Extracting Bilingual Persian Italian Lexicon from Comparable Corpora Using Different Seed Dictionaries

Ebrahim Ansari,[†‡] **M.H. Sadreddini,**[§] **Mahsa Radinmehr,**[†] **and Ziba Khosravan**[†]
[†] Department of Computer Science and Information Technology,
Institute for Advanced Studies in Basic Sciences
[‡] Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University
[§] Computer Science and Engineering Department, Shiraz University
{ansari,radinmehr,zibakh}@iasbs.ac.ir
sadredin@shirazu.ac.ir

Abstract

Bilingual dictionaries are very important in various fields of natural language processing. In recent years, research on extracting new bilingual lexicons from non-parallel (comparable) corpora have been proposed. Almost all use a small existing dictionary or other resources to make an initial list called the "seed dictionary". In this paper, we discuss the use of different types of dictionaries as the initial starting list for creating a bilingual Persian-Italian lexicon from a comparable corpus. Our experiments apply state-of-the-art techniques on three different seed dictionaries; an existing dictionary, a dictionary created with pivot-based schema, and a dictionary extracted from a small Persian-Italian parallel text. The interesting challenge of our approach is to find a way to combine different dictionaries together in order to produce a better and more accurate lexicon. We propose two different novel combination models and examine the effect of them on various comparable corpora that have differing degrees of comparability. We conclude our work with a new weighting schema to improve the extracted lexicon. The experimental results show the efficiency of our proposed models.

1 Introduction

Bilingual lexicons are a key resource in a multilingual society. The availability of translation resources varies depending on the languages pairs. Therefore, bilingual dictionaries for languages with fewer native speakers are scarce or even non-existent. Though automatic lexicon creation methods often have drawbacks such as including noise in the form of erroneous translations of some words, they are still popular because the alternative – manually constructing a dictionary – is very time-consuming. Automatic methods are often used to generate a first noisy dictionary that can be cleaned up and extended by manual work (Sjbergh, 2005).

A pivot language (bridge language) is useful for creating bilingual resources such as bilingual dictionaries. The Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries that share a common language. For example, using the Persian-English and the English-Italian dictionaries to build a new Persian-Italian lexicon. In recent years, some approaches based on this idea have been proposed (Tanaka and Umemura, 1994; Sjbergh, 2005; Istvn and Shoichi, 2009; Tsunakawa et al., 2008, 2013; Ahn and Frampton, 2006). In the last decade, some research has been proposed to acquire bilingual lexicons from non-parallel (comparable) corpora. These methods are based on this assumption: there is a correlation between co-occurrence patterns in different languages (Rapp, 1995). For example, if the words teacher and school co-occur more often 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 German corpus (Rapp, 1995). Most of the approaches share a standard strategy based on context similarity. The basis of these methods is finding the target words that have the most similar distributions with a given

Proceedings of the 12th Workshop on Building and Using Comparable Corpora at RANLP 2019, Varna, Bulgaria.
Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp (eds.)

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. This starting list, or initial dictionary, is named the seed dictionary (Fung, 1995) and is usually provided by an external bilingual dictionary (Rapp, 1999; Chiao and Zweigenbaum, 2002; Fung and McKeown, 1997; Fung and Yee, 1998). Some of the recent methods use small parallel corpora to create their seed list (Otero, 2007) and some other use no dictionary for starting phases (Rapp and Zock, 2010). Sometimes there are different types of dictionaries, with each having its own accuracy. (Ansari et al., 2014) propose two simple methods to combine four different dictionaries (one existing dictionary and three dictionaries extracted using pivot based method) to increase the accuracy of the output. They use three languages English, Arabic, and French to create their pivot based lexicons. In this work, we use three different types of dictionaries and then combine them to create our seed dictionaries. The first dictionary is a small existing Persian-Italian dictionary. The second dictionary is extracted from a pivot-based method. The third dictionary is created from our small parallel Persian-Italian corpus. Using these dictionaries, we propose different combination strategies and a new weighting method to use on these different dictionaries.

2 Related works

In this Section, we discuss approaches and implementations in three parts and show their relation to our work.

