2008

Experiments with English-Persian Text Retrieval

Abolfazl Ale Ahmed
University of Tehran

Hadi Amiri
University of Tehran

Masoud Rahgozar
University of Tehran

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

Publication Details

This conference paper was originally published as AleAhmad, A, Amiri, H, Rahzogar, M and Oroumchian, F, Experiments with English-Persian text retrieval, in Proceedings of iNews'08 - the 2nd ACM workshop on improving non english web searching, 30 October 2008, California, 77- 80.

Experiments with English-Persian Text Retrieval

Abolfazl AleAhmad, Hadi Amiri, Masoud Rahgozar
Database Research Group
School of Electrical and Computer Engineering
Campus #2, University of Tehran, North Kargar St., Tehran, Iran
+982188338507
{a.aleahmad,h.amiri}@ece.ut.ac.ir, rahgozar@ut.ac.ir

Farhad Oroumchian
Department of Computer Science
University of Wollongong in Dubai
Knowledge Village, Dubai, U.A.E
+97143672400
oroumchian@acm.org

ABSTRACT
As the number of non-English documents is increasing dramatically on the web nowadays, the study and design of information retrieval systems for these languages is very important. The Persian language is the official language of Iran, Afghanistan and Tajikistan and is also spoken in some other countries in the Middle East, so there are significant amount of Persian documents available on the web. In this study, we will present and compare our English-Persian cross language text retrieval experiments on Hamshahri text collection. Also, we will present Combinatorial Translation Probability (CTP) calculation method for query translation that estimates translation probabilities based on the collection itself.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information filtering, Query formulation, Retrieval models, Search process.

General Terms
Algorithms, Performance, Experimentation, Languages.

Keywords
Persian English cross language, Farsi bilingual text retrieval.

1. INTRODUCTION
The Persian language (also know as Farsi) is the official language of Iran, Tajikistan and Afghanistan and also second language of some other countries in the Middle East. As a result of the special and different nature of the Persian language compared to other languages like English and even Arabic, the design of information retrieval systems in Persian requires special considerations. But in spite of special characteristics of the Persian language, little efforts have been focused on retrieval of Persian text compared to other languages.

The difference between the Persian text retrieval and retrieval of other languages on the web is not just for linguistic characteristics of the languages but also because of the culture and behavior of the Persian users on the web. As an example, it is pertinent to mention the recall problem with which modern search engines are faced while queries are formulated by a Persian user. Persian uses Arabic-like script for writing and consists of 32 characters that are written mostly continuously from right to left. The problem is about two specific characters that are Kaf and Ya, each of which has four different shapes, depending on their position in words that are shown in table 1.

Table 1. Kaf and Ya Arabic words

Name	Position in Words			
	Final	Medial	Initial	Isolated
Kaf كاف	ک or ك	ک	ک	ک or ك
Ya يا	ی or ي	ي	ي	ی or ي

Kaf has two forms that are ك and ک with Unicode 1603 and 1705, also Ya has two forms that are ي (with two dots underneath) and ی (without dots) with Unicode 1610 and 1740 respectively. There is a consensus among Arab writers on the web about just using the first form of the two characters. But Persian authors of the web documents use the two forms of the characters equivalently. In order to clarify the problem, we tried a number of queries on two different search engines on 20 July 2008 and the results are summarized in table 2.

Table 2. Queries tried with different search engines

Query No.	Query Title	Search Engine	No. of documents found
1	الحليب (using ي)	Google	1,580,000
		MSN	199,000
2	الحليب (using ی)	Google	431
		MSN	10
3	شير (using ي)	Google	8,710,000
		MSN	259,000
4	شیر (using ی)	Google	1,660,000
		MSN	40,200

In table 2, queries number 1 and 2 are Arabic translation of the word 'Milk' that we ran as a sample query on the search engines. Also, queries number 3 and 4 are Persian translation of the same word. But queries 1 and 3 are written with the first form of the Ya character. One can conclude from table 2 that although queries 3 and 4 are the same from the users' perspective but the search engines return completely different set of results, however this problem can be solved by a simple preprocessing step. We also

iNEWS'08, October 30, 2008, Napa Valley, California, USA.

tried the queries on Yahoo and it returned consistent results that show it processes Persian queries well.

In this paper we will present our method and results of English-Persian text retrieval on Hamshahri collection. This collection is the largest Persian collection built so far which have been used by many researchers. In addition, as there is no parallel or comparable English-Persian corpora available for research, we will present combinatorial translation probability calculation method that uses terms cooccurrence in the collection for calculation of translation probabilities. Remaining parts of this paper are organized as follows: section 2 is an overview of other works on the Persian language, Section 3 describes our method for English-Persian text retrieval and the experimental results are discussed in section 4. Finally we conclude our paper in section 5.

