Solving the Polysemy Problem of Persian Words Using Mutual Information Statistics

Tayebeh Mosavi Miangah[1]

Abstract

In recent years, large monolingual, comparable and parallel corpora have played a very crucial role in solving various problems of computational linguistics including machine translation, information retrieval, natural language processing, and the like. This paper tries to solve the problem of polysemy of Persian words while translating them into Persian by the computer. We use Mutual Information statistics obtained from a very large monolingual corpus of Persian. Mutual information values are calculated based on co-occurrence frequencies of words and used to measure the correlation between words.

Using mutual information statistics the occurrence or co-occurrence frequencies of different equivalents of an ambiguous word in the target language is calculated and the most probable equivalent for every ambiguous word is selected. When mutual information value is high, the word associations are strong and provide dependable results for translational disambiguation and vice versa.

The method discussed in this paper not only can be directly applied in the system of Persian-English machine translation, but also it can certainly increase performance effectiveness of the retrieval tasks, especially in cross-language information retrieval.

1. Introduction

Automatic translation of texts from one language into another one faces various problems among which we can name lexical ambiguity. Lexical ambiguity refers to a case in which either a lexical unit belongs to different part-of-speech categories with different senses, or to a lexical unit for which there is more than one sense, while these different senses fall into the same part-of-speech category (Mosavi Miangah, 2000). Our concern in this study is solving the second type of lexical ambiguity, that is, those lexical ambiguities in which the different senses of a word fall into the same lexical category. Polysemy refers to a case where a word or phrase has multiple, related meanings. That is, a word or phrase is considered polysemous if it has more than one senses which are related.

During the last decades the application of corpus-based approaches in solving linguistic problems as well as in machine translation has been rapidly growing. In recent years large monolingual, comparable and parallel corpora have played a very crucial role in solving various problems of computational linguistics such as part of speech tagging (Brill, 1995), word sense disambiguation (Mosavi Miangah and Delavar khalafi, 2005), language teaching (Aston, 2000; Leech, 1997; Nesselhauf, 2004), phrase recognition (Cutting et al., 1992), information retrieval (Braschler and

[1] Payame Noor University of Yazd, Yazd, Iran
e-mail: mousavi-t@lit.sku.ac.ir; mosavit@hotmail.com

Schauble, 2000), statistical machine translation (Brown et al., 1990) and some other problems.

Monolingual and bilingual dictionaries as main translation tools as well as terminologies and encyclopedias are now available in different forms whether on paper or in electronic format. Dictionaries follow a synthetic approach to lexical meaning (via a definition), while corpora follow an analytic approach (via multiple contexts). In most cases target monolingual corpora alongside target monolingual dictionaries can be used by translators to check the meaning and usage of possible translation alternatives in the target contexts.

In this paper we tried to solve the problem of polysemy of Persian words while translating them into Persian by the computer. We use Mutual Information statistics obtained from a very large monolingual corpus of Persian. Mutual information values are calculated based on co-occurrence frequencies of words and used to measure the correlation between words. As the aim of disambiguating translational problems is to select the most appropriate choice among many alternatives, mutual information is one of the best methods to measure the degree of association between two co-occurring items within a certain text boundary. In other words, mutual information shows some degree of semantic association between words. Those two words which have the highest mutual information value and hence most strongly associated with each other are most likely to be the correct translations of the query items. It is based on the assumption that when two words co-occur in the same query, they are probably to co-occur in the same in the same affinity in documents. And those words that do not co-occur in the same affinity are not probably to appear in the same query.

2. Related Work

In recent years, the importance of corpora in the field of translation has become noticeable to trainers and researchers. So, some of these researchers believe that the analysis of corpora should be integrated into translator education. There have been a number of studies on monolingual corpora (general and specialized) and various kinds of exploitation of such corpora like collocation extraction.

The majority of the latest research in translation knowledge acquisition is based on parallel corpora (Brown et al.1993). However, since large aligned bilingual corpora are hard to obtain, some researches have tried to exploit translation knowledge from non-parallel corpora such as comparable corpora or monolingual corpora. One of the best known large-scale monolingual corpora is the British National corpus (BNC), a 100 million-word collection of samples of written and spoken language from wide range of sources. However, the BNC has, despite its large size, serious limitations as a translation aid if you are translating contemporary specialized text (Wilkinson, M. 2006).