2.1 Using Pivot languages

Over the past thirty years different approaches have been proposed to build a new source-pivot lexicon using a pivot language and consequently source-pivot and pivot-target dictionaries (Tanaka and Umemura, 1994; Istvn and Shoichi, 2009; Tsunakawa et al., 2008, 2013; Ahn and Frampton, 2006). One of the most known and highly cited methods is the approach of Tanaka and Umemura (Tanaka and Umemura, 1994) 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 done with machines. Tanaka and Umemura 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 (Tanaka and Umemura, 1994). Sjobergh presented another well-known method in this field (Sjbergh, 2005). 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.

2.2 Using Parallel Corpora

Another way to create a bilingual dictionary is to use parallel corpora. Using parallel corpora to find a word translation (i.e. word alignment) started with primitive methods of (Brown et al., 1990) and continued with some other word alignment approaches such as (Gale and Church, 1991, 1993; Melamed, 1997; Ahrenberg et al., 1998; Tiedemann, 1998; Och et al., 1999). These approaches share a basic strategy of first having two parallel texts aligned in pair segments and second having word co-occurrences calculated based on that alignment. This approach usually reaches high score values of 90% precision with 90% recall, (Otero, 2007). Many studies show that for well-formed parallel corpora high accuracy rates of up to 99% can be achieved for both sentence and word alignment. Currently, almost the entire task of bilingual dictionary creation and especially the creation of a probability table for any word pairs could be done with well-known statistical machine translation software, GIZA++ (Och and Ney, 2003). Using Parallel corpora as the input of the dictionary creation process is attractive in two ways. First, alignment between sentences and words is very accurate as a natural characteristic of parallel corpora and these methods do not need any other external knowledge to build a bilingual lexicon. Second, no external bilingual dictionary (seed dictionary) is required. The main problem of creating a parallel corpus lexicon is the lack of extensive language pairs, therefore reliance on just using parallel corpora to build accurate bilingual dictionaries is impossible.

2.3 Using Comparable Corpora

There is a growing interest in the number of approaches focused on extracting word translations from comparable corpora (Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Djean et al., 2002; Kaji, 2005; Otero, 2007; Otero and Campos, 2010; Rapp and Zock, 2010; Bouamor et al., 2013; Irimia, 2012; E. Morin and Prochasson, 2013; Emmanuel and Hazem, 2014). 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 (Rapp, 1995). For example, if the words teacher and school co-occur more often than expected by chance in a corpus of English, 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 the 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. This starting list is called the seed dictionary. The seed dictionary is usually provided by an external bilingual dictionary. (Djean et al., 2002) uses one multilingual thesaurus as the starting list instead of using a bilingual dictionary. In (Otero, 2007) the starting list is provided by bilingual correlations previously extracted from a parallel corpus. In (Rapp et al., 2012), the authors extract a bilingual lexicon without using an existing starting list. Although they use no seed dictionary, their results are acceptable. Another interesting issue considered in recent years evaluating the effect of the degree of comparability on the accuracy of extracted resources (Li and Gaussier, 2010; Sharoff, 2013)

As described before, it is assumed that there is a small bilingual dictionary available at the beginning. Most methods use an existing dictionary (Rapp, 1999; Chiao and Zweigenbaum, 2002; Fung and McKeown, 1997; Fung and Yee, 1998) or build one with some small parallel resources (Otero, 2007). Entries in the dictionary are used as an initial list of seed words. Texts in both source and target languages are lemmatized and part-of-speech (POS) tagged with function words are removed. A fixed window size is chosen and it is determined how often a pair of words occurs within that text window. These windows are called the "fixed-size window" and word order does not take into account within a window. R. Rapp observed that word order of content words is often similar between languages, even between unrelated languages such as English and Chinese (Rapp, 1996). In approaches considering word order, for each lemma, there is a context vector whose dimensions are the same as the starting dictionary but in different window positions with regard to that lemma. For instance, if the window size is 2, the first context vector of lemma A, where each entry belongs to a unique seed word, shows the number of co-occurrences two positions to the left of A for that seed word. Three other vectors should also be computed, counting co-occurrences between A and the seed words appearing one position to the left of A and the same for two right hand positions following lemma A. Finally, all four vectors of length n are combined (where n is the size of the seed lexicon) into a single vector of length $4n$. This method takes into consideration the word orders to define contexts. In this paper, the efficiency of considering the word order schema is evaluated. Moreover, In the computation of the log-likelihood ratio, the simplified formula from Dunning and Rapp (Dunning, 1993) is used:

