Cross Language Experiments at Persian@CLEF 2008

Abolfazl AleAhmad, Ehsan Kamalloo, Arash Zareh, Masoud Rahgozar

Database Research Group
School of Electrical and Computer Engineering
{a.aleahmad, e.kamalloo, a.zareh}@ece.ut.ac.ir,
rahgozar@ut.ac.ir

Farhad Oroumchian

Department of Computer Science
University of Wollongong in Dubai
oroumchian@acm.org

Abstract
In this study we will discuss our cross language text retrieval (CLIR) experiments of Persian ad hoc track at CLEF 2008. Two teams from University of Tehran were involved in cross language text retrieval part of the track using two different CLIR approaches that are query translation and document translation. For query translation we used a method named Combinatorial Translation Probability (CTP) calculation for estimation of translation probabilities. In the document translation part we used the Shiraz machine translation system for translation of documents into English. Then we create a Hybrid CLIR system by score-based merging of the two retrieval system results. In addition, we investigated N-grams and a light stemmer in our monolingual experiments.

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

Keywords
Persian English cross language, Farsi bilingual text retrieval.

1. Introduction
The Persian language is categorized as a branch of Indo-European languages and is the official language of Iran, Afghanistan and Tajikistan and is also spoken in some other countries in the Middle East. Morphological analysis of the language is relatively hard because of its grammatical rules. For example the word "خبر" is an Arabic word that is used in Persian. This word has two plural forms in Persian "اخبار" and "خبرها", the first plural form obeys Arabic grammatical rules and the second plural form is obtained by use of Persian rules.
After creation of 50 new bilingual topics and standardization of Hamshahri collection according to CLEF standards, we could investigate CLIR on Persian. Persian@CLEF 2008 is our first attempt to evaluate cross language information retrieval on the language. Our aim is to investigate two main approaches of cross language text retrieval on Persian that are query translation and document translation.
We used the Hamshahri collection [7] for evaluation of our retrieval methods. Documents of this collection are actually news articles of Hamshahri newspaper from year 1996 to 2002. The collection contains 160,000+ documents from variety of subjects. The documents size varies from short news (under 1 KB) to rather long articles (e.g. 140 KB) with the average of 1.8 KB. Also we used Apache Lucene [8] and Lemur toolkit [5] for indexing and retrieval on the collection.
The remaining parts of this paper are organized as follows: section 2 introduces our monolingual experiments, section 3 discusses our query translation method and its results, section 4 contains document translation experimental results and finally we will conclude our paper in section 5.

2. Experiments on Monolingual Persian Text Retrieval
We had no efficient morphological analyzer for Persian, so in our monolingual experiments we tried to investigate some alternative methods like n-grams. Also, we used a stop word list in monolingual part of our experiments to improve retrieval results.
In order to create the stop word list we manually inspected most frequent words of the collection and extracted actual stop words. Then we added some other words from the Bijankhan Persian corpus [6] that were marked with tags like proposition and conjunction. The final stop word list contains 796 items.
In our monolingual experiments, we submitted top 100 retrieved documents of six monolingual runs that are summarized in table 1 and their description is as follows:

- *Run #1*: Vector space retrieval model using a light stemmer
- *Run #2*: Term based vector space model retrieval

- *Run #3*: Using 3-grams with Language Modeling retrieval
- *Run #4*: Using 4-grams with Language Modeling retrieval
- *Run #5*: Using 5-grams with Language Modeling retrieval
- *Run #6*: Term-based Language Modeling retrieval

Table 1. Persian monolingual retrieval systems

Run#	Run Name	tot-ret	rel-ret	MAP	Retrieval Model	Retrieval System
1	SECMLSR	5161	1967	26.89	Vector Space	Lucene
2	SECMLUSR	5161	1991	27.08	Vector Space	Lucene
3	UTNLPDB1M3G	5161	1901	26.07	Language Modeling	Lemur
4	UTNLPDB1M4G	5161	1950	26.70	Language Modeling	Lemur
5	UTNLPDB1M5G	5161	1983	27.13	Language Modeling	Lemur
6	UTNLPDB1MT	5161	2035	28.14	Language Modeling	Lemur

In all of these runs we used just title part of the 50 Persian topics that was made available at CLEF 2008. In the first run, we used a light Persian stemmer that works like the Porter algorithm but it could not improve our results because of the simple algorithm of the stemmer. As an example consider the word "فیلم" that was a term in topic no 559. This word is a noun that means 'film' in English but our light stemmer considers the final 'م' letter of the word as a suffix and converts it to 'فیل' that means 'elephant' in English.
Also, it worth mentioning that we do not cross word boundaries for building N-grams. For example 4-gram of the word "ویمبلدون" is " ویمب+ یمبل+ مبلد+ بلدو+ لدون" by use of our method.

