Indian Journal of Science and Technology, *Vol 7(9), 1279–1288, September 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Combining Different Seed Dictionaries to Extract Lexicon from Comparable Corpus

Ebrahim Ansari[1,2], M. H. Sadreddini[1*], AlirezaTabebordbar[1] and Mehdi Sheikhalishahi[3]

[1]Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran; sadredin@shirazu.ac.ir, tabebordbar@tuv.ac.ir
[2]Department of Informatica, Universita di Pisa, Pisa, Italy; ansari@di.unipi.it
[3]Department of Electronics, Informatics and Systems, University of Calabria, Rende, Italy; alishahi@unical.it

Abstract

In recent years, many studies on extracting new bilingual lexicons from non-parallel (comparable) corpora have been proposed. Nearly all apply an existing small dictionary or other resource to make an initial list named seed dictionary. In this paper we discuss on using different types of dictionaries and their combinations as the initial starting list to produce a bilingual Persian-Italian lexicon from a comparable corpus. Our experiments applied state of the art techniques on four different seed dictionaries; an existing dictionary and three dictionaries created with pivot-based schema considering three different languages as pivot. We have used English, Arabic and French as pivot languages to extract these three pivot based dictionaries. An interesting challenge in our approach is proposing a method to combine different dictionaries together producing a better and more accurate lexicon. In order to combine seed dictionaries we proposed two novel combination models and examine the effect of them on comparable corpora which are collected from News Agencies. The experimental results exploited by our implementation show the efficiency of our proposed combinations.

Keywords: Bilingual Lexicon, Comparable Corpus, Pivot Language

1. Introduction and Related Works

In the last decade, some research has been proposed to acquire bilingual lexicons from non-parallel (comparable) corpora. A comparable corpus consists of sets of documents in several languages dealing with a given topic, or domain when documents have been composed independently of each other in different languages. Contrary to parallel corpus, comparable corpora are much easier to build from commonly available documents, such as news article pairs describing the same event in different languages. Therefore, there is growing interest in acquiring bilingual lexicons from comparable corpora. These methods are based on the assumption, which there is a correlation between co-occurrence patterns in different languages[1]. For example, if the word *teacher* and *school* co-occur more frequently than expected by chance in an English corpus then the German translations of *teacher* and *school*, *Lehrer* and *schule*, should also co-occur more often than expected in a corpus of German[1].

The starting point of their strategy is a list of bilingual expressions that are used to build the context vectors of all words in both languages. This starting list, or initial dictionary, is named the seed dictionary[2] and is usually provided by an external bilingual dictionary[3–6]. Some of recent methods use small parallel corpora to create their seed list[7] and some of them use no dictionary for starting phases[8]. Sometimes there are different types of dictionaries, each with its own accuracy. In this study, we use four different dictionaries and then their compositions as our seed dictionaries. The first dictionary is a small existing Persian-Italian dictionary. Other three dictionaries are extracted from a pivot-based method using English, French and Arabic as the pivot language individually.

*Author for correspondence

1.1 Using Pivot Languages to Create Bilingual Lexicon

Different approaches using a pivot language and consequently source-pivot and pivot-target dictionaries to build a new source-pivot lexicon have been proposed over the past twenty years[9–14]. One of the most known and highly cited methods is the approach of Tanaka and Umemura[11] where they only use dictionaries to translate into and from a pivot language in order to generate a new dictionary. These pivot language based methods rely on the idea that the lookup of a word in an uncommon language through a third intermediated language could be executed with machines. Tanaka and Umemura[11] use bidirectional source-pivot and pivot-target dictionaries (harmonized dictionaries). Correct translation pairs are selected by means of inverse consultation. This method relies on counting the number of pivot language definitions of the source word, which identifies the target language definition[11]. Sjobergh[10] presented another well-known method in this field. He generated his English pivoted Swedish-Japanese dictionary where each Japanese-to-English description is compared with all Swedish-to-English descriptions. The scoring metric is based on word overlaps, weighted with inverse document frequency and consequently the best matches are selected as translation pairs. The basis of most of other ideas and approaches proposed in recent years is based on those two described approaches[10,11]. Compared to other implementations, our approach needs some small and reliable extracted dictionaries as a part of our seed input. In our work, the method of Sjöbergh[10] is used because of its simplicity in implementation. Moreover as we needed only top translations with the highest scores and the generality of a selected method was not a factor.

1.2 Using Comparable Corpora

There is a growing interest in the number of approaches focused on extracting word translations from comparable corpora[3–8,15–19]. Most of these approaches share a standard strategy based on context similarity. All of them are based on an assumption that there is a correlation between co-occurrence patterns in different languages[1]. For example, if the words "معلم"|teacher| and "مدرسه" |school| co-occur more often than expected by chance in a corpus of Persian, then the Italian translations of them, "*insegnante*" |teacher| and "*scuola*" |school| should also co-occur in a corpus of Italian more than expected by chance. The general strategy extracting bilingual lexicon from the comparable corpus could be described as follows:

Word target t is a candidate translation of word source s if the words with which word t co-occur within a particular window in the target corpus are translations of the words with which word s co-occurs within the same window in the source corpus.