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

Therefore:

$$loglike(A,B) = K_{11} \log \frac{K_{11} * N}{C_1 * R_1} + K_{12} \log \frac{K_{12} * N}{C_1 * R_2} + K_{21} \log \frac{K_{21} * N}{C_2 * R_1} + K_{22} \log \frac{K_{22} * N}{C_2 * R_2} \quad (2)$$

Where:

$$C_1 = K_{11} + K_{12} \quad (3)$$

$$C_2 = K_{21} + K_{22} \quad (4)$$

$$R_1 = K_{11} + K_{21} \quad (5)$$

$$R_2 = K_{12} + K_{22} \quad (6)$$

$$N = C_1 + C_2 + R_1 + R_2 \quad (7)$$

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*

For any word in a source language, the most similar word in a target language should be found. First, using a seed dictionary all known words in the co-occurrence vector are translated to the target language. Then, With consideration of the result vector, 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 (Rapp, 1999). Different similarity scores have been used in the variants of the classical approach (Rapp, 1999). In (Laroche and Langlais, 2010) the authors presented some experiments for different parameters like context, association measure, similarity measure, and seed lexicon. Some of the famous similarity metrics are included in the Appendix of this paper. We decided to use `diceMin` similarity score in our work which has been used previously in (Curran and Moens, 2002; Plas and Bouma, 2005; Otero, 2007). The `diceMin` score is the similarity of two vectors, X and Y, and is computed using the below similarity measure.

$$diceMin(X,Y) = \frac{2 \cdot \sum_{i=1}^{n} min(X_i, Y_i)}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i} \quad (8)$$

3 Our Approach

Our experiments to build a Persian-Italian lexicon are based on the comparable corpora window approach discussed in Section 2.3. 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 with and without considering the differences in their weights.

3.1 Building Seed Dictionaries

We have used three different dictionaries and their combinations as the seed dictionaries. The first dictionary is a small Persian-Italian dictionary named `DicEx`. For each entry, only the first translation is selected to create lemmas. While `DicEx` is a manually created dictionary, it is the most accurate dictionary in our experiments, and its size is the smallest in comparison with the other dictionaries. The second dictionary is created based on the pivot-based method presented in (Sjbergh, 2005), which contains top entries with the highest score. In contrast to the Sjobergh's implementation where the main focus is creating a dictionary with very large coverage, our goal is creating a small dictionary with more accuracy for use as a seed dictionary in the main system. Therefore, we select the top 40,000 translations from all translations and named it `DicPi`. Finally, the third dictionary is built using two little parallel Persian-Italian corpora which is named `DicPa`. When there is more than one translation for an entry in the primary dictionary, we should select one translation. Most standard approaches select the first translation in the existing dictionary or the candidate with the highest score in the extracted (created) dictionary. However, in (Irimia, 2012), several definitions for one word based on their scores could be selected in the seed dictionary generation step. Like other standard methods, we selected the first translation among all the candidates.

3.2 Using seed dictionaries to extract lexicon from Comparable Corpora

Mathematics and theoretical points of our approach were discussed in Section 2.3. Given that there are large differences between Persian and Italian words in syntax and grammar, the window-based approach is preferred. The baseline of the method implemented in our study is an adaptation of (Rapp, 1999). Based on our proposed idea, the seed dictionary could be an existing dictionary, an automatically created dictionary, or a combination of them. Previous approaches show the need for

replacing the co-occurrence frequency in the matrix by measures that are able to eliminate word-frequency effects and consequently to favor significant word pairs. Therefore we use the log-likelihood ratio (i.e. Formula 1 (Dunning, 1993)) in our approach described in Section 2.3. To see its effect, we also carried out our tests without this metric by using the simple frequency matrix. In this experiment, we use `diceMin` similarity score as the preferred score. In Section 3.5 of this paper, a new similarity score, `newdiceMin` is proposed by the authors to weight dictionaries when different seed dictionaries are combined together.

