Collocation Extraction Using Parallel Corpus

Kavosh Asadi Atui[1] *Heshaam Faili*[1] *Kaveh Assadi Atuie*[2]
(1)NLP Laboratory, Electrical and Computer Engineering Dept., University of Tehran, Iran
(2)Electrical Engineering Dept., Sharif University of Technology, Iran
kavosh.asadi@ece.ut.ac.ir,hfaili@ut.ac.ir,kassadi@ee.sharif.ir

ABSTRACT

This paper presents a novel method to extract the collocations of the Persian language using a parallel corpus. The method is applicable having a parallel corpus between a target language and any other high-resource one. Without the need for an accurate parser for the target side, it aims to parse the sentences to capture long distance collocations and to generate more precise results. A training data built by bootstrapping is also used to rank the candidates with a log-linear model. The method improves the precision and recall of collocation extraction by 5 and 3 percent respectively in comparison with the window-based statistical method in terms of being a Persian multi-word expression.

KEYWORDS: Information and Content Extraction, Parallel Corpus, Under-resourced Languages

Proceedings of COLING 2012: Posters, pages 93–102,
COLING 2012, Mumbai, December 2012.

1 Introduction

Collocation is usually interpreted as the occurrence of two or more words within a short space in a text (Sinclair, 1987). This definition however is not precise, because it is not possible to define a short space. It also implies the strategy that all traditional models had. They were looking for co-occurrences rather than collocations (Seretan, 2011). Consider the following sentence and its Persian translation[1]:

"Lecturer issued a major and also impossible to solve problem."

مدرس یک مشکل بزرگ و غیر قابل حل را عنوان کرد.

"مدرس"/modarres/"lecturer"

"یک"/yek/"a"

"مشکل"/moshkel/"problem"

"بزرگ"/bozorg/"major"

"و"/va/"and"

"غیر قابل حل"/gheyreghabelehal/"impossible to solve"

"عنوان کرد"/onvankard/"issued"

This sentence emphasizes the action of "issuing a problem" which is a collocation, because it is a common way of saying that someone warned about a problem. Figure 1 shows that a verb-object relation between "issued" and "problem" and the alignments between the sentences imply that there is a corresponding relation between "مشکل" /moshkel/"problem" and "عنوان کرد"/onvan kard/"issued" in the Persian language.

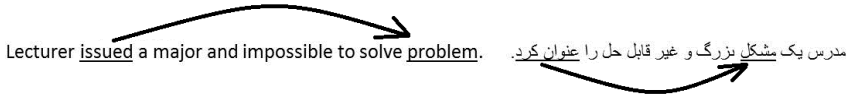

FIGURE1 – Example of a collocation: The relation between مشکل and عنوان کرد in the Persian sentence corresponds to the relation between issued and problem in English sentence.

Noticeably, window-based method cannot extract the collocation, because of the vagueness in the definition of short space. Moreover, the window cannot be expanded to include the words constructing the collocation. It is proved that expanding the window to more than 5 words is impractical (Dias, 2003). Besides, another flaw of the classical methods is that many false positive samples are mistakenly classified as collocation. This problem occurs especially in pairs having a small number of occurrences in the corpus (Seretan, 2011). While the latter problem can be solved (Basili et al.,1994), what this paper presents is another strategy which does not insist on classical approaches.

Recent improvements on the accuracy of parsers motivate modern approaches to analyze the sentences first (Seretan, 2006). Although that is the case in many languages like English, a lot of

[1] Persian uses Arabic script as its writing formalism which is written from right to left direction.

efforts have to be done in order to obtain an admissible accuracy in parsing the sentences of under-resourced languages like Persian.

This study accepts an alternative definition for collocation: "an expression consisting of two or more words that corresponds some conventional way of saying things" (Manning & Schütze, 1999). This definition does not have the vagueness the window-based method has.

Collocation has a deep influence on many other tasks of NLP such as MT systems (Orliac & Dillinger, 2003) and Text Summarization (Seretan, 2011) which makes it essential to find an alternative solution. From the next part of the paper, a process of extracting the collocations of Persian language will be presented.

The parallel corpus used in this study is Tehran English-Persian Parallel Corpus (Pilevar et al., 2011). The Corpus is comprised of more than 500000 pairs of sentences. The sentences are aligned by the IBM model3 using Giza++. In IBM model3 it is possible to have many to many alignments. This model is selected because it provides the extraction of collocations including more than two words.