The goal is to find the target words having most similar distributions with a given source word. The starting point of this strategy is a list of bilingual expressions that are used to build the context vectors of all words in both languages. We named this the starting list the seed dictionary. The seed dictionary is usually provided by an external bilingual dictionary.

Otero and Campos[18] proposed a method using comparable corpora in order to validate the dictionary created from a pivot-based model. The method is based on two main tasks: First, a new set of bilingual correspondences is generated from two available bilingual dictionaries and second, the generated correspondences are validated by making use of a bilingual lexicon automatically extracted from non-parallel corpora. Irimia[16] uses comparable corpus to build an English-Romanian dictionary and uses the Rapp (1995)'s model as the core of her the implementation. Hazem and Morin[20] extracts bilingual lexicon from comparable corpora by using a statistical method, Independent Component Analysis (ICA). Bouamor et al.[21] present an extension of the classical approach using a Word Sense Disambiguation process. Their main focus is on resolving the word ambiguity problem revealed by the seed dictionaries used to transfer source context vectors to target language vector.

There are two approaches to create bilingual lexicon from comparable corpora: window based approach and syntax based approach. The difference is in the way the word contexts are defined. In Window-based methods, a fixed window size is chosen and it is determined how often a pair of words occurs within a text window. These windows are called the "fixed size window". Rapp[22] observed that word order of content words is often similar between languages, even between unrelated languages such as English and Chinese, and since this may be a useful statistical clue, we have modified the common approach in the way proposed by Rapp[22]. For a word *A*, several co-occurrence vectors is considered and calculated, one for each position within the window, instead of computing a single one.

Simple context frequency and additional weights such as inverse document frequency can be considered in both window and syntax based approaches. Well-known and widely used weighting for these approaches is log-likelihood[6]. In our implementation we use and consequently compare both simple context frequency and log-likelihood frequency individually. In computation of the log-likelihood ratio, the following formula from Dunning[23] and Rapp[6] is used:

$$loglike(A,B) = \sum_{i,j \in \{1,2\}} K_{ij} * \frac{\log(K_{ij} * N)}{C_i * R_j} \qquad Formula\ 1$$

Therefore:

$$\frac{loglike(A,B) = K_{11}\log(K_{11} * N)}{C_1 * R_1} + \frac{K_{12}\log(K_{12} * N)}{C_1 * R_2} + \frac{K_{21}\log(K_{21} * N)}{C_2 * R_1} + \frac{K_{22}\log(K_{22} * N)}{C_2 * R_2}$$

where,

$$C_1 = K_{11} + K_{12}, C_2 = K_{21} + K_{22}$$

$$R_1 = K_{11} + K_{21}, R_2 = K_{12} + K_{22}$$

$$N = C_1 + C_2 + R_1 + R_2$$

With parameters K_{ij} expressed in terms of corpus frequencies:

K_{11} = frequency of common occurrence of word *A* and word *B*
K_{12} = corpus frequency of word A - K_{11}
K_{21} = corpus frequency of word B - K_{11}
K_{22} = size of corpus (no. of tokens) - corpus frequency of word *A* - corpus frequency of word *B*

All numbers are normalized in our experiments.

For any word in source language, the most similar word in target language should be found. First, using seed dictionary all known words in the co-occurrence vector are translated to target language. Then, with considering the result vector, a similarity computation is performed to all vectors in the co-occurrence matrix of the target language. Finally, for each primary vector in the source language matrix, the similarity values are computed and the target words are ranked according to these values. It is expected that the best translation will be ranked first in the sorted list[6].

Different similarity scores have been used in the variants of the classical approach; Rapp[6] used *city-block* as their preferred similarity vector. The *cosine* similarity is used by Fung and McKeown[4], Chaiao and Zweigenbaum[3] and Saralegui et al.[19] and the *lin* similarity metric is used by Lin[24]. The other well-known similarity metrics are *dice* and *jaccard*[3,19]. In both *dice* and *jaccard* metrics, the association values of two lemmas with the same context are joined using their product. There are two different forms of *jaccard* and *dice*; the *jaccardMin* metric[25,26] and *diceMin*[7,27,28]. Only the smallest association weight is considered for both of these lemmas. Laroche and Langlais[29] presented some experiments for different parameters like context, association measure, similarity measure, and seed lexicon.

2. Our Approach

Our experiments to build a Persian-Italian lexicon are based on the comparable corpora window-based approach. In Section 2-1, we will describe our method to collect and create seed dictionaries and consequently, our implementation to use them independently is explained. Afterwards in Section2-2, we will describe the usage of comparable corpora to build a new Persian-Italian lexicon. An interesting challenge in our work is to combine different dictionaries with varying accuracies and use all of them as the seed dictionary for comparable corpora based lexicon generation. We address this problem using different strategies: First, combining dictionaries with some simple priority rules, and then, using all translations together without considering the differences in their accuracy. These combination strategies are discussed in Sections 2-3 and 2-4 respectively.