3.3 Using simple combination

In this section, the process of creating the bigger seed dictionary by using a simple combination rule is discussed. `DicEx` has the highest accuracy and the accuracy of `DicPi` is higher than the dictionary created from the parallel corpus (i.e. `DicPa`). Based on the accuracy of dictionaries, a priority order is defined to create the final seed dictionary:

$$\texttt{DicEx} > \texttt{DicPi} > \texttt{DicPa}$$

Our simple combination rule is:

> Suppose that the priority of Dic_i is more than the priority of Dic_j; if a word w is in both Dic_i and Dic_j, its translation is selected from Dic_i (i.e. the dictionary with higher priority)

By applying the above priority rule, a new Persian-Italian dictionary with more than 65,000 unique entries is created. We name this newly created dictionary `DicCoSi`. Apparently, all the words in `DicEx` are included in `DicCoSi`. The experimental results show an improvement in the extracted lexicon when this new dictionary `DicCoSi` is used as the main system's seed dictionary in comparison with using our three simple dictionaries individually.

3.4 Using independent word combination

In our simple priority-based combination which is described in Section 3.3, there is an important issue that should be discussed. Given two words, where the first one appears in all three dictionaries and the second one just appears in one dictionary. In our simple approach, there is no difference between these words. Therefore, a new advanced combination method is proposed. Our advanced combination method is based on the assumption that one word in two different dictionaries should be considered independently as two different words. 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 `DicCoIn` 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 indicators of corresponding dictionaries.

3.5 New weighting method

There is another problem in our proposed advanced combination. Even though some dictionaries are more accurate than others, there is no difference in dealing with initial seed dictionaries. In order to ease this problem, a new weighting model for similarity scores is introduced. This new metric relies on two following aspects:

(1) We could change the effect of each seed dictionary in order to consider the higher weight for the more accurate dictionary. All weights could be tuned manually.

(2) If a word appears in two dictionaries, then it is not necessary to count it twice as a double-count would produce an unfair skew. We could consider its weight a little bit more than a normal occurrence weight and then divide it between different dictionaries.

If there are k different dictionaries in our proposed independent word-based combination, to calculate the similarity scores between bilingual lemmas we could use the proposed equation:

$$newdiceMin(X,Y) = \frac{2 \cdot \sum_{j=1}^{k} \sum_{X_i \in Dic_j} min(X_i, Y_i) \cdot w_j}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i} \quad (9)$$

where n is the size of the new combined dictionary and w_j is the weight of dictionary j. In our experiments, the size of k is equal to three. The new weighting method is based on this assumption that the dictionary with higher accuracy should affect the extracted lexicon more. In our experiments, two different sets of w_j are studied and the results are evaluated in Section 5.1.

4 Preparing The Inputs

As stated prior, two primary inputs are needed to perform comparable corpora-based lexicon generation: seed dictionary and comparable corpus/corpora. Three different seed dictionaries are used in our experiments. Table 1 shows some characteristics of three dictionaries.

To evaluate the result, a test dataset is needed. The evaluation of the test is performed by two annotators. The first evaluator is one of the authors, who is a native Persian speaker and fluent in Italian and the second one is a Persian native who teaches the Italian language. If both of the evaluators agree on a translation word, it is accepted as a true translation, otherwise, the translation is considered false. We selected 400 Persian objective test words from Nabid Persian-English dictionary [1]. The frequencies of all the selected words in our corpora (general corpus and specific domain corpus) were greater than 100.