2. RELATED WORKS ON PERSIAN

In this section we will review other researches that we have found in the literature. The authors in [10] proposed the design and testing of a Fuzzy retrieval system for Persian (FuFaIR) with support of Fuzzy quantifiers in its query language and their experiments showed that the retrieval system outperforms the vector space model. Also experiments in [2] on Hamshahri collection [7] suggest the usefulness of language modeling techniques for Persian. Furthermore in [1] the authors evaluated vector space model on Persian text with different weighting schemes and show that N-gram vector space model using Lnu-ltu weighting with slope 0.25 produces good results.

Furthermore, Compression of Persian text for web was assessed on Hamshahri collection in [12] in which the authors could obtain up to 52% reduction in size. Persian text classification is investigated on the collection in [3] and in [8] the authors have applied a reranking method, called local cluster analysis, after an initial retrieval step on the collection that shows considerable improvements on precision of the retrieval system.

3. ENGLISH-PERSIAN TEXT RETRIEVAL

There exists some monolingual Persian text retrieval studies in the literature [1, 2, 10, 8] but we could find no cross language text retrieval research on Persian. In this section we present our method for cross language text retrieval on Persian texts. More precisely, we mean the retrieval of Persian documents based on queries formulated by a human using the English language. There exist three different approaches for bilingual text retrieval according to Oard et al [11]:

1- Thesaurus-based approaches
2- Corpus-based approaches
3- Modular use of machine translation

Our method goes into the first category because we use a bilingual dictionary for translation of the query terms. Let M be the number of query terms, then we define users query as:

$$Q = \{q_i\} \ (i = 1,...,M)$$

As the users query is expressed in English and the collection's documents are in Persian, we use an English-Persian dictionary with 50,000+ entries that we have prepared in our lab for translation of the query terms. Also we use a light stemmer to help matching of the query terms with the dictionary entries. If we define T as the translation function that returns Persian translations set of a given English term q_i, then we have $|T(q_1)|\times|T(q_2)|\times\cdots\times|T(q_M)|$ different translations for the query Q and as one can expect $|T(q_i)|>1$ for most of query terms. So, we need a retrieval model which enables us to take translation probabilities into consideration. This model is introduced in section 3.1 and in section 3.2 we propose our method for translation probability calculation. Also, our experimental results are presented in section 4.

3.1 Probabilistic Structured Query Method

Information retrieval systems rely on two basic statistics: the number of occurrences of a term in a document (Term Frequency or TF) and the number of documents in which a term appears (Document Frequency or DF). In case of bilingual text retrieval, when no translation probabilities are known, Pirkola's "structured queries" have been repeatedly shown to be among the most effective known approaches when several plausible translations are known for some query terms [13]. The basic idea behind Pirkola's method is to treat multiple translation alternatives as if they were all instances of the query term. Darwish and Oard later extended the model to handle the case in which translation probabilities are available by weighting the TF and DF computations, an approach they called probabilistic structured queries (PSQ) [6]. They found that Pirkola's structured queries yielded declining retrieval effectiveness with increasing numbers of translation alternatives, but that the incorporation of translation probabilities in PSQ tended to mitigate that effect. In our bilingual text retrieval experiments we use the PSQ method [6] in which TF and DF are calculated as follows:

$$TF(e, D_k) = \sum_{f_i} p(f_i \mid e) \times TF(f_i, D_k)$$
$$DF(e) = \sum_{f_i} p(f_i \mid e) \times DF(f_i) \qquad (9)$$

Where $p(f_i|e)$ is the estimated probability that e would be properly translated to f_i. Our method for calculation of the translation probability is presented in the next section.

3.2 Combinatorial Translation Probability

Translation probability is generally estimated from parallel corpus statistics. But as no parallel corpus is available for Persian, in this section we introduce a method which estimates English to Persian translation probabilities by use of the Persian collection itself. As most user queries contain more than two terms (e.g. in Hamshahri collection all queries has two or more terms), the main idea is to use co-occurrence probability of terms in the collection for translation probability calculation of adjacent query terms.

Consider M as the number of user's query terms then we define the users query as $Q = \{q_i\}$ $(i=1,...,M)$. For translation of Q, we look up Q members in an English to Persian dictionary to find their Persian equivalents. Considering T as the translation function, then we define set of translations of Q members as:

$$E = \{T(q_1), T(q_2), ..., T(q_M)\}$$

Then the probability that two adjacent query terms q_i and q_{i+1} are translated into $E[i,x]$ and $E[i+1,y]$ respectively, is calculated from the following equation:

$$P(q_i \rightarrow E[i,x] \wedge q_{i+1} \rightarrow E[i+1,y]) = \frac{|D_{qi} \cap D_{qi+1}|}{c + Min(|D_{qi}|,|D_{qi+1}|)} \quad (x = 1..|T(q_i)|, y = 1..|T(q_{i+1})|) \quad (10)$$