2.1 Building Seed Dictionaries

We have used four different dictionaries and their combinations as the seed dictionaries. The first dictionary is a small Persian-Italian dictionary, the three other dictionaries are created based on the pivot-based method presented by Sjobergh[10], which contain top entries with highest score. Like other standard methods, we just select the first translation among the all candidates. In next two sub sections, we describe the process of creating our dictionaries.

2.1.1 Existed Dictionary–DicEx

We used one small Persian-Italian dictionary as the existing dictionary named *DicEx*. For each entry, only first translation are selected and lemmatized. Although *DicEx*

is manually created dictionary and is our most accurate one, it has a small size in comparison with the rest.

2.1.2 Dictionaries Created by a Pivot based Method–DicPi-en, DicPi-fr and DicPi-ar

We used the method introduced by Sjobergh[10] as the baseline of the Pivot based dictionary creation. Translations with the highest scores are selected and results with lower score are taken out. We used three different languages English, French and Arabic as the pivot language. For each dictionary, a Persian-Pivot dictionary and a Pivot-Italian dictionary are selected as this step's inputs. So we needed six different input dictionaries; Persian-English, Persian-French, Persian-Arabic, English-Italian, French-Italian and Arabic-Italian dictionaries.

For all three pivot languages, English, French and Arabic, the following process is done individually:

All stop-words and all non-alphabet characters are removed from Pivot sides of these six dictionaries. Then the inverse document frequency is calculated for the remaining Pivot words as follow:

$$idf(w) = \log\left(\frac{|Pr|+|It|}{Pr_w + It_w}\right)$$

where w is the word we calculate the weight for, $|Pr|$ is the total number of dictionary entries in the Persian-Pivot dictionary, $|It|$ the same for Pivot-Italian dictionary, Pr_w is the number of descriptions in the Persian-Pivot dictionary the word w occurs and It_w is this number for Pivot-Italian.

Afterwards, all the Pivot language descriptions in the first dictionary must be matched to all descriptions in the second. Matches are scored by word overlaps that are weighed by predefined inverse document frequencies. In the counting phase, a word is only counted once for more than one occurrence in a same description. Based on Sjöbergh's[10] method scores are calculated as follow:

$$score = \frac{2 * \sum_{w \in Pr \cap It} idf(w)}{\sum_{w \in Pr} idf(w) + \sum_{w \in It} idf(w)}$$

Where Pr is the text in the translation part of Persian-Pivot lexicon and it is the same for Pivot-Italian Dictionary. When all scores are calculated, candidates with the highest score will be selected to build our new Persian-Italian dictionary. Considering three pivot languages English, French and Arabic, We have three extracted dictionaries and in final step, we just selected top 40,000 translations from all translations and named them *DicPi-en*, *DicPi-fr* and *DicPi-ar* respectively.

2.2 Using Seed Dictionaries to Extract Lexicon from Comparable Corpora

Because of large differences between Persian and Italian terms in syntax and grammar, the window-based approach is used, instead of the syntax based. Therefore, the columns of the weighting matrix are words and not lemmas. Based on our proposed consumption, the seed dictionary could be an existent dictionary, an independent dictionary created automatically or a combination of them.

2.3 The Core System

In this section we present our window-based approach. There are two types of input: the seed dictionary, and the bilingual comparable corpus. Weighting vectors must be created based on corpora and lexicons. Before creation of matrices for both Persian and Italian languages, the stop words of corpora are deleted and lemmatized. Two co-occurrence matrix sets are created for the Persian and Italian corpora: one set for simple approach and another for ordered base approach. In the order-based method, matrices must save the placement of each word with the pivot word in addition to saving the frequency in one window. In order to calculate the similarity scores we transferred our matrices from the source language to target language. A possible translation is a row in transferred matrix corresponding with a row in target matrix. Therefore the value of similarity scores are calculated and sorted between any row in the transferred matrix and all the rows in target matrix. In our experiment we use *DiceMin* similarity as the preferred similarity score:

$$diceMin(X,Y) = \frac{2 * \sum_{i=1}^{n} \min(X_i, Y_i)}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}$$

To build a new lexicon, for each word (i.e. row) in the source vector, the best matches in the target vector could be considered as the translation. Therefore, for each entry, we select the word corresponding to target vectors where the similarity score is more than the rest.

2.4 Using Simple Combination

In this Section, the process of creating the bigger seed dictionary by using a simple combination rule is discussed. The reliability of the existed dictionary, *DicEx* is highest among others and the accuracy of *DicPi-en*, the dictionary created using English as the pivot is higher than the dictionary created using French language as the pivot, *DicPi-fr*. The Dictionary created using Arabic language has less accuracy in comparison by the others. Based on these observations, a priority order is defined to create the final seed dictionary:

DicEx>DicPi-en>DicPi-fr>DicPi-ar

Our simple combination rule is:

Suppose that Dic_i's priority is more than Dic_j's; if there is the word A in both of Dic_i and Dic_j the translation is selected from Dic_i, the dictionary with higher priority.

