephritis nephron	portal hypertension supreme superlative varix varicose vein varicose veins esophagitis esophagus oesophagus portosystemic by pass
Word by World translation and disambiguation	diagnostic examination hypertension patient slim weak inpatient kidney	portal hypertension varicose veins oesophagus esophagus portosystemic by pass
Phrase Based Translation	workup hypertension inpatient patient weak slim kidney	portal hypertension varicose veins esophagus oesophagus faced with portosystemic by pass

Table 2. Average precision at the top-k retrieved documents with and without stemming

		No stemming			**Using stemming**		
Top Retrieved Docs	English Queries	WbW Translation	WbW and Disambiguation	Phrase Based Translation	WbW Translation	WbW and Disambiguation	Phrase Based Translation
5	0.3623	0.1283	0.1283	0.1509	0.1811	0.2264	0.2340
10	0.3142	0.1189	0.1189	0.1406	0.1689	0.2170	0.2198
20	0.2660	0.1090	0.1090	0.1288	0.1434	0.1797	0.1788
100	0.1419	0.0526	0.0526	0.0662	0.0775	0.0944	0.0970
200	0.1002	0.0362	0.0362	0.0471	0.0553	0.0650	0.0692

For our cross language experiments, a fluently English speaking medical doctor has translated the 106 queries first into Greek. The Greek queries are then translated back into English by our automatic methods.

Table 3. Average precision at the top-k retrieved documents from two different stemmers

	AvgPr(%) Word by Word Translation			
Top Retrieved Docs	English Queries	Stemmer-1	Stemmer-2	Improvement
5	36.23	18.11	20.00	5%
10	31.42	16.89	19.34	7.8%
20	26.60	14.34	16.70	8.8%
100	14.19	7.75	9.07	9.3%
200	10.02	5.53	6.19	6.6%

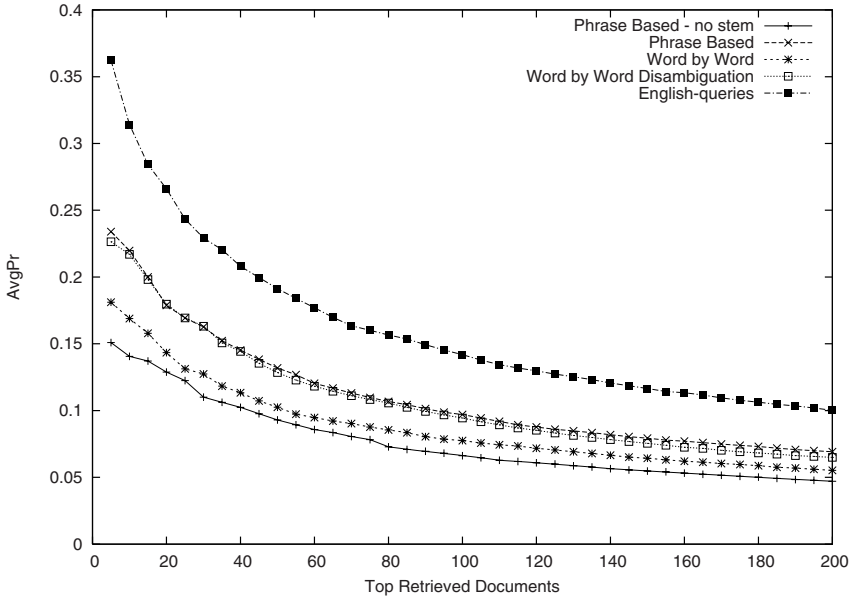

Fig. 3. Average precision plot with respect to the top-k retrieved documents

For the evaluation of the performance the well-known measure of precision was used on the top retrieved documents. Precision is defined by [7]:

$$Precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Number\ of\ documents\ retrieved} \tag{4}$$

In tables 2 and 3 we present values of the averaged precision for the top-k ranked documents over all the queries. The performance of the cross language retrieval is evaluated against the same system running in a monolingual mode (English collection English queries), which serves as the base line of our evaluation. Figure 3 presents visually the performance of the algorithms.