Where D_{qi} is a subset of collection's documents, D that contains the term q_i and the constant c is a small value to prevent the denominator to become zero. In the next step we create translation probability matrix W_k for each pair of adjacent query terms:

$$W_k = \{w_{m,n}\} \ (m = 1..|T(q_k)|, n = 1..|T(q_{k+1})|)$$

Where $w_{m,n}$ is calculated using equation (10). Then Combinatorial Translation Probability (CTP) is a $|T(q_1)| \times |T(q_M)|$ matrix that is calculated by multiplication of all of the W_k matrices:

$$CTP(Q) = W_1 \times ... \times W_{k-1} \ (k = 1..M)$$

In other words, CTP matrix contains translation probability of Q members into its different possible translations in Persian. Then given $CTP(Q) = W_1 \times ... \times W_{k-1} \ (k = 1..M)$, the algorithm in table 3 returns the TDimes matrix which contains dimensions of $E = \{T(q_1), T(q_2), ..., T(q_M)\}$ matrix that correspond to top n translations of the query $Q = \{q_i\}$ $(i=1,...,M)$.

Having TDimes matrix, we are able to extract different translation of the users query from $E = \{T(q_1), T(q_2), ..., T(q_M)\}$ and their weight from CTP. For example if we consider an English query that has three terms then the most probable Persian translation of the query terms would be *E[1,TDimes [1,1]]*, *E[2,TDimes [1,2]]* and *E[3,TDimes [1,3]]* respectively and the translated query's weight would be *CTP[TopColumns[1],TopRows[1]]*.

Table 3. Calculation of the TDimes matrix

1. Let *TopRows[n]* be the row number of *n* largest members of CTP
2. Let *TopColumns[n]* be the column number of *n* largest members of CTP
3. For i ← [1,…,n]
   - 3.1. Let R = TopRows [i]
   - 3.2. Let C = TopColumns [i]
   - 3.3. TDimes[i,M] = C
   - 3.4. For j ← [M-1,…,1]
     - If (j=1)
       - Let TDimes[i,j] = R
     - else
       - Let TDimes [i,j] = the culomn number of the largest element of Rth row of W_{i-1}
4. Output the *TDimes* matrix

4. EXPERIMENTAL RESULTS

4.1 Test Collection

Experimentations of this paper are accomplished by use of Hamshahri collection [7] that is now included in cross language evaluation forum (CLEF) [5] collections and to the best of our knowledge it is the largest Persian test collection. Hamshahri collection contains more that 160,000 documents which are actually news articles of the Hamshahri newspaper from year 1996 to 2002. Table 4 depicts some attributes of the collection. It should be noted that this collection have already been preprocessed in order to overcome the recall problem that we have discussed in section 1.

Table 4. Attributes of Hamshahri collection

Attributes	Value
Collection size (Unicode)	564 MB
No. Of documents	166,774
No. Of unique terms	417,339
Average length of documents	380 Terms
No. Of topics in the third topic set	50

In our experiments we use the third topics set of the collection which are prepared based on TREC specifications. TREC uses a technique called pooling [4] in which a pool of subset of documents is created for each topic and is judged for relevance by the topic author. The pooling technique prevents the relevance judgment process from biasing toward a retrieval system [14]. The topic set was also used as training set in cross language evaluation forum 2008 and contains 50 queries in both English and Persian and their relevance judgments. Each topic consists of tree parts: Title, Description and Narrative. Title contains up to two or three words which give the main idea of the topic. Description contains a full sentence or question describing the topic in short. Narrative contains a broader description of the topic including examples and perhaps mentions aspects that should not be counted as relevant.

4.2 Bilingual Text Retrieval Results

This section presents our bilingual English-Persian information retrieval results. In our experiments we just use the title of the topics. In addition, we used the Lemur toolkit for implementation of our algorithm (http://www.lemurproject.org/). The default retrieval model of the lemur's retrieval engine (Indri) is the language modeling.

Our bilingual experimental results include three different runs that are summarized in table 5. Figure 1 depicts the precision-recall graph of their results that are calculated by use of the Treceval tool [9]. In the first run we retrieve documents based on Persian title of the 50 topics as a base line to compare other results with it. Our second run uses the English version of the 50 topics' titles and we translate them by use of the English-Persian dictionary. Then we formulate a query in Persian by concatenation of all meanings of each of the query terms.

The Third run uses the retrieval model of section 3.1 and calculation of translation probabilities with method of section 3.2. In other words, we use the English version of the 50 topics' titles and apply the method of section 3.2 for translation of the topics and calculation of translation probabilities. Then we formulate Persian equivalents of the 50 English queries based on top 5 translations with highest probability.