By applying the above priority rule, a new Persian-Italian dictionary with about 73K unique entries is created. We named this new created dictionary which using the simple combination rules, *DicCoSi*. Apparently, all the words in *DicEx* are included in *DicCoSi*. According to Table 1 which presents a small view of three existed or extracted dictionaries *DicEx*, *DicPi-en, DicPi-fr* and *DicPi-ar*, Table 2 shows the combined dictionary with using our simple priority rule. All words in both tables are selected from the real test case.

2.5 Using Independent Word Combination

In simple priority based combination described in Section 3.2.2, there is a point should be discussed. Consider two words when first one appears in all four dictionaries and the second one just appears in one dictionary. In our simple approach, there is not any difference between these words. Therefore, a new combination method is proposed to deal with this flaw. Our advanced combination method is based on the assumption that one similar word in two different dictionaries could be considered independently. For example if a word appears in both dictionaries Dic_1 and Dic_2, it may have two independent columns in our vector matrix (i.e. it has two different weights in the transferred vectors). Therefore, the new dictionary named *DicCoAdv* is created where its size is equal to the sum of our three dictionary's sizes. In this new dictionary if the word *X* occurs in two dictionaries, there are two different entries for it named x_i and x_j where *i* and *j* are the indicator of corresponding dictionaries. An example of creating this new seed dictionary is presented in Table 3. In this example the creation phase is based on four primary dictionaries were defined in Table 1.

An example presented in Figure 1-A shows the lemma vectors for Persian words with simple combination method and Figure 1-B shows them after creation of *DicCoAdv*. Both of them are created based on dictionaries defined in Table 1.

Table 1. An example of four dictionaries

Persian word	Italian in *DicEx*	Italian in *DicPaEn*	Italian in *DicPaFr*	Italian in *DicPaAr*
سلام [hi]	Ciao	ciao		
خداحافظ [bye]		Ciao	Arrivederci	
دلقک [joker]			Buffone	burlone
شیر [milk]	Latte [milk]	Leone [lion]		Leone [lion]
زیبا [beautiful]			piacevole	
سگ [dog]		cane	cane	cane
ایران [Iran]				Iran
نان [bread]	Pane	pane	Pane	pane

Table 2. Combined dictionary with using simple combination rule based on dictionaries introduced at the Table 1

Persian word	Italian in *DicCoSi*
سلام [hi]	Ciao [*DixEx*]
خداحافظ [bye]	Ciao [*DicPaEn*]
دلقک [joker]	Buffone [*DicPaFr*]
شیر [milk]	Latte [*DixEx*]
زیبا [beautiful]	Piacevole [*DicPaFr*]
سگ [dog]	Cane [*DicPaEn*]
ایران [Iran]	Iran [*DicPaAr*]
نان [bread]	Pane [*DixEx*]

3. Preparing the Inputs

As explained before, two primary inputs are needed to perform comparable corpora based lexicon generation: first, seed dictionary and second comparable corpus/corpora. The procedures to prepare these needed data have been described in sections 4.1 and 4.2. Another needed input in our experiments is test words as our

testing dataset. The evaluation of test study is performed by two Persons. The first evaluator was one of the authors, who is native Persian and fluent in Italian and second one was Persian native who teaches Italian language. If both of the evaluators agree in one translation term, it is accepted as a true translation and otherwise, the translation is considered false. We selected 400 Persian objective test words from *Nabid*[30] Persian-English dictionary. The frequencies of all the selected words in our comparable corpus were more than 100.

Table 3. Combined dictionary with independent words method

Persian word	Italian in *DicCoAdv*	Persian word	Italian in *DicCoAdv*
سلام [hi]	Ciao	زیبا [beautiful]	Piacevole
سلام [hi]	Ciao	سگ [dog]	Cane
خداحافظ [bye]	Ciao	سگ [dog]	Cane
خداحافظ [bye]	Arrivederci	سگ [dog]	Cane
دلقک [joker]	Burlone	ایران [Iran]	Iran
دلقک [joker]	Buffone	نان [bread]	Pane
شیر [milk]	Latte [milk]	نان [bread]	Pane
شیر [milk]	Leone [lion]	نان [bread]	Pane
شیر [milk]	Leone [lion]	نان [bread]	Pane

3.1 Seed Dictionaries

Four different seed dictionaries are used in our experiments. The first one was a small preexisting Persian-Italian dictionary named *DicEx*. The second, third and fourth dictionaries, *DicPi-en*, *DicPi-fr* and *DicPi-ar* are dictionaries extracted by the pivot-based approach. These dictionaries are created considering English, French and Arabic as the pivot language respectively. Therefore, three source-pivot and three pivot-target dictionaries are needed. The Persian-English dictionary we used contains about 100,000 Persian index terms, The Persian-French contains about 80,000 Persian index terms and Persian-Arabic dictionary contains 85,000 index terms. The English-Italian, French-Italian and Arabic-Italian dictionaries contain about 130,000 words, 100,000 words and 75,000 words respectively.