Yarowsky uses Roget's Thesaurus to disambiguate word senses of English words using statistical models of major categories. By searching the hundred surrounding words for indicators of each category, the most probable category of a word can be determined. During training, by examining the hundred surrounding words for indicators of each category, these indicator words are obtained and weighted. Yarowsky's system needs a large untagged training corpus and a thesaurus. A list of indicator words for each category along with their weights are created, and all these words are reduced to their root forms to achieve more useful statistics by greater occurrence counts. The log of a word's salience for each category is defined as

a weight. Salience is Pr (w|cat)/ Pr (w), that is, the probability that a word appears in the context of a word from a given category, divided by the probability of the word's occurrence in the corpus as a whole. Naturally, the log of salience or the weight will be greater than one for useful words.

Yarowsky's system is not limited to particular vocabulary and works in a wide domain. When testing with ambiguous words previously used for testing other disambiguation systems, this system achieves accuracy of between 72 and 99% (Yarowsky, 1992). This system can cope best with the problem of disambiguation of concrete nouns whose senses can be distinguished by the broad context. Also the system cannot disambiguate topic-independent distinction words that occur in many topics. Another problem with the system is that it does not take account of the distance of words in the contexts it handles. It might be better to consider such natural units like sentences and weight words by their sentence distance from the word in question, rather than a hundred-word context.

Another method for disambiguation of multiple-meaning words presented by Dagan and Itai (1994) tries to select the most probable sense of a word using frequencies of the related word combinations in a target language corpus. In this method the word combinations fall in the limits of the syntactic tuples in the target language. However, first of all the system identifies syntactic relations between words using a source language parser and maps those relations to several possibilities in the target corpus using a bilingual lexicon. Training corpus selection is done using a statistical model and a constraint-propagation algorithm that ensures ambiguities dependent on others are handled properly and simultaneously (Dagan and Itai, 1994).

Dagan and Itai did not evaluate performance using a complete system because some of the required elements (parser and lexicons for the source language) were not available. Two tests were done: one using Hebrew sentences and the other using German sentences. The applicability of the system for Hebrew and German were 68 and 50%, respectively, and the accuracy of the system was 91 and 78% for Hebrew and German, respectively.

As far as the writer of this article is aware there has not been any program for disambiguating Persian polysemous words in terms of their English translation.

3. What is a Corpus?

A corpus is simply defied as a large collection of linguistic evidence mainly naturally occurring data either written texts or a transcription of recorded speech. Such data in the form of corpora can be exploited for a range of research purposes in a number of disciplines.

According to the EAGLES text typology elaborated by John Sinclair (1996) we can make a general distinction between Monolingual and Multilingual (including Bilingual) corpora. Monolingual corpora contain samples of only one language. Multilingual corpora are of two types: comparable and parallel. Comparable corpora contain the same text-types in different languages, while parallel corpora contain the same texts translated into different languages (Hunston, 2002; Kennedy, 1998; McEnery and Wilson, 1996; Meyer, 2002). In bilingual parallel corpora the texts in one language are aligned with their translation in another language.

A large variety of corpora in English and in other languages have been compiled in electronic format for various purposes over the past few decades. The website "Gateway to Corpus Linguistics on the Internet" at http://www.corpus-

linguistics.de/ provides a useful summary of many of the best-known corpora, including information on when and by whom they were compiled, as well as their size, contents, and accessibility.

One of the best-known mega-corpora of British English is the *British National Corpus (BNC),* a 100 million-word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English. It was first released in 1995. The written part (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text (Wilkinson, 2006).

4. Corpus Used in This Study

As far as monolingual corpora for particular text types are very rare especially for low density languages like Persian, some translators find it necessary to construct such corpora to improve their translation performance. Since the texts in the corpus are entirely written by native speakers, the occurrence of the words, collocations and patters are expected to be authentic.

There are some reasons in favor of constructing and exploiting specialized monolingual corpora verses general ones. General monolingual corpora like BNC provide us with too wide range of search patterns and their usages in context, most of which are irrelevant to the task at hand. This problem can be easily solved by using corpora consisting only of the texts of the same or very similar types these kinds of corpora can be considered as a sub-corpora of general corpora, easier to be constructed, handled, and the search results in these specialized corpora are more informative due to the a higher lexical density and repetition for the texts of the same type. Since the frequencies of different meanings of polysemous words are not the same in different text types, it will be less likely to encounter the irrelevant meanings of a certain item in a certain text type. In fact, larger the size of the specialized corpus and the greater the variability of the text- type to be represented, better and more precise the results. Friedbichler and Friedbichler suggest that for English, authoritative specialized corpora of 500,000 to 5 million words (according to the variability of the text-type) should provide solutions to 97% of the translator's questions (Friedbichler, I. and M. Friedbichler, 1997).