4.1 Seed Dictionaries

Dictionary Name	Entries	Mutual words
DicEx	13,309	N/A
DicPi	40,000	6,954
DicPa	40,000	4,220

Table 1: Number of entries and mutual words with DicEx of dictionaries used in our Experiments

In our experiments, three different types of comparable corpora are gathered: The first one is a small set of Wikipedia[2] articles in Persian and Italian. In order to skip those articles which are famous and well described in one of our languages (e.g. an article about an Italian village) we selected those article pairs where the difference between their sizes is not more than 50%. After applying this criterion, 6,500 articles are selected in both languages: about 150,000 sentences for Persian and 176,000 sentences in Italian. Both groups of sentences were tokenized and lemmatized. The resulting corpus is called `WikiCorpus` in our studies. This corpus is the most comparable corpus among our corpora (The comparability degree is more than the rest). The second corpus is the international sport-related news gathered from different Persian and Italian news agencies. We used the ISNA[3] and the FARS [4] for the Persian part, and the news agency CORRIERE DELLA SERA[5] and the Gazzetta dello Sport[6] for the Italian part. The numbers of selected articles are about 12,000 and about 15,000 from Persian and Italian resources, respectively. We named this corpus SportCorpus. We combined SportCorpus and `WikiCorpus` and used them together in our experimental results. We call this new combined corpus `SpeCorpus` (Specific domain-based corpus). The third corpus is based on international news gathered from different Persian and Italian news agencies. The difference between this corpus and `SpeCorpus` is that the former was gathered from sport-related news and this one is gathered from general subjects. This is our biggest corpus but obviously has a very low comparability degree in comparison with `SpeCorpus`. The number of articles in the Persian version was about 108,000 and for the Italian version was about 140,000 articles. We used ISNA and FARS news agencies for Persian version and CORRIERE DELLA SERA as the Italian resource. We named this corpus `GenCorpus`.

5 Experimental Results

All experiments described in this paper were applied on two types of comparable corpora: (1) the combination of `WikiCorpus` and `SportsCorpus` which we named `SpeCorpus`. (2) `GenCorpus` as a big, general, and less comparable corpus. The characteristics of these corpora were discussed in Section 4. In our experiments and for each test, two different result sets 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 resulting lexicon, divided by the number of test words.

In the first phase of our experiments, all three previously mentioned dictionaries are used individually as the seed lexicon. These are the preexisting dictionary (`DicEx`), the pivot base extracted dictionary (`DicPi`) and the parallel corpus-based dictionary (`DicPa`). Figures 1 summarizes the

[1] Nabid Dictionary, written by Hani Kaabi, Iran, 2002
[2] https://www.wikipedia.org/
[3] https://isna.ir
[4] https://www.farsnews.com
[5] https://www.repubblica.it/
[6] La Gazzetta dello Sport, Italian, http://www.gazzetta.it/

evaluation results using these three seed dictionaries with and without using word order on SpeCorpus, the corpus with higher comparability degree. Figure 2 shows that using corpus with higher comparability degree increases the accuracy in both Top-1 and Top-10 results significantly. As it is expected, this difference for Top-1 results is more than the Top-10 measure. According to the results, the DicEx has better outcome despite its small size compared with the other dictionaries. A reason is the high accuracy of DicEx as it is a handmade dictionary. We could consider it a 100% accurate dictionary. The experimental results show that DicPi has a slightly better efficiency in comparison with parallel corpora based dictionary DicPa. The authors conclude that the reason is the limitation of our parallel Persian-Italian corpus used to create the translation table.

In the second part of our experiments, we evaluated our ideas of combining different dictionaries together. Table 2 shows the results of this study. According to this table, the best results for Top-1 measure belong to the simple combination model when all dictionaries are combined together. The best Top-10 results belong to the advanced combination model combining all dictionaries. In advanced combination, all words in all dictionaries are selected in the lexicon generation phase, and this generally gives us better Top-10 results. An important issue for our advanced combination is that all translations in different dictionaries have the same weight and this may decrease the effect of DicEx. Although it is our most accurate dictionary, it is also the smallest one. This problem is tackled in the next section by using our weighting lemma.