We checked 200 randomly translated words in *DicPi-en*, the dictionary created using English as the pivot language and 84% of them are translated with acceptable tag. This accuracy is near but slightly less than the best results in famous pivot based approaches described in Section 1.1. Table 4 shows some characteristics of three explained dictionaries.

3.2 Comparable Corpora

The comparable corpus used in our experiment is the international sport related news gathered from different

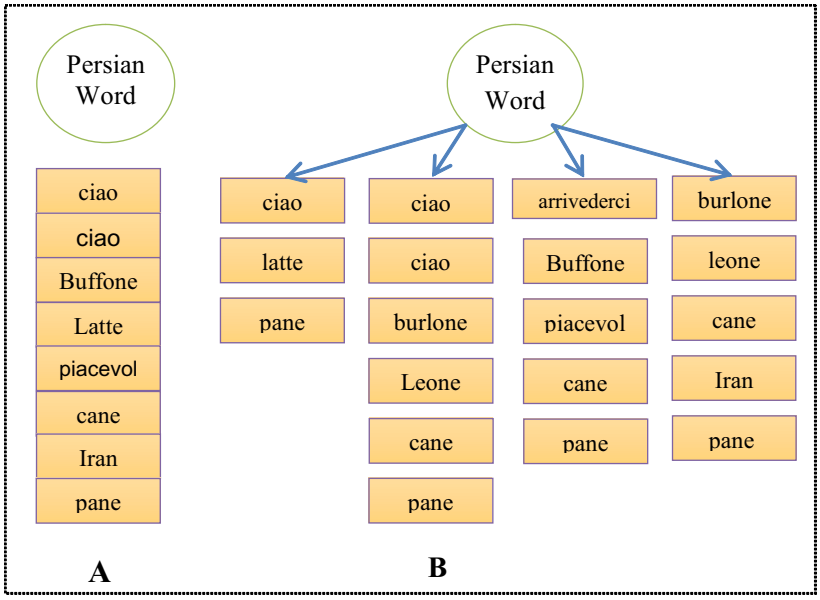

Figure 1. Combination vectors. Figure 1-A shows co-occurrence vector for a Persian lemma in simple combination and Figure 1-B uses independent words method for combination.

Persian and Italian news agencies. We used the *ISNA*[31] and the *FARS*[32] for the Persian part, the news agency *CORRIERE DELLA SERA*[33] and the *Gazzettadello Sport*[34] for Italian part. The numbers of selected articles are about 12K and about 15K from Persian and Italian resources respectively. While international sport news is very similar in different agencies, the comparability degree is not too small.

4. Experimental Results

In our experiments and for each test, two different result set are calculated. The Top-1 measure is the number of times when the test word's acceptable translation is ranked first, divided by the number of test words. The Top-10 measure is equal to the number of times a correct translation for a word appears in the top 10 translations in the result lexicon, divided by the number of test words.

As discussed in Section 3.2, In order to see the effect of using order-based windows, we studied both simple window and considering word order windows separately. The results show that taking ordering into account is not very effective to extract Persian-Italian lexicons and just in some cases, it has a slightly positive effect. In our approach all window size set to five and we have calculated both simple frequency and log-likelihood ratio. Despite our expectation, in a few cases using simple co-occurrence has a better efficiency with comparison of using log-likelihood ratio. While this difference is very small, at most demonstrated figures in this paper, simple frequency ratio is not considered and only log-likelihood ratio is shown. All experiments in this paper applied on gathered comparable corpora introduced in 4-2. Finally, different experiments are executed in order to evaluate and compare our combination models. In the first subsection, we use the four prior mentioned dictionaries as the seed lexicon individually. Then our two different proposed combination strategies are studied.

Table 4. The used corpora in our experiments

Dictionary name	Entries	Mutual words with *DicEx*
DicEx	13309	*NA*
DicPi-en	40000	6954
DicPi-fr	40000	5935
DicPi-ar	40000	5430

4.1 Using Independent Dictionaries

In first phase of our experiments, all four prior mentioned dictionaries are used as the seed lexicon individually. These dictionaries are the existed dictionary (*DicEx*) and three pivot base extracted dictionaries. First one considers English as the pivot (*DicPi-en*), second one uses French as the pivot language (*DicPi-fr*) and in the latest one Arabic is considered as the pivot language (*DicPi-ar*).

Figure 2 summarize the evaluation results considering these four seed dictionaries with and without using words order issue. The goal of this experiment is to see the effect of some general issues about our primary dictionaries.

