# A Method for Cross-Language Retrieval of Chunks using Monolingual and Bilingual Corpora

Tayebeh Mosavi Miangah
English Language Department
Payame Noor University
Yazd, Iran

Amin Nezarat
Computer Engineering Department
Shiraz University
Shiraz, Iran

## ABSTRACT
Information retrieval (IR) is a crucial area of natural language processing (NLP) and can be defined as finding documents whose content is relevant to the query need of a user. Cross-language information retrieval (CLIR) refers to a kind of information retrieval in which the language of the query and that of searched document are different.

One of the fundamental issues in bilingual retrieving of information in search engines seems to be the way and the extent users call for phrases and chunks. The main problem arises when the existing bilingual dictionaries are not able to meet the users actual needs for translating such phrases and chunks into an alternative language and the results often are not reliable.

In this paper it has been tried to report the findings extracted from an experiment carried out in this respect to deal with this problem. In this project a heuristic method for extracting the correct equivalents of source language chunks using monolingual and bilingual linguistic corpora as well as text classification algorithms is to be introduced. For this purpose we use a statistical measure known as Association Score (AS) to compute the association value between every two corresponding chunks in the corpus.

The results gained from the experiment carried out in this respect to examine the effectiveness of the heuristic method on extracting all possible chunks in Persian language and finding the most appropriate equivalents for them in English are very encouraging.

## Keywords
chunk retrieval, cross-language information retrieval, linguistic corpora, text classification, Persian language.

## 1. INTRODUCTION
Unlimited and public approachability to this bulk of information have become one of the greatest challenges the specialists in the field of computer sciences need to tackle with. Searching keywords in Internet using search engines and gaining the required outputs and resources in a language other than the search language is a growing need of most users. In this research we tried to use some statistical models based on monolingual and bilingual linguistic corpora and their combination to obtain a disambiguation method for various chunks as keywords entered by users in the search engines.

The traditional method for cross-language retrieval was based on one or more bilingual dictionary used in the time of searching and displaying the outputs. This method, though rapidly spread out, has its own various drawbacks including: limited number of words in a dictionary, incompatibility of the existing words in these dictionaries with the most recent and current words of a language, the various equivalents of a given word in the target language and the way of selecting the most appropriate one, and some others. In fact, the last one of the above mentioned drawbacks seems to be the most important one as so many words and phrase have more than one translation depending on the subject area to which the word or phrase belongs. This problem can frequently be observed in studying the different languages of Middle East and East Asian due to the lexical richness of such languages. To make more clear, the Persian phrase " دوره های مقدماتی" (introductory courses), when searching the Google translation engine, gives back "fundamental courses" as its equivalent. This type of translation which is based on a bilingual dictionary is carried out without considering the subject matter and the context to which the phrase belongs and its contemporary usage, as a result of which the search output is not what is expected to be.

In the present research we are going to demonstrate the effectiveness of our novel heuristic method in automatic extraction of all possible and valid chunks in Persian language, and at the same time selecting the most appropriate translation of each chunk among those equivalents presented by the bilingual parallel corpus. We believe that the unstructured but complete information available in linguistic corpora can provide more precise and relevant responses in retrieval tasks compared to the structured but incomplete information from the existing monolingual and bilingual dictionaries.

## 2. RELATED WORKS
In recent years many researchers have tried to develop high-efficiency systems of information retrieval using various methods, almost all of which were based on linguistic dictionaries. There are quite a number of non-Persian studies carried out in this respect, some of which are to be mentioned here. A research taking the dictionary approach and using retrieval precision demonstrate that word-to-word translating of the queries leads to a 40% - 60% decrease in retrieval efficiency comparing to phrase-to-phrase translation of the same queries [3]. Chen has also examined effect of phrase translation in cross-language information retrieval between Chinese and English. Using the methodology of program evaluation, he found that translation by phrase is more successful than translation by single words. His findings showed that phrase translation with 53% of efficiency compared to word translation with 42% efficiency had a better performance in information retrieval task. He added that the rate of efficiency could be enhanced in case of exploiting some

complementary resources. Doing that, he achieved 83% of efficiency for monolingual information retrieval [2].