In this method, by having a parallel aligned corpus and also parsed sentences of the source language, dependencies between the words of the target language are extracted initially. Direct Projection Algorithm (Hwa et al., 2005) is employed. It uses the alignments between the source and target sentences and the dependencies of the source language. In order to rank pairs of words by a log-linear classifier, a reasonable training data is then provided using bootstrapping with a small initial training set. Afterwards, the log-linear model is trained to sort and classify the candidates. Finally, the validation phase is implemented by the means of mutual dependency of two constituents to validate them based on their frequency of occurrence in another Persian corpus. Figure 2 demonstrates the architecture of this system.

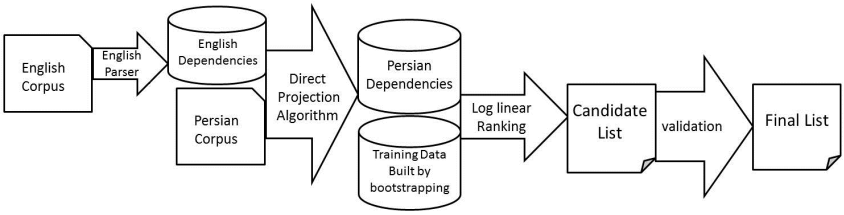

FIGURE 2 – Simplified architecture of system.

Briefly, the contribution of the paper is as follows:

- Employing the initial syntactical analysis without a parser for target language
- Using bootstrapping to build up a training data for log-linear classifier
- Developing the first dependency-based collocation extraction approach for Persian language.

In the next section, previous work related to this study is discussed. Section 3 consists of 4 separate parts and explains the method. Section 4 reports a comparative evaluation between our method and a classical window-based method as a baseline.

2 Related work

In the past decades, many studies regarding the collocation extraction have been undertaken. Classically, all approaches are consisted of two parts: Candidate Identification and Candidate Ranking. All earliest approaches devoted most of their efforts to find a suitable association measure (AM) in order to perform the ranking phase. One of the earliest measures is the z-score (Smadja, 1993) which assumes the data to be normally distributed. Log likelihood ratio (LLR) is another measure that is used in the recent efforts (Orliac & Dilinger, 2003). Still, the most common measure of collocation extraction is Pointwise Mutual Information (Church & Hanks, 1990). There is not an agreement on the best AM, but recent studies suggest that Mutual Information is the best possible measure (Pecina, 2010).

As mentioned above, the first phase of the architecture is identifying possible candidates. This phase is consisted of a linguistic preprocessing of the sentences (Seretan, 2011). The phase could be varied from lemmatization to deep parsing. Obviously, collocation deals with lemmas, not with word surface. Combining all inflected forms of a unique lemma leads to more competitive results (Evert, 2004). POS tagging is another preprocessing attempt to identify the potential candidates more precisely. There is a considerable improvement in the results of the system by performing POS tagging (Church & Hanks, 1990).

The common drawback of all these approaches is that they are not able to capture long distance dependencies. There is a solution to overcome this problem (Charest et al.,2007; Pecina,2010). Although less convenient to apply for under-resourced languages, deep parsing could be used to preprocess the text (Lu & Zhou, 2004).

Using monolingual corpora and word alignment is another recently common approach. In this approach, the monolingual corpus is replicated to generate a parallel corpus of the same sentences. Then, the aligned words are ranked, and pairs with higher scores are extracted as collocation (Liu et al., 2011). Another option is to use a multilingual parser to obtain more accurate results (Seretan & Wehrli, 2006). It is also unavailable in under-resourced languages.

The classical approaches do not lead to the competitive results, and recent approaches are based on accurate parsers. This paper introduces a novel method that not only eliminates the drawbacks of classical approaches, but also employs the syntactical analysis of the corpus without the need for a parser for the Persian (target side) or any other under-resourced language.

3 Methodology

This section introduces the novel method of extraction of the collocations of the Persian language. The method is divided into four steps: dependency projection, candidate generation, candidate ranking, and validation. Each step is explained in the following parts.

3.1 Dependency projection

Having a parsed English corpus, a list of relations between pairs of words is provided. In this method Dependency Parsing is used. It provides the relations between pairs of constituents. An algorithm is needed to transfer these relations to the target language.

Direct Projection Algorithm (Hwa et al., 2005) is employed in this step. This algorithm needs a list of the alignments between source and target words. Having a pair of sentences formed by an arbitrary number of words named e_1 through e_N in the source language and f_1 through f_M in the target language and also alignments between the words, five different scenarios are possible:

1. one to one: if two words of the source sentence named e_i and e_j are aligned to unique words f_i and f_j in the target language, relation (e_i,e_j) results in a new relation between f_i and f_j in the target language.
2. unaligned: if there is no corresponding word for e_i, a new null node is created. Relations including the unaligned word form relations having that null node in one part of the relation in the target language.
3. one to many: if more than one word i.e. f_x ,…, f_y are aligned to a unique word in the source language, a new node as the parent of these words is created and the alignment is modified to form a one to one alignment.
4. many to one: if e_i ,…, e_j are all aligned to a single word in the target language, all the alignments between them and the unique target word is eliminated except for the alignment containing the head of these words (which could be extracted from the set of the dependencies).
5. many to many: in this case, first one to many and then many to one scenario happen.