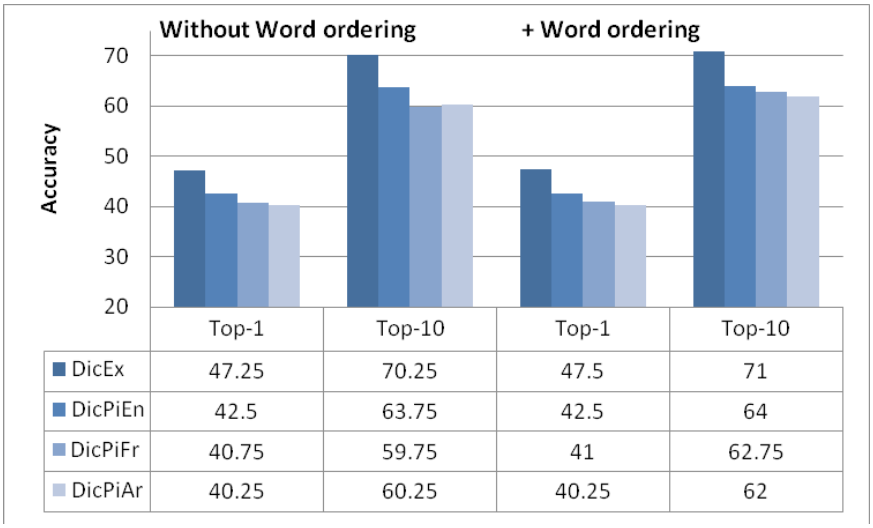

Figure 2. Results of using independent dictionaries with and without considering word orders. All results are based on log-likelihood measurement using our comparable corpus.

According to results and our expectation, the *DicEx*has better outcome despite its small size in comparison with others. The reason is higher accuracy of *DicEx* where it is a handmade dictionary and we can consider it 100%. The experimental results show that *DicPi-en* has a slightly better efficiency in comparison with two other created dictionaries. Based on retrieved statistics in section 4-1 (Table 4), *DicPi-en* has more mutual words with existed dictionary in comparison with *DicPaFr* and *DicPaAr's* mutual words with *DicEx* and this could be used to predict the accuracy order.

In Figure 3, the effect of using log-likelihood in comparison with using the simple frequency vectors is shown. For each experiment, we used two different schemas: with considering and without considering word orders. Based on our data sets and our results, and with considering the noise effects, this hypothesis could be supported that none of these schemas has a better efficiency in comparison with other.

4.2 Using Composite Dictionaries

In this section, we evaluate our ideas to combine different dictionaries together. As described before, two different types of combination are used in our experiments. The simple combination creates a dictionary with using a simple priority rule and advanced combination combines all dictionaries with considering all translations of any word. Table 5 shows the results of these studies. According to this table, the best results for Top-1 measure belong to simple combination model when all dictionaries are combined together and the best Top-10 results belongs to advanced combination model using all dictionaries together. In advanced combination, all the words in all dictionaries are selected in lexicon generation phase, and this generality could give us the better top-10 results.

Finally, Figure 4 shows a brief illustration to see the effect of our combination methods in comparison with classic approaches when they used just the existing dictionary, *DicEx* (the most accurate independent dictionary in our study) as the seed dictionary. In all results, log-likelihood ratio with considering word ordering issue are used to extract bilingual lexicons from our comparable corpus. In legends of this Figure, AC means advanced combination model.

Table 5. The effect of different dictionary combinations sung different methods

Dictionary name	Top-1		Top-10	
	Simple	Advanced	Simple	Advanced
DicEx + DicPi-en	50.00	49.50	75.00	75.50
DicEx + DicPi-fr	49.25	48.25	74.00	74.75
DicEx + DicPi-ar	48.75	48.00	73.75	74.50
All Pivot based*	44.50	44.75	71.75	72.50
All Dictionaries	***50.50***	50.00	75.50	***76.75***