Among the works which have been done in cross-language information retrieval for Persian is a study during which Alizade and his colleagues evaluate a system in which only a machine-readable bilingual dictionary was used. Their findings were specified as follows: higher efficiency when 1) using the first equivalent compared to using all equivalent of a given query; 2) morphological processing of all query words before their translation compared to the lack of any kind of processing; 3) adopting the phrase translation procedure compare to word translation of the queries [1]

In another study on cross-language information retrieval, Mosavi Miangah made an attempt to use a bilingual parallel corpus to extract suitable equivalents for query words. There, she reported very encouraging findings regarding the use of a bilingual corpus instead of a bilingual dictionary [4]. The present research is, in fact, in the line of the mentioned study improving methodology in order to deal with phrases and chunks using a large monolingual corpus as well.

# 3. METHODOLOGY

## 3.1 Corpora Used in This Study

For the purpose of this experiment we tried, as the very first stag, to revise and complete our already existing 110-million monolingual corpus of Persian to reach about 200 million words.

In order to increase the number of sentences indexed in the monolingual corpus, a Web Agent Software was provided to act as a Robot which looks through the Persian Web environments and extracts Persian sentences among these Persian pages. Subject classification of each sentence is automatically determined mapping the subject of the website to which the given sentence belongs. For example, when our browser enters a News Agency Website, all sentences extracting from it are indexed as "politics" in the corpus.

The corpus is comprehensive in the sense that it has been divided into different sub-corpora of various text types as politics, medicine, poetry, sport, literature, art, idioms and proverbs, religion, science, culture, history, economics, and miscellaneous. These texts are mainly extracted from books, journals, interviews, reports, written news, etc. but the main contribution goes with the online version of Hamshahri newspaper[1]. All the texts are to be processed before entering to the corpus. That is, all tables, pictures, figures or diagrams are to be deleted from the texts to be ready for the corpus. Moreover, the texts should be converted to an XML[2] format to be suitable for use on Internet sites.

In this stage the texts entered the pre-designed corpus distinguished by the text type and then all the texts are automatically segmented at the sentence level. While entering each text, its type or specialized field is determined and indexed in order for the type to be appeared next to the sentences when retrieving them (to which type each sentence belongs).

[1] - http://www.Hamshahrionline.ir

[2] - Extensible Markup Language

## 3.2 Bilingual Parallel Corpus

The English-Persian parallel corpus has been compiled as a bilingual textual database consisting of aligned original English texts and their translations into Persian, and of original Persian texts and their translations into English. Although the availability of bilingual texts involving Persian is subject to some limitations due to the low density of this language around the world and the unavailability of texts in some specific genres and domains, we first succeeded in collecting a relatively large number of texts, totaling about 3,500,000 words in English and Persian [5]. For the purpose of t is experiment, the number of words in this corpus reached about 4,500,000 words resorting to various methods such as hiring translators to do the job and the like.

All texts in this corpus have been manually aligned at the sentence level. Although we could use various automatic methods for aligning sentences, we prefer to do this manually in order for the corpus to be highly reliable and without any noise. The corpus includes a variety of genres such as literature, sport, medicine, politics, science, culture, and the like.

After complementing the two monolingual and bilingual corpora in the above mentioned manner, we set to exploit them as the main material of the present experiment.

## 3.3 Automatic Extraction of Chunks

In order to disambiguate the search inputs, first the association score among all components of a sequence of words should be determined. Then, the gained association score will decide on the probability degree of the sequence as an acceptable chunk or phrase occurring in the given language. For calculating the association score, we made use of text classification methods. In these methods, the degree of association or disassociation of several words in a sentence is introduced as probability functions. For instance, the association score of d € X, where X is a set of sentences, to the class of

$C = \{c_1, c_2, \ldots c_j\}$ can be written as follows:

< d, c > = < Tehran joins the World Trade Organizational, Iran >

In order to examine the association score between d and c, a formula for calculating association probability of $X^2$ is to be used. In this formula, the association degree of c and d is calculated using probability functions of P(DC) = P(D) P(C) as follows:

$$x^2(d,c) = \sum_{e_d \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_d e_c} - E_{e_d e_c})^2}{E_{e_d e_c}}$$

in which $e_d$ indicates the occurrence of the chunk d in the sentence, and $e_c$ indicates the occurrence of the chunk c in the sentence. The quantity of E is also calculated as follows:

$$E_{e_d e_c} = N \times \frac{(N_{e_d e_c} + N_{e_d 0})}{N} \times \frac{(N_{e_d e_c} + N_{0\ e_c})}{N}$$

where N is the total number of sentences in the corpus, and $N_{t,c}$ is the number of occurrence of the chunks t and c in the corpus, so that $N_{11}$ equals the number of simultaneous occurrence of both chunks in the sentence, and $N_{02}$ equals the number of occurrence of the second chunk without the first one. Now, for calculation of

the association score between the two chunks c and t, the following formula is applied [6]:

$$x^2 = \frac{(N11 + N10 + N01 + N00) \times (N11\,N00 - N10\,N01)2}{(N11 + N01) \times (N11 + N10) \times (N10 + N00) \times (N01 + N00)}$$