The very first stage towards constructing a specialized corpus is collecting a relatively large volume of linguistic data. In light of this, it has been tried to collect as many Persian texts in the field of politics as possible (150 MB, or over 5 million words). These texts are mainly extracted from political articles, journals, interviews, etc. found in the Internet and preprocessed before entering to the corpus. That is, all tables, pictures, figures or diagrams are to be deleted from the texts to be ready for the corpus. Moreover, the texts should be converted to an XML format to be suitable for use on Internet sites. In this stage the texts can be entered into the corpus to be used by translators trying to translate political texts from English into Persian.

5. The Experiment

The method presented by this paper tries to solve the problem of polysemy of Persian words while translating them into Persian by the computer. We use Mutual Information statistics obtained from a very large monolingual corpus of Persian. This corpus is about 150 MB in size and unannotated. This can be regarded as one of the advantages of our method since finding a tagged or annotated corpus for a low-density language such as Persian is nearly impossible.

While reading an English text, we frequently encounter words for which there are more than one Persian equivalents. A straightforward way to find translations of the given terms is to use a bilingual dictionary, however, this method alone faces some problems due to one-to-many correspondences in a bilingual dictionary. Consider the English phrase "*Nuclear talks resumption"* in which all the three words can be translated into multiple Persian words based on an English-Persian dictionary as follows:

Nuclear: اتمي ، هسته اي ، مغزي
Talk: گفتگو ، صحبت ، حرف ، مذاكره
Resumption: ازسرگيري ، ادامه ، تجديد ، شروع

As we can see, each English word has multiple Persian words as its translation which are semantically related and similar but contextually different. In Persian, texts are written from right to left and the order of translation of subsequent nouns in a noun phrase is from the last noun to the first one. So, the correct translation of the above English noun phrase is "ازسرگيري مذاكرات هسته اي". For solving this ambiguity problem we apply a word disambiguation technique using the co-occurrence information extracted from the collection of source language words, here Persian corpus. That is, the mutual information statistics between pairs of words are used to determine the most suitable English equivalent of the ambiguous Persian word in a certain context. The mutual information *MI (x,y)* which is calculated based on word co-occurrence statistics and used as a measure to calculate correlation between words is defined as the following formula (Church and Hanks, 1990):

$$MI(x,y) = Log_2 \frac{P(x,y)}{P(x)P(y)} = Log_2 \frac{Nf_w(x,y)}{f(x)f(y)}$$

Here x and y are the given words in context. The probabilities *p(x)* and *p(y)* are calculated estimated by counting the number of occurrence of x and y in a corpus, *f(x)* and *f(y),* and N is the size of the corpus. $P(x,y)$ is calculated by counting the number of times that *x* is followed by *y* in a window of n words. For this experiment n is between 0 and 6.

Using this formula we can choose the pairs of words that are most strongly associated with each other, thereby eliminating those equivalents that are not likely to be suitable. Table 1 shows the significant *MI* values calculated for the corresponding word pairs (in bold face) as well as the MI values for all the other possible pairs.

The algorithm looks for the first Persian pair with the highest mutual information value and select it the best equivalent for the English pair. This selection will naturally limit the number of subsequent pairs whose mutual information is going to be calculated. The process will go on for the other pairs in the phrase or

collocation. If there were additional pairs to be compared, the same process would be applied to the rest of the network.