* *DicPi-en + DicPi-fr + DicPi-ar*

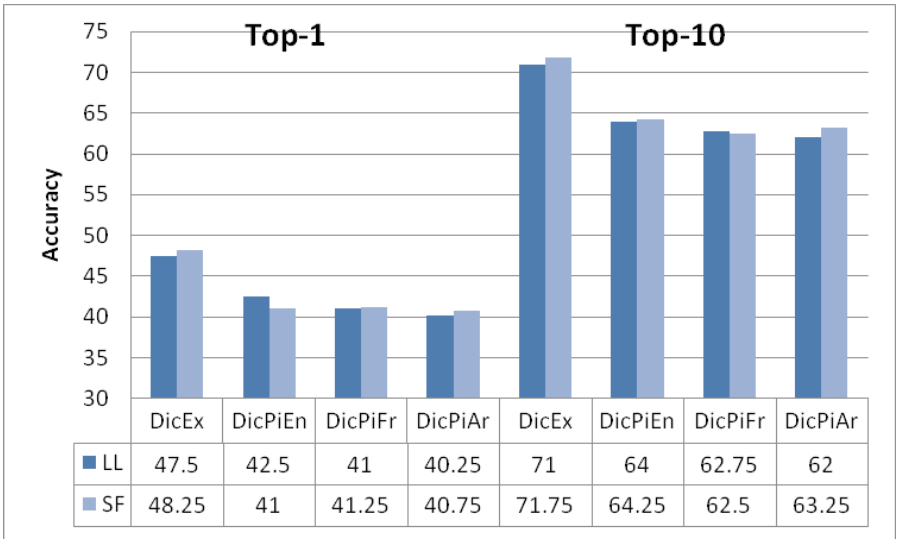

Figure 3. The effect of log-likelihood

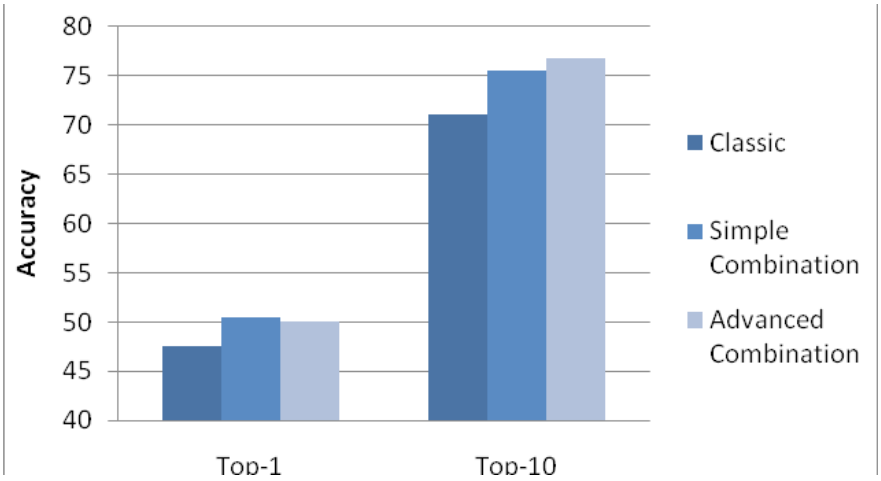

Figure 4. The effect of different introduced combinations.

5. Conclusion and Future Works

In the last decade, some methods have been proposed to extract bilingual lexicons from comparable corpora. To create a Persian-Italian lexicon, we decided to implement a comparable corpora-based lexicon generation method. This type of methods usually needs a small dictionary as their starting seed dictionary. In our study, four different seed lexicons (and their combination) are used, one pre-existing dictionary and three extracted dictionaries. The extractions of these three dictionaries are pivot based with considering three different languages English, French and Arabic as the pivot. In first part of our study, the effects of using these dictionaries on our comparable corpora are evaluated.

A new and interesting challenge introduced in our work was combining different dictionaries to create the seed dictionary. We used two different strategies: First, composing dictionaries with some priority rules; second, using all dictionaries together with considering similar words is two dictionaries as the different words in result dictionary. Both of these strategies were studied and based on our experimental results these novel dictionary combinations could improve the accuracy extracted lexicon.

6. Acknowledgement

The authors gratefully acknowledge the contribution and helps of Daniele Sartiano, VahidPooya, Amir Onsori and Dr. M. N. Makhfif in the completion of this work.

7. References

1. Rapp R. Identifying word translations in non-parallel texts, Proceedings of the 33rd annual meeting on Association for Computational Linguistics; 1995 Jun 26–30; Cambridge, Massachusetts. 981709. Association for Computational Linguistics; 1995. p. 320–2.
2. Fung P. Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese corpus. Proceedings of the Third Annual Workshop on Very Large Corpora; 1995 Jun; Boston, Massachusettes. p. 173–83.
3. Chiao Y-C, Zweigenbaum P. Looking for candidate translational equivalents in specialized, comparable corpora, Proceedings of the 19th international conference on Computational linguistics; Taipei, Taiwan. 1071904. Association for Computational Linguistics; 2002; 2:1–5.
4. Fung P, McKeown K. Finding Terminology Translations from Non-parallel Corpora. Proceedings of the Fifth Workshop on Very Large Corpora; 1997 Aug 18; Hong Kong. p. 192–202.
5. Fung P, Yee LY. An IR approach for translating new words from nonparallel, comparable texts. Proceedings of the 17th International Conference on Computational Linguistics—Volume 1; 1998 Aug 10–16; Montreal, Quebec, Canada. 980916. Association for Computational Linguistics; 1998. p. 414–20.
6. Rapp R. Automatic identification of word translations from unrelated English and German corpora, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics; 1999 Jun 20–26; College Park, Maryland. 1034756. Association for Computational Linguistics; 1999. p. 519–26.
7. Otero PG. Learning bilingual lexicons from comparable English and Spanish corpora, Proc of the Machine Translation Summit (MTS 2007); Copenhagen, Denmark. 2007. p. 191–8.