After computing the frequency $N_{dc}$ and putting it in the above formula and calculating $x^2(d, c)$, the highness or lowness of the association level of the chunk elements is determined using the table 1, and threshold of $x^2$ which equals 6.63.

**Table 1. CRITICAL VALUES OF THE $x^2$ DISTRIBUTION WITH ONE DEGREE OF FREEDOM**

| P | $X^2$ (Critical Value) |
|---|---|
| 0.1 | 2.71 |
| 0.05 | 3.84 |
| 0.01 | 6.63 |
| 0.005 | 7.88 |
| 0.001 | 10.83 |

Therefore, if $x^2$ is smaller than the threshold of 6.63, it means there is some degree of association between two given elements and other quantities greater than the threshold would be rejected as implying no significant association in this respect.

## 4. IMPLEMENTING THE METHOD ON MONOLINGUAL CORPUS

In order to determine the association score between different components of a chunk searched by a user, different relationships inside the chunk are to be valuated using $x^2$ formula. Consider, for instance, the chunk "یک روز در میان" (every other day) is entered as a search input by a user. As the first stage, the association score between every two words of the given chunk are calculated one by one using their relative frequencies in the monolingual corpus of Persian. Then, using the $x^2$ formula the degree of association between all components of the chunk is calculated and the degree of probability of the chunk as an acceptable one in Persian language is gained referring to threshold table above. After calculating $x^2$ , using the software implemented on the monolingual corpus of Persian, the critical value of 4,94 was gained which is smaller than the threshold and thus can be accepted as a valid and acceptable chunk in Persian language.

## 5. EXTRACTING THE CORRECT EQUIVALENTS OF CHUNKS USING BILINGUAL CORPUS

When the valid chunks were extracted using information gained from the monolingual corpus of Persian, it's time to find the most suitable equivalent of this Persian chunk in English using the English-Persian bilingual parallel corpus. Again, the association score is to be used. For this purpose, first all Persian sentences including such a chunk are retrieved from the bilingual corpus, and then for each single sentence the association probability of the chunk having different combinations of words inside the sentence is calculated. Take, for example, the following Persian sentence and its corresponding English sentence found in bilingual corpus:

1. کمیسیون یک روز در میان تشکیل جلسه می دهد.
2. The committee convenes every other day.

Now, using a rather complicated algorithm designated by the authors, we set to break the English and Persian sentences into all possible chunks shown roughly in Tables 2 and 3. Gaining the association scores of al possible chunks presented in table 2, the valid ones have been starred among which No. 6 in table 2 and No. 18 in table 3 are translations of each other.

**TABLE 2.**
**ALL POSSIBLE CHUNKS FOR THE PERSIAN SENTENCE 1 FOR CALCULATING $x^2$**

| No | Chunk |
|---|---|
| 1 | The committee |
| 2 | The committee convenes |
| 3 | The committee convenes every |
| 4 | committee convenes |
| 5 | committee convenes every |
| 6 | committee convenes every other |
| 7 | convenes every |
| 8 | convenes every other |
| 9 | convenes every other day |
| 10 | every other |
| **11** | **every other day** |
| 12 | other day |

Then, $x^2$ is calculated for every corresponding English and Persian chunks separately. Afterwards, a software, designed for this purpose, begins to delete improbable cases (based on the threshold defined in table 1). Finally, among those remaining probable cases, the chunk with the greatest $x^2$ (association score) is selected as the most appropriate equivalent for the given Persian chunk.