Importantly, in order to extract the collocations with more than two words, many to many alignments are necessary. The next step is generation of the candidates.

3.2 Candidate generation

In this step, a list of the potential collocations is generated. Dependency parser provides the relations between pairs of the constituents and their directions. Dependencies listed in Table 1 are primary candidates to construct collocation if they satisfy the following conditions:

1. not having a proper noun in one of its two parts.
2. not containing a null node created by Direct Projection.
3. not being an erroneous dependency e.g. dependency between a verb and an object without having any verbs at the both sides of the relation.
4. not including an auxiliary or modal verb.
5. not including uncommon constituents between the source and target languages. An example is a dependency having "را"/ra/ in the Persian language. This word indicates that there is an object right before it, while there is no such word in the English language.

Type	Example
Verb - Adverb	Sleep – Deeply
Verb - Object	Issue – Problem
Verb - Subject	Shine – Gold
Noun - Adjective	Game – Full
Adjective - Adverb	Fully – Optimistic

TABLE 1 – List of types of collocations accepted in this paper and their corresponding examples.

After identifying potential pairs, candidate ranking is performed. The next part describes the method of sorting the candidates.

3.3 Candidate ranking

In this phase the method of ranking the candidates is introduced. This ranking is based on a set of features and a log linear model. As mentioned earlier, sorting the pairs depends on some set of features. Importantly, the type of the dependency and two phrases are not the only information used to perform the ranking. It is crucial to include the results of Direct Projection Algorithm to better define discriminative features.
Following is the list of the features:

- Length of the target sentence
- Difference between the length of two sentences
- Total number of null nodes created by Direct Projection Algorithm in the sentence
- Type of the dependency
- Relation-specific features. An example is whether the verb imposes an object in a verb-object relation

Having these features, a training data is needed. This is provided by bootstrapping. This is obtained by having only a small initial training data. In each step of the algorithm best decisions made by the algorithm are selected and are added to the initial training data. This process results in a large training data which is necessary to train the log-linear model. The most important requirement of this phase is to have a measure to evaluate each decision made by the algorithm in all iterations.

It is now possible to build up a log linear model and estimate the weights for each one of the features form the training data derived from bootstrapping. Equation 1 denotes the possibility of each class.

$$p(c|x) = \frac{e^{\Sigma_{i=0}^{N}(w_i f_i)}}{\Sigma_{c=1}^{2}(e^{\Sigma_{i=0}^{N} w_i f_i})} \qquad (1)$$

Here, $p(c|x)$ denotes the probability of constructing a collocation for every pairs of words x or belonging to other class. In the next step the validation phase is discussed.

3.4 Validation

In order to exclude outliers and noisy samples that remained in the list after the two previous sections, validation is essential. We should note that this step is equally applied to the window-based method which is selected as the baseline for collocation extraction. For validation, an association measure (AM) is needed. AM is interpreted as a formula that computes an association score in a pair type's contingency table (Evert, 2004).Among many measures defined to test the dependency between pairs of words or more generally pairs of constituents of a sentence, mutual dependency (MD) is used. As a notification, the measure is defined as equation 2.

$$D(w1,w2) = \log_2 \frac{p(w1,w2)^2}{P(w1)\,P(w2)} \qquad (2)$$

The measure is maximized for the pairs that are dependent. Note that this measure could be replaced by any other measure estimating the probability of co-occurrence within a sentence. Since

the candidates that this measure is trying to test their co-occurrence are the result of Min Direct Algorithm, unrelated pairs are not verified.

4 Evaluation

In order to evaluate this method we performed a comparison between our method and the classical baseline for Collocation Extraction which is the window-based method. Our evaluation approach obviates the necessity of setting a threshold. To have a better baseline, we performed a part of speech tagging to eliminate some noisy pairs from the list of collocations resulted at the end of the process. Maximum size of the window is 5 with expansion to 7 words based on part of speech of the two outmost words in the rare cases. The final results of both two methods are judged manually by three different referees. Every pair not verified by two or three of our referees was not counted as a true sample. Table 2 shows the agreement rate for 500 best results.