8. Rapp R, Zock M. Utilizing citations of Foreign words in Corpus-based Dictionary Generation. Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010); 2010 Aug; Beijing. p. 50–9.
9. István V, Shoichi Y. Bilingual dictionary generation for low-resourced language pairs, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing; 2009 Aug 06–07; Singapore. 1699625. Association for Computational Linguistics; 2009. p. 862–70.
10. Sjobergh J. Creating a free digital Japanese-Swedish lexicon, Proceedings of PACLING 2005. p. 296–300.
11. Tanaka K, Umemura K. Construction of a bilingual dictionary intermediated by a third language, Proceedings of the 15th Conference on Computational Linguistics— Volume 1; Kyoto, Japan. 991937. Association for Computational Linguistics; 1994. p. 297–303.
12. Tsunakawa T, Okazaki N, TsujiiJi. Building bilingual lexicons using lexical translation probabilities via pivot languages, Proceedings of the 6th International Conference on Language Resources and Evaluation; 2008 May 28–30; Mansour Eddahbi. p. 1664–7.
13. Tsunakawa T, Yamamoto Y, Kaji H. Improving calculation of contextual similarity for constructing a bilingual dictionary via a third language, International Joint Conference on Natural Language Processing; 2013 Oct 14–18; Nagoya, Japan. p. 1056–61.
14. Saralegi X, Manterola I, Vicente IS. Building a Basque-Chinese dictionary by using English as pivot. LREC 2012, Eighth International Conference on Language Resources and Evaluation; 2012 May 21–27; Istanbul, Turkey. p. 1443–7.
15. Dejean H, Gaussier E, Sadat F. Bilingual terminology extraction: an approach based on a multi-lingual thesaurus applicable to comparable corpora, COLING 2002; 2002 Aug 24–30; Tapei, Taiwan.
16. Irimia E. Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for: English-Romanian language pair, The 5th Workshop on Building and using Comparable Corpora: Language Resources for Machine Translation in Less-Resourced Languages and Domains, LREC 2012 Workshop; 2012 May 26; Istanbul, Turkey. p. 49–55.
17. Kaji H. Extracting Translation Equivalents from Bilingual Comparable Corpora. IEICE-Trans Inf Syst. 2005; E88-D(2):313–23.
18. Otero PG, Campos JRP. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora, Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing; 2010 Mar 21–7; Iaşi, Romania. 2175399. Springer-Verlag; 2010. p. 473–83.
19. Saralegui XISV, Gurrutxaga A. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. LREC 2008 Workshop on Building and using Comparable Corpora; 2008 May 31.
20. Hazem A, Morin E. ICA for Bilingual Lexicon Extraction from Comparable Corpora, BUCC 2012: the 5th Workshop on Building and using Comparable Corpora with special topic Language Resources for Machine Translation in Less-Resourced Languages and Domains co-located with LREC 2012; 2012 May 26; Istanbul, Turkey. p. 126–33.
21. Bouamor D, Semmar N, Zweigenbaum P. Building specialized bilingual lexicons using word sense disambiguation, International Joint Conference on Natural Language Processing; 2013 Oct 14–18; Nagoya, Japan. p. 952–6.
22. Rapp R. Die Berechnung von Assoziation ein korpuslinguistischer Ansatz. Hildesheim Zürich New York Olms; 1996.
23. Dunning T. Accurate methods for the statistics of surprise and coincidence. Comput Linguist. 1993; 19(1):61–4.
24. Lin D. Automatic retrieval and clustering of similar words. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2; 1998 Aug 10–14; Montreal Quebec Canada. 980696. Association for Computational Linguistics; 1998. p. 768–74.
25. Grefenstette G. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers: 1994.
26. Kaji H, Aizono T. Extracting word correspondences from bilingual corpora based on word co-occurrences information, Proceedings of the 16th conference on Computational linguistics - Volume 1; 1996; Copenhagen, Denmark. 992636. Association for Computational Linguistics; 1996. p. 23–8.
27. Curran JR, Moens M. Improvements in automatic thesaurus extraction, Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9; Philadelphia, Pennsylvania. 1118635. Association for Computational Linguistics; 2002. p. 59–66.
28. PlasLvd, Bouma G. Syntactic contexts for finding semantically similar words, The 16th Meeting of Computational Linguistics in the Netherlands (CLIN)2005, 2005 Dec 16; Amsterdam. p. 173–86.
29. Laroche A, Langlais P. Revisiting context-based projection methods for term-translation spotting in comparable corpora, Proceedings of the 23rd International Conference on Computational Linguistics; 2010 Aug 23–7; Beijing, China. 1873851. Association for Computational Linguistics; 2010. p. 617–25.
30. Kaabi H. Nabid Dictionary; 2002.
31. ISNA, Iranian students News Agency, International News part, Persian. Available from: http://isna.ir/fa/service/World
32. Fars News Agency, International News part, Persian. Available from: http://www.farsnews.com/newsv.php?srv=6
33. CORRIERE DELLA SERA, International news. Available from: Italian, http://www.repubblica.it
34. La Gazzetta dello Sport. Available from: Italian, http://www.gazzetta.it