**TABLE 3.**
**ALL POSSIBLE CHUNKS FOR THE ENGLISH SENTENCE 2 FOR CALCULATING $x^2$**

| $x^2$ < 6.63 | Chunk | No |
|---|---|---|
| | کمیسیون یک | 1 |
| | کمیسیون یک روز | 2 |

| | | |
|---|---|---|
| | کمیسیون یک روز در | 3 |
| * | یک روز | 4 |
| | یک روز در | 5 |
| * | **یک روز در میان** | 6 |
| | روز در | 7 |
| | روز در میان | 8 |
| | روز در میان تشکیل | 9 |
| | در میان | 10 |
| | در میان تشکیل | 11 |
| | در میان تشکیل جلسه | 12 |
| | میان تشکیل | 13 |
| | میان تشکیل جلسه | 14 |
| | میان تشکیل جلسه می دهد | 15 |
| * | تشکیل جلسه | 16 |
| * | تشکیل جلسه می دهد | 17 |
| | جلسه می دهد | 18 |

## 6. THE EXPERIMENT

An experiment using our monolingual corpus of Persian as well as the English-Persian parallel corpus described in section 3 has been done to demonstrate the effectiveness of the heuristic method on extracting all possible chunks in Persian language and finding the most appropriate equivalents for them in English. For this purpose, we carried out an experiment using a small fraction of the monolingual corpus and tried to extract a collection of one hundred Persian chunks and phrases to be passed through the subsequent stage of the experiment. In the second phase, all possible equivalents for every chunk extracted from the previous stage are generated from which the most appropriate one is selected as the correct equivalent for the given chunk. The results of this experiment are very encouraging and support our initial claim that the unstructured but complete information available in linguistic corpora can provide more precise and relevant responses in retrieval tasks compared to the structured but incomplete information from the existing monolingual and bilingual dictionaries.

## 7. CONCLUSION AND FURTHER DEVELOPMENTS

One of the consequences of the present project is to enhance the precision of the information retrieval systems in search machines using the databank accessible through two linguistic corpora, far richer the dictionaries.

Considering the fact that the researchers of this project had a rather easier accessibility to Persian corpora, they set the basis of the present methods and software, and algorithms on Persian language. However, as the computations and the algorithms of the present research do not depend on any specific language, there is such a possibility to apply this methodology to other languages. The only difference is the linguistic corpora used in this research which must be compiled for any language or languages involved in the study. The Robot algorithm for browsing monolingual sites and extracting the indexed sentences can also be used for similar research on other languages provided the referable addresses change according to the given language.

Considering the parameters of $x^2$ formula which are used in calculating the number of occurrence of the lexical items t and c in the corpora, we notice that the degree of computational complexity of the algorithm would be higher and higher as the corpora get bigger and bigger adding more records. As a result, the computation time gets longer. In order to optimize the algorithm of the $x^2$ formula, it is suggested that the corpora are divided into smaller components. After calculating the $x^2$ formula in each part and introducing the new formula, the association degree of the $x^2$ s are calculated and then the association between d and c is recalculated.

As almost all users expect the retrieval systems in search machines to be able to quickly and accurately respond their needs, it seems more economical to use a sorting algorithm in order to sort out all possible chunks in the monolingual corpus of the given language. Calculating the association score of all possible translations extracted from a bilingual parallel corpus, we would be able to create a bilingual dictionary of possible chunks for the pair of languages involved in the study.

Building a translation memory system is still another technology as language tool for translators which would be possible using the knowledge obtaining from the two existing corpora.

## 8. ACKNOWLEDGMENT

This study is partly supported by the Research Affairs of Payame Noor University.

## 9. REFERENCES

[1] H. Alizade, et al. , "Studying the efficiency of the existing methods in cross-language information retrieval using a machine-readable bilingual dictionary". Iranian Information and Documentation Centre, Vol. 25, No. 1, pp. 53-70, 2009.

[2] H. H. Chen, "Chinese information extraction Techniques". Presented at the SSIMIP, Singapore, 2002.

[3] D. Hull, and G. Grefenstette, "Querying Across Languages; A Dictionary –Based Approach to Multilingual Information Retrieval". In Proceedings of the 19th Annual International ACM Sigir, 49-57. Zurich, Switzerland, 1996.

[4] T. Mosavi Miangah, "Automatic term extraction for cross-language information retrieval using a bilingual parallel corpus". Proceedings of the 6th International Conference on Informatics and Systems (INFOS2008), PP. 81-84, Cairo, Egypt. 2008.

[5] T. Mosavi Miangah, "Constructing a large-scale English-Persian Parallel Corpus". META, 54 (1), pp. 181-188, 2009

[6] C. D. Manning, P. Raghavan, and H. Schütze, "An Introduction to Information Retrieval", Cambridge University Press, Cambridge, England, 2009.