	Window method	Our method
Referees 1 and 2	85.0	82.1
Referees 2 and 3	76.5	77.3
Referees 1 and 3	89.2	88.5
Referees 1 and 2 and 3	70.6	69.8

TABLE 2 – Agreement Rate among referees.

At each level, N best pairs are picked and the precision is calculated. Every pair is required to be validated by two out of the three referees. Table 3 shows the precision of both methods in terms of being a sub-part of a Persian MWE.

N	Window Method	Our method
100	77.0	82.0
200	73.5	76.5
300	62.3	69.0
400	61.7	66.5
500	60.2	64.2

TABLE 3 – Precision for N best samples: Each raw shows the precision for N best results of both methods.

To compare the recall of these methods, 200 pairs validated by all of our three referees and 500 pairs validated by two out of the three referees are selected. Table 4 shows the results of the comparison.

	Window Method	Our method
Accepted by two referees	68.3	71.4
Accepted by All referees	69.1	72.9

TABLE 4 – Recall of the methods in each condition. First raw considers the pairs that are accepted by minimum of two judges and the second raw shows the recall in pairs accepted by 3 referees.

Table 5 shows why our method has a better recall in comparison to the window-based method. The results are showed for 100 best results picked by our method. The method is able to capture long distance dependencies. Hence, a noticeable improvement in the recall of the system is occurred. Besides, our method generates less false positive samples.

Distance between pairs	1 or 2	2 or 3	4 or 5	More than 5
Total number of collocations	42	35	14	9

TABLE 5 – 23 out of 100 best pairs are 3 words away from each other and 9 of them more than 5 which makes it impossible for window-based methods to have a reasonable recall.

Conclusion

This paper introduced a method to extract the collocations of the Persian language with a preprocessing phase by means of a dependency parser for the English language. The results suggested that syntactical analysis makes the method more accurate, even if it is implemented in a novel approach. What is important tough is that the accuracy of whole system tightly depends on the accuracy of the parser as well as the alignments between words. Having a noisy parser makes it impossible to achieve competitive results. In other words, it can diminish the benefits of employing the syntactical analysis.

It was concluded that although it is still impossible to have an accurate parser for many languages such as Persian, initial syntactical analysis of corpus is such indispensable that it can lead to a better precision and recall in extracting the collocations even in this kind of implementation.

With no doubt, preprocessing is both essential and beneficial in collocation extraction. Achieving more accurate results is not hindered by the fact that many languages such as the Persian language are under-resourced. The method presented in this paper simultaneously solved the problem of missing long-distance collocations and generation of false positive samples in the earlier methods.

Acknowledgements

We would like to express our deepest gratitude towards Professor Maryam S. Mirian for her valuable advice on our research.

References

Hwa, R. , Resnik, P. , Weinberg, A. , Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3): 311-325.

Pilevar, M. T., Faili, H., and Pilevar, A. H. (2011). TEP: Tehran English-Persian parallel corpus. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing* - Volume Part II, Tokyo, Japan.

Manning, C. D., and Schèutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Liu, Z., Wang, H., Wu, H., and Li, S. (2011). Two-word collocation extraction using monolingual word alignment method. *ACM Trans. Intell. Syst. Technol., 3*(1): 1-29.

Seretan, V., and Wehrli, E. (2006). Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia.

Rasooli, M. S. , Faili, H. , and Minaei-Bidgoli, B. (2011). Unsupervised identification of persian compound verbs. In *Proceedings of the 10th Mexican international conference on Advances in Artificial Intelligence*, Puebla, Mexico.

Sinclair, J.M. (1987). Collocation: a progress report. In R. Steele and T. Treadgold (eds.), Essays in honour of Michael Halliday, (pp. 319-331). Amsterdam: John Benjamins.

Seretan, V. (2011). *Syntax-based collocation extraction,*Berlin: Spriger.

Basili, R. , Pazienza, M. T. and Velardi, P. (1993). Semi-automatic extraction of linguistic information for syntactic disambiguation. *Applied Artificial Intelligence,* **7**, 339-364.

Orliac, B., and Dillinger, M. (2003). Collocation extraction for machine translation. *Machine Translation Summit., 9*: 292-298.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143–177.

Evert, S., and Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France.

Church, K. , and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1): 22–29.

Pecina, P., and Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL*, Sydney, Australia.

LU, Y., and ZHOU, M. (2004). Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Liu, Z., Wang, H., Wu, H., and Li, S. (2009). Collocation extraction using monolingual word alignment method. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Liu, Z., Wang, H., Wu, H., and Li, S. (2010). Improving statistical machine translation with monolingual collocation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Dias, G. (2003). Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, Sapporo, Japan.