The Indri retrieval engine supports structured queries, so we could easily implement the PSQ method using CPT for weighting.

Table 5. Runs and their description

Run#	Run Name	Description
1	Monolingual	Monolingual LM retrieval model
2	All Meanings	Bilingual English-Persian LM retrieval model using all meanings of each of the query terms
3	PSQ CPT Top5	Bilingual English-Persian LM retrieval model using methods of 3.1, 3.2 sections

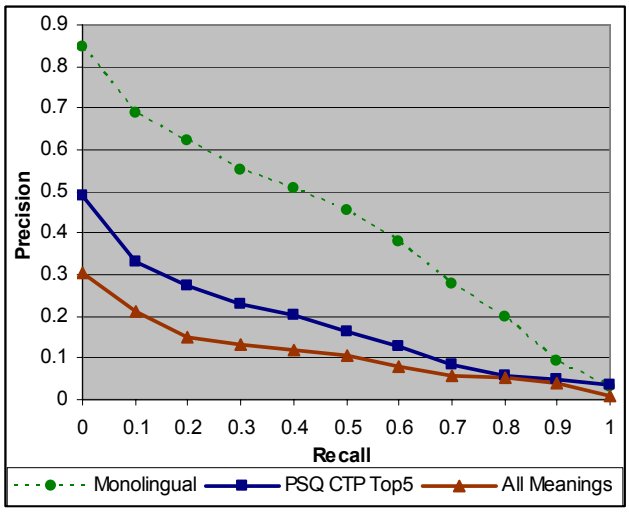

Figure 1. Precision-Recall of the three runs

Comparison of the three runs shows that performance of the PSQ CTP Top5 run is considerably better that the All Meanings run and our proposed method can work well in case of lack of language resources like parallel corpora. Also, performance of PSQ CTP Top5 is comparable with performance of other cross language text retrieval results that other researchers have reported on other languages [6].

5. CONCLUSION AND FUTURE WORKS

In this paper we proposed a method for calculation of translation probabilities based on statistics of the collection itself. The results that we obtained form this method shows that it is able to improve cross language retrieval performance in cases that we do not have enough language resources for calculation of translation probabilities.

Hamshahri collection is now standardized according to CLEF standards and 50 new bilingual queries are developed for the collection recently that facilitates study of bilingual English-Persian text retrieval. So, one of our future work would be investigation of other aspects of cross language information retrieval on the Persian language.

6. ACKNOWLEDGEMENT

This research was partially supported by Iranian Telecommunication Research Center (ITRC). Also, we would like to thank members of DBRG lab at University of Tehran that helped us in this research.

7. REFERENCES

[1] A. Aleahmad, P. Hakimian, F. Mahdikhani, F. Oroumchian. N-Gram and Local Context Analysis for Persian Text Retrieval. International Symposium on Signal Processing and its Applications ISSPA 2007, Sharjah, UAE, 2007.

[2] H. Amiri, A. AleAhmad, F. Oroumchian, C. Lucas, M. Rahgozar, Using OWA Fuzzy Operator to Merge Retrieval System Results. In Computational Approaches to Arabic Script-based Languages (CAASL 2007), Stanford University, USA, 2007.

[3] B. Bina, M. Rahgozar, A. Dehmoubed. Automatic Classification of Persian Text. In 13th International CSI Computer Conference (CSICC) 2008, Kish island, Iran, 2008.

[4] N. Craswell, D. Hawking, R. Wilkinson, and M.Wu. Overview of the TREC 2004 web track. In Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), 2004.

[5] Cross Language Evaluation Forum (CLEF). http://www.clef-campaign.org/

[6] Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In Proceedings of the 21st Annual 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 338–344. ACM Press, July 2003.

[7] Hamshahri Collection. http://ece.ut.ac.ir/dbrg/hamshahri/

[8] A.H. JadidiNezhad, H. Amiri. Local Cluster Analysis as a Basis for High-Precision Information Retrieval. The 6th International Conference on Informatics and Systems, INFOS 2008, Cairo, Egypt, 2008.

[9] National Institution of Standards and Technology, http://trec.nist.gov/trec_eval/

[10] A. Nayyeri, F. Oroumchian. FuFaIR: a Fuzzy Persian Information Retrieval System. IEEE International Conference on Computer Systems and Applications, pp. 1126-1130, 2006.

[11] D. W. Oard & B. J. Dorr, A Survey of Multilingual Text Retrieval, UMIACS TR-96-19, University of Maryland, College Park, MD, 1996.

[12] F. Oroumchian, E. Darrudi. Experiments with Persian Text Compression for Web. World Wide Web 2004, pp. 478-479, New York, New York, USA, May 2004.

[13] Ari Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 55–63. ACM Press, August 1998.

[14] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 307–314, Melbourne, Australia. ACM Press, New York, August 1998.