5.1 Using new weighting

Two different heuristics are considered to adjust weights in our weighting schema. The first one is to tune weights based on dictionaries accuracy. The accuracies could be collected from Top-10 scores calculated in our experiments. In the first set, the weights for DicEx, DicPi and DicPa are 0.7, 0.64 and 0.59, respectively. In the second heuristic set, the weights are calculated based on both accuracy and the dictionary size. This weight set is constructed based on the assumption that the bigger dictionary should have a lower effect on the final result. We used the following formula to calculate the weights.

$$w_i = accuracy_i \cdot \frac{MaxSize}{size_i} \qquad (10)$$

Based on the second heuristic, and with considering the results in our study the weights are:

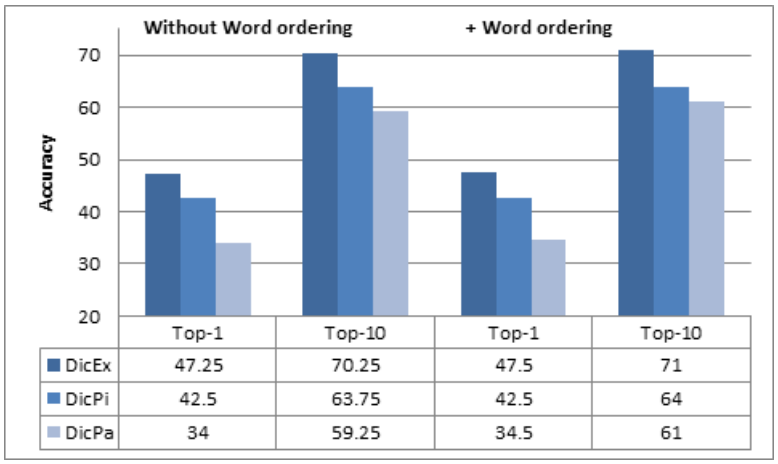
$$\begin{aligned} W_{\texttt{DicEx}} &= 2.10, \\ W_{\texttt{DicPi}} &= 0.64, \\ W_{\texttt{DicPa}} &= 0.59. \end{aligned}$$

The results of these experiments based on different weighting sets are shown in Table 3. W_i = 1 presents the classic approach without using the proposed weighting system.

Finally, Figure 3 shows a brief demonstration 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, the log-likelihood ratio with considering word ordering schema are used to extract bilingual lexicons from SpeCorpus, our corpus with high comparability degree. AC stands for advanced combination model.

6 Conclusion

In the last decade, some approaches have been proposed to extract bilingual lexicons from comparable corpora. In order to create a Persian-Italian lexicon, we decided to implement a comparable corpora-based lexicon generation method. In our study, three different seed lexicons (and combinations) are used consisting of one pre-existing dictionary and two extracted dictionaries. The first extracted dictionary is based on parallel-corpora dictionary creation methods and the second one is extracted by pivot language models. While for a seed dictionary a small dictionary is needed, we just selected the top translations from these created dictionaries. In the first part of our study, the effects of using these dictionaries on different types of comparable corpora are evaluated. A new and interesting challenge which is introduced in this paper was creating a new seed by combining some different dictionaries. We used two different strategies: First, composing dictionaries with some priority rules; second, using all dictionaries together considering similar words in two dictionaries as a different word. Both of these strategies were studied and based on our experimental



	Top-1	Top-10	Top-1	Top-10
DicEx	47.25	70.25	47.5	71
DicPi	42.5	63.75	42.5	64
DicPa	34	59.25	34.5	61

Figure 1: Results of using independent dictionaries with and without considering word orders. All results are based on log-likelihood measurement using SpeCorpus (in-domain corpus)