3. CLIR by Query Translation

This section illustrates our query translation experiments at Persian ad hoc track of CLEF 2008. As the users query is expressed in English and the collection's documents are written in Persian, we used an English-Persian dictionary with 50,000+ entries for translation of the query terms. In addition, we inserted some proper nouns into the dictionary. The query translation process is accomplished as follows.
Let M be the number of query terms, then we define users query as:

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

Then we looked each q_i up in the dictionary and after finding translations of q_i we split the translations into its constituent tokens. Then we eliminate those tokens that are included in our Persian stop word list.
If we define T as the translation function that returns Persian translations set of a given English term q_i as described above, then we have $|T(q_1)|\times|T(q_2)|\times\cdots\times|T(q_M)|$ different possible 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 briefly introduced in section 3.1 and in section 3.2 we propose our method for translation probability calculation. Then our query translation CLIR experimental results are presented in section 3.3.

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 [1].
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) [2]. 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 [2] in which TF and DF are calculated as follows:

$$TF(e,D_k) = \sum_{f_i} p(f_i \mid e) \times TF(f_i, D_k)$$
$$DF(e) = \sum_{f_i} p(f_i \mid e) \times DF(f_i) \qquad (1)$$

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

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

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:

$$P(q_i \rightarrow E[i,x] \wedge q_{i+1} \rightarrow E[i+1,y]) = \frac{|D_{qi} \cap D_{qi+1}|}{c + Min(|D_{qi}|, |D_{qi+1}|)} \quad (2)$$

$$(x = 1..|T(q_i)|, y = 1..|T(q_{i+1})|)$$

Where D_{qi} is a subset of collection's documents 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 (2). 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:

$$CPT(Q) = W_1 \times \ldots \times W_k \ \ (k = 1 \ldots M-1)$$

In other words, CTP matrix contains probability of translation of Q members into their different possible translations in Persian. Given the $CTP(Q)$ matrix, the algorithm in table 2 returns the TDimes matrix which contains dimensions of $E = \{T(q_1), T(q_2), \ldots, T(q_M)\}$ matrix that correspond to top n most probable translations of the query $Q = \{q_i\}\ (i=1,\ldots,M)$.

Table 2. 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

Having TDimes matrix, we are able to extract different translation of the users query from $E = \{T(q_1), T(q_2), \ldots, T(q_M)\}$ 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]]*.

3.3. Query Translation Experimental Results

We translated the queries through term lookup in an English-Persian dictionary as described before and using methods of section 3.1 and 3.2. All of our query translation experiments were run using title of the English version of the 50 topics except run #8 in which we used title + description of the topics. In this part of our experiments we had eight runs that are summarized in table 3 and their description is as follows:

- *Run #1*: In this run we concatenate all meanings of each of the query terms to formulate a Persian query.
- *Run #2*: The same as previous run but uses top 5 Persian meanings of each of the query words for query translation.
- *Run #3*: The same as previous run but uses the first Persian meaning of each of the query words for query translation.
- *Run #4*: Uses all Persian meanings of query terms for query translation for calculating CTP. Then we used the PSQ method with top 10 most probable Persian translations of the query.
- *Run #5*: In this run we first look up top 5 meanings of query terms in the dictionary and then we convert them into 4-grams for calculating CTP. Then we use PSQ method with top 10 most probable Persian translations of the query to run 4-gram based retrieval.
- *Run #6*: The same as previous run but we use 5-grams instead of 4-grams.
- *Run #7*: This run is the same as run #3 but in this run we use the Lucene vector space retrieval model.
- *Run #8*: This run is the same as run #7 but in this run we use title + description. We eliminate common words such as 'find', 'information', from the topics description.

We used the Lemur toolkit [5] for implementation of our algorithm for run #1 to run #5. The default retrieval model of the lemur's retrieval engine (Indri) is language modeling. The Indri retrieval engine supports structured queries and we could easily implement the PSQ method using CPT for translation probability estimation. Also, run #7 and run #8 are implemented by use of the Lucene retrieval engine.