In table 3 results are presented from the word-by-word translation method with the two stemmers mentioned above. From these results it is evident that stemmer-2 performs best in the case of word-by-word translation. In the other two cases, (phrase-based translation and word by word with disambiguation) both stemmers perform equivalently. This sounds reasonable since the disambiguation step removes the uncertaintly introduced by the use of an aggressive stemmer while when larger units are used in the translation, like phrases, the ambiguity is kept to the minimum.

According to the results it is apparent that phrase-based retrieval is best performing achieving a performance between 65%-70% of the corresponding monolingual retrieval.

5 Conclusions-Extensions

The main issue addressed here is the evaluation of an approach to remove disambiguation introduced by the stemming. According to our results it is apparent that stemming is an important part on a CLIR system. Query words are morphologically reduced to their root forms and then substituted by their counterparts in the target language through the dictionary. To reduce the ambiguity due to morphology we have introduced a double translation filter and to reduce the ambiguity due to the translation we used phrases as basic units for translation. Although we are making use of resources freely available in the web, the resulting performance of the algorithms tested is quite fair and comparable to other published results from counterpart approaches.

The retrieval model we have described at its present state is the simplest one and it makes no use of semantic knowledge of terms. Certainly the effectiveness of retrieval on a specific domain, such as medicine, can be improved when domain knowledge is used. Such knowledge may contain synonymous terms and phrases, broader or narrower terms, related terms etc. This is an ongoing research, and we are currently translating a part of the MeSH metathesaurus in the cardiovascular domain, that will be used for query expansion.

Acknowledgements

We thank Dr. D. Soulis for his excellent job of translating the OHSUMED-database queries into Greek.

This work was partially funded by EU, ASIA ICT/TIME project and partially by the Greek Secretariat of Research and Technology, Image, Speech and Language Processing, Action 3.3, MedAS project.

References

1. Hollink, V., Kamps, J., Monz, C., de Rijke, M.: Monolingual document retrieval for european languages. Information Retrieval 7, 33–52 (2004)
2. Kalamboukis, T.: Suffix stripping with modern greek. Program 29(3), 313–321 (1995)

3. Ballesteros, L., Croft, W.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of the 20th ACM SIGIR Conference, pp. 84–91 (1997)
4. Hull, D., Grefenstette, G.: Querying across languages: A dictionary–based approach to multilingual information retrieval. In: H.P., F., D., H., P., S., R., W.(eds.) Proceedings of the 19th International Conference on Research and Development in Information Retrieval (ACM SIG/IR 1996), pp. 49–57 (1996)
5. Dumais, S., Letsche, T., Littman, M., Landauer, T.: Automatic cross–language retrieval using latent semantic indexing. In: Hull, D., Oard, D. (eds.) 1997 AAAI Symposium on Cross–Language Text and Speech Retrieval (1997), http://www.clis.umd.edu/dlrg/filter/sss/papers/dumais.ps
6. Sheridan, P., Wechsler, M., Schauble, P.: Cross language speech retrieval. In: Belkin, N., Narasimhalu, A., Willett, P. (eds.) Proceedings of the 20th International Conference on Research and Development in Information Retrieval (ACM SIGIR 1997), pp. 99–109 (1997)
7. Salton, G., Wu, H., Yu, C.: Measurement of term importance in automatic indexing. J. Am. Soc. Inf. Sci. 32, 175–186 (1981)
8. Pirkola, A., Hedlund, T., Keskustalo, H., Jarvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval 4, 209–230 (2001)
9. Holton, D., Mackridge, P., Filippaki-Warburton, E.: Greek Grammar. Patakis Editions (2006)
10. Hersh, W., Buckley, C., Leone, T., Hickam, D.: Ohsumed: An interactive retrieval evaluation and new large test collection for research. In: Croft, B., van Rijsbergen, C. (eds.) Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval (ACM SIG/IR 1994), pp. 192–200 (1994)
11. Ballesteros, L., Croft, W.: Dictionary methods for cross-lingual information retrieval. In: 7th Conference and Workshop on Database and Expert Systems Applications, pp. 791–801 (1996), http://ciir.cs.umass.edu/info/psfiles/irpubs/ir.html