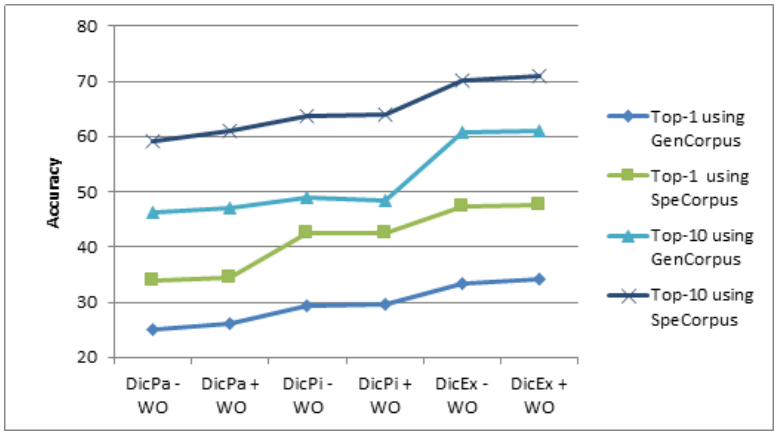


Figure 2: Effect of using different corpora in with different comparability degree

results these novel dictionary combinations could improve the efficiency of the results. Furthermore, the effect of comparability degree of the initial comparable corpus is studied using different types of comparable corpora. Finally, a new weighting method has been proposed to increase the efficiency of our dictionary combination. This new weighting method uses the assumption that the effect of a more accurate seed dictionary should have a better result in comparison with others.

Acknowledgments

The authors gratefully acknowledge the contribution and help of Dr. Fatemeh Alimardani, Dr. Daniele Sartiano, Vahid Pooya, Dr. Amir Onsori, S. M. H. Mirsadeghi, and Dr. Mahshid Nikravesh to this work. The research was partially supported by OP RDE project No. CZ.02.2.69/0.0/0.0/16_027/0008495, International Mobility of Researchers at Charles University.

References

Kisuh Ahn and Matthew Frampton. 2006. Automatic generation of translation dictionaries using intermediary languages. Association for Computational Linguistics, 1608848, pages 41–44.

Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. Association for Computational Linguistics, 980851, pages 29–35. https://doi.org/10.3115/980451.980851.

Ebrahim Ansari, M. H. Sadreddini, Alireza Tabebord-

Dictionary Name	Top-1		Top-10	
	Simple	Advanced	Simple	Advanced
DicEx + DicPi	50.00	49.50	75.00	75.50
DicEx + DicPa	48.75	48.00	74.00	74.75
DicPi + DicPa	42.50	43.00	66.75	67.50
All Dictionaries	**50.25**	49.75	75.25	**76.75**

Table 2: The effect of different dictionaries in combination with different methods on SPECORPUS for advanced combination

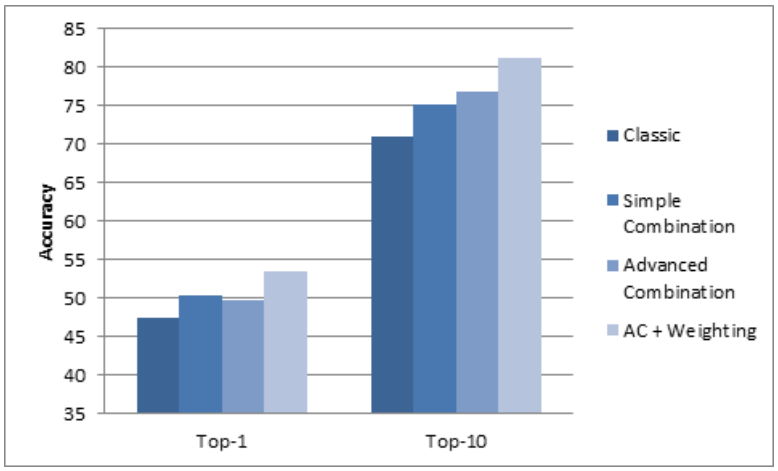


Figure 3: The effect of different introduced combinations

	Top-1	Top-10
W_i=1	50.25	76.75
Weight 1	52.50	78.25
Weight 2	**53.75**	**81.25**

Table 3: The effect of new weighting schema on accuracy of extracted dictionary (In all tests, the combination of three dictionaries is used and the comparable corpus is SPECORPUS)

bar, and Mehdi Sheikhalishahi. 2014. Combining different seed dictionaries to extract lexicon from comparable corpus. *Indian Journal of Science and Technology* 7(9):1279–1288.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Building specialized bilingual lexicons using word sense disambiguation. pages 952–956.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.* 16(2):79–85.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. Association for Computational Linguistics, 1071904, pages 1–5. https://doi.org/10.3115/1071884.1071904.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. Association for Computational Linguistics, 1118635, pages 59–66. https://doi.org/10.3115/1118627.1118635.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19(1):61–74.