Table 3. English-Persian query translation experiments

Run#	Run Name	tot-rel	rel-ret	MAP	Dif	Retrieval Model	Retrieval System
1	UTNLPDB1BA	5161	758	6.73	baseline	Language Modeling	Lemur
2	UTNLPDB1BT5	5161	974	10.19	+ 3.46	Language Modeling	Lemur
3	UTNLPDB1BT1	5161	930	12.4	+ 5.67	Language Modeling	Lemur
4	UTNLPDB1BA10	5161	1150	14.07	+ 7.34	Language Modeling	Lemur
5	UTNLPDB1BT4G	5161	1196	14.46	+ 7.73	Language Modeling	Lemur
6	UTNLPDB1BT5G	5161	1166	14.43	+ 7.70	Language Modeling	Lemur
7	CLQTR	5161	677	8.93	+ 2.20	Vector Space	Lucene
8	CLQTDR	5161	592	6.01	- 0.72	Vector Space	Lucene

Also Figure 1 depicts the precision-recall graph of the eight runs for top 100 retrieved documents that are calculated by use of the Trec_Eval tool. According to the 'comparison of median average precision' figure that was released at Persian@CLEF 2008, this method could over perform monolingual retrieval results for some topics like topic no 570. This is because of the implicit query expansion effect of this method. The topic's title is 'Iran dam construction' and after its translation into Persian, the CTP method adds the word 'آب' to the query that means water in English.

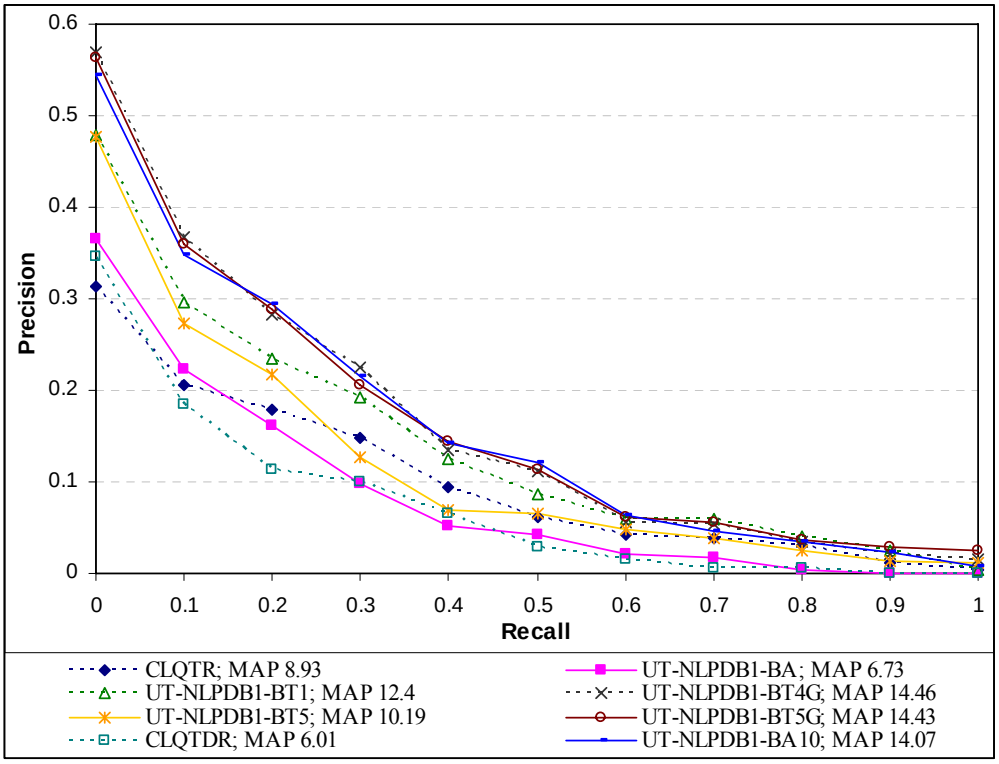

Figure 1. Precision-Recall of the six query translation runs

4. CLIR by Document Translation

In order to translate the Hamshahri collection's documents from Persian into English, we used the Shiraz machine translation system that is prepared at the New Mexico State University [3]. The Shiraz machine translation system is an open source project that is written with the C language [4]. This system uses a bilingual Persian to English dictionary consisting of approximately 50,000 terms, a complete morphological analyzer and a syntactic parser. The machine translation system is mainly targeted at translating news material.

Document translation is not a popular approach because this approach of CLIR is not computationally efficient. This fact was also apparent in our experiments. We ran the Shiraz machine translation on a PC with 2G of RAM and an Intel 3.2G CPU and it took more than 12 days to translate nearly 80 percent of the collection. Finally we could translate 134165 out of 166774 documents of the collection and we skipped translation of long documents to save time. In our document translation experiments we had one run, named CLDTDR, by use of document translation that is described below:

- *Run #9*: In this run we use the English version of the 50 topics of Persian@CLEF 2008. Then we retrieved translated documents of the collection using the Lucene vector space retrieval engine. This run utilizes title + description part of the topics.

Furthermore, we tried a hybrid CLIR method by score-based merging of the results of query translation and document translation methods. For this purpose we used merge results of the CLDTDR and UTNLPDB1BT4G runs. The two runs used different retrieval engines and hence their retrieval scores were not in the same scale. To address this problem we used the following equation to bring the scores of the two retrieval lists into the same scale:

$$Score_i = \frac{x_i - Min(L_{i,q})}{Max(L_{i,q}) - Min(L_{i,q})}$$

In which x_i and $Score_i$ are the old and the normalized scores, $Min(L_{i,q})$ and $Max(L_{i,q})$ are the minimum and maximum scores in the i^{th} retrieved list for the query q (i=1,2 for the two runs). This normalization normalizes the scores into the range [0, 1]. Then for obtaining the merged results we chose top 100 documents with highest weight from the two lists.

Table 4 and Figure 2 show performance of our query translation, document translation and hybrid CLIR systems and compare them with one of our monolingual systems as a baseline.

Table 4. Comparison of CLIR retireval experiments

Run Name	tot-rel	rel-ret	MAP	CLIR/Mono	Retrieval Model	Retrieval System
SECMLUSR	5161	1967	27.08	baseline	Vector Space	Lucene
UTNLPDB1BT4G	5161	1196	14.46	53 %	Language Modeling	Lemur
CLDTDR	5161	1234	12.88	48 %	Vector Space	Lucene
Hybrid CLIR	5161	1478	16.19	60 %	LM + Vector Space	Lemur + Lucene

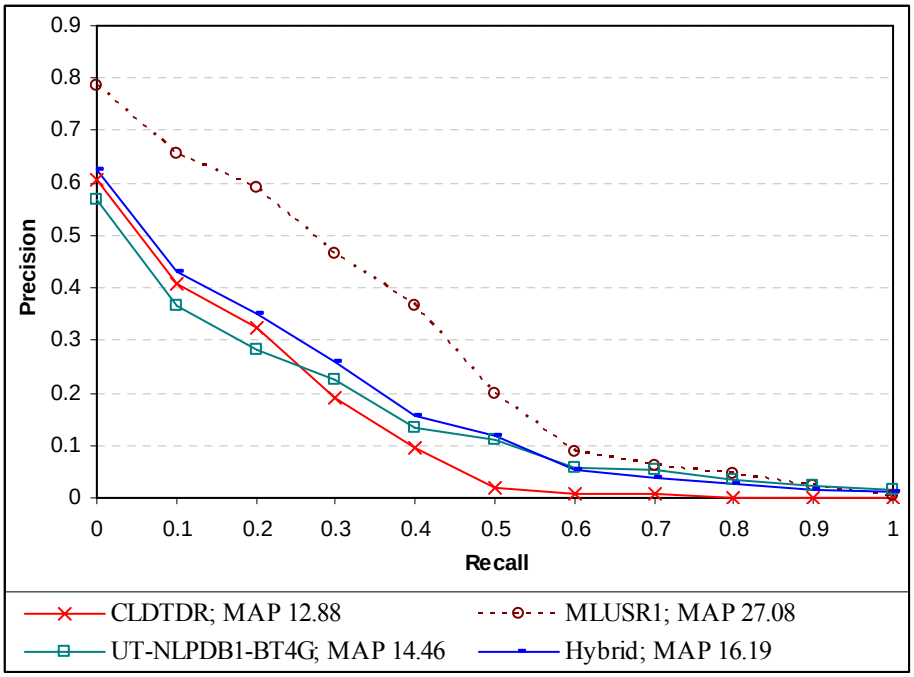

Figure 2. Precision-Recall of CLIR experiments

5. Discussion and Future works

In Persian ad hoc track of ninth CLEF campaign in addition to some monolingual retrieval systems, we evaluated a number of cross language information retrieval systems. In monolingual part of our experiments we evaluated N-grams and a light stemmer on the Persian language and in cross language part we evaluated query translation and document translation approaches of English-Persian cross language information retrieval. We used combinatorial translation probability method for query translation that uses statistics of the target language for estimating translation probabilities. Result of our hybrid cross language information retrieval experiments also suggests usefulness of combining document translation and query translation.

Acknowledgements

We would like to thank CLEF 2008 organizers for their supports in development of the Hamshahri collection.

References

[1] 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.

[2] 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.

[3] Amtrup, Jan W., Hamid Mansouri Rad, Karine Megerdoomian and Rémi Zajac (2000). *Persian-English Machine Translation: An Overview of the Shiraz Project*. NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-319).
[4] Shiraz Project, http://crl.nmsu.edu/Research/Projects/shiraz
[5] Lemur Toolkit, http://www.lemurproject.org/
[6] Bijankhan Corpus, http://ece.ut.ac.ir/dbrg/bijankhan/
[7] Hamshahri Collection, http://ece.ut.ac.ir/dbrg/hamshahri
[8] Apache Lucene project, http://lucene.apache.org/