Herv Djean, ric Gaussier, and Fatia Sadat. 2002. Bilingual terminology extraction: an approach based on a multi-lingual thesaurus applicable to comparable corpora.

B. Daille E. Morin and E. Prochasson. 2013. Bilingual terminology mining from language for special purposes comparable corpora. In *Building and Using Comparable Corpora*. Springer.

Morin Emmanuel and Amir Hazem. 2014. Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 1284–1293.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. pages 173–183.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations fromNon-parallel corpora. volume 1, pages 192–202.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. Association for Computational Linguistics, 980916, volume 1, pages 414–420. https://doi.org/10.3115/980451.980916.

William A. Gale and Kenneth W. Church. 1991. Identifying word correspondence in parallel texts. Association for Computational Linguistics, 112428, pages 152–157. https://doi.org/10.3115/112405.112428.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.* 19(1):75–102.

Elena Irimia. 2012. Experimenting with extracting lexical dictionaries from comparable corpora for: English-romanian language pair. pages 49–55.

Varga Istvn and Yokoyama Shoichi. 2009. Bilingual dictionary generation for low-resourced language pairs. Association for Computational Linguistics, 1699625, volume 2, pages 862–870.

Hiroyuki Kaji. 2005. Extracting translation equivalents from bilingual comparable corpora. *IEICE - Trans. Inf. Syst.* E88-D(2):313–323. https://doi.org/10.1093/ietisy/E88-D.2.313.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. Association for Computational Linguistics, 1873851, pages 617–625.

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. Association for Computational Linguistics, 1873854, pages 644–652.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. Association for Computational Linguistics, 979656, pages 305–312. https://doi.org/10.3115/979617.979656.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.* 29(1):19–51. https://doi.org/10.1162/089120103321337421.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. pages 20–28.

Pablo Gamallo Otero. 2007. Learning bilingual lexicons from comparable english and spanish corpora. pages 191–198.

Pablo Gamallo Otero and Jose Ramom Pichel Campos. 2010. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. Springer-Verlag, 2175399, pages 473–483. https://doi.org/10.1007/978-3-642-12116-6_40.

Lonneke van der Plas and Gosse Bouma. 2005. Syntactic contexts for finding semantically similar words. pages 173–186.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. Association for Computational Linguistics, 981709, pages 320–322. https://doi.org/10.3115/981658.981709.

Reinhard Rapp. 1996. Die berechnung von assoziationen: ein korpuslinguistischer ansatz. *Hildesheim; Zrich; New York: Olms* .

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. Association for Computational Linguistics, 1034756, pages 519–526. https://doi.org/10.3115/1034678.1034756.

Reinhard Rapp, Serge Sharoff, and Bogdan Babych. 2012. Identifying word translations from comparable documents without a seed lexicon. pages 460–466.

Reinhard Rapp and Michael Zock. 2010. Utilizing citations of foreign words in corpus-based dictionary generation .

Serge Sharoff. 2013. Measuring the distance between comparable corpora between languages. In *BUCC: Building and Using Comparable Corpora*. Springer.

Jonas Sjbergh. 2005. Creating a free digital japanese-swedish lexicon. pages 296–300.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. Association for Computational Linguistics, 991937, pages 297–303. https://doi.org/10.3115/991886.991937.

Jrg Tiedemann. 1998. Extraction of translation equivalents from parallel corpora.

Takashi Tsunakawa, Naoaki Okazaki, and Junichi Tsujii. 2008. Building bilingual lexicons using lexical translation probabilities via pivot languages. pages 1664–1667.

Takashi Tsunakawa, Yosuke Yamamoto, and Hiroyuki Kaji. 2013. Improving calculation of contextual similarity for constructing a bilingual dictionary via a third language. pages 1056–1061.