In this way, we conducted an experiment using our disambiguation method on a collection of 500 English phrases and collocations consisting polysemous words and gained the accuracy rate of 91.24% which is very encouraging.

x	y	$f(x)$	$f(y)$	$f(x,y)$	$MI(x,y)$
ازسرگيري	گفتگو	67	25467	16	5.5510
ازسرگيري	صحبت	67	21459	9	4.9680
ازسرگيري	حرف	67	9236	3	4.5993
ازسرگيري	**مذاكره**	67	3256	18	**8.6884**
ادامه	گفتگو	32453	25467	1352	3.0319
ادامه	صحبت	32453	21459	1765	3.6635
ادامه	حرف	32453	9236	89	0.5701
ادامه	مذاكره	32453	3256	892	5.3994
تجديد	گفتگو	3429	25467	65	1.8959
تجديد	صحبت	3429	21459	52	1.8210
تجديد	حرف	3429	9236	21	1.7291
تجديد	مذاكره	3429	3256	35	3.9703
شروع	گفتگو	22398	25467	1209	3.4056
شروع	صحبت	22398	21459	1278	3.7327
شروع	حرف	22398	9236	56	0.4367
شروع	مذاكره	22398	3256	78	2.4189
مذاكره	اتمي	3256	6443	255	5.9254
مذاكره	**هسته اي**	3256	6398	432	**6.6961**
مذاكره	مغزي	3256	259	2	3.5678

Table 1: Mutual information values for word pairs in Persian translation of English phrase "*Nuclear talks resumption*"

6. Conclusion

In this paper we tried to solve the problem of polysemy of Persian words while translating them into Persian by the computer. We use Mutual Information statistics obtained from a very large monolingual corpus of Persian. Mutual information values are calculated based on co-occurrence frequencies of words and used to measure the correlation between words. In other words, mutual information shows some degree of semantic association between words.

We carried on an experiment based on our disambiguation method and the results were very encouraging for the pair of languages, namely, Persian and English. We examined the method using 500 English ambiguous phrases and collocations and gained the accuracy rate of 91.24%. The method discussed in this paper not only can be directly applied in the system of Persian –English machine translation, but also it can certainly increase performance effectiveness of the retrieval tasks, especially in cross-language information retrieval.

References

Aston, G. (2000). "I corpora come risorse per la traduzione e l'apprendimento". In Silvia Bernardini and Federico Zanettin (eds.) I corpora nella didattica della traduzione. Bologna: CLUEB, 21–29.

Bowker, L., 1998, "Using specialized monolingual native-language corpora as a translation resource: a pilot study", Meta, 43/4, pp. 631–51.

Braschler, M. and Schauble, P. (2000). Using corpus-based approaches in a system for multilingual information retrieval. Information Retrieval, 3, PP. 273–84.

Brill, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. In 2nd Workshop on Large Corpora, Boston, USA.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. and Roosin, P. (1990). A Statistical Approach to Machine Translation. Computational Linguistics 16:2, 79–85.

Brown P.F., Pietra, S.A.D., Pietra, V. J. D., and Mercer R. L. 1993. The mathematics of machine translation: parameter estimation. Computational Linguistics, 19(2): 263–313.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, Vol. 16, No. 1, pp. 22–29.

Cutting, D.; Kupiec, J.; Peterson, J. and Sibun. P. (1992). A practical part of speech tagger. In proceeding of 3rd Conference on Applied Computational Linguistics, Trento, Italy, PP. 133–40.

Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. Association for Computational Linguistics, 20(4): 563–96.

Friedbichler, I. and M. Friedbichler, 1997, "The potential of domain-specific target-language corpora for the translator's workbench", available online, http://www.sslmit.unibo.it/cultpaps/fried.htm

Kennedy, G. (1998). An introduction to corpus linguistics, London: Longman.

Hunston, S. (2002). Corpora in applied linguistics, Cambridge: Cambridge University Press.

Leech, G. (1997). Teaching and language corpora: A convergence. In: A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), Teaching and language corpora (1–23). New York: Addison Wesley Longman.

McEnery T. and Wilson A. (1996). Corpus linguistics. Edinburgh: Edinburgh University Press.

Meyer, C. F. (2002). English corpus linguistics: An introduction. Cambridge: Cambridge University Press.

Mosavi Miangah, T. (2000). Ambiguity problem in English–Persian machine translation. Problems of Language Theory and Translation Science, 4: 88–98.

Mosavi Miangah, T. and Delavar Khalafi, A. (2005). Word sense disambiguation using target language corpus in a machine translation system. Literary and Linguistic Computing, 20(2), 237–49.

Myung-Gil Jang, Sung Hyon Myaeng and Se Young Park (……). Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In: J. McH. Sinclair (Ed.), How to use corpora in language teaching (125–152). Amsterdam: Benjamins.

Sinclair, J. McH. (1996) EAGLES Preliminary recommendations on Corpus

Typology, EAG–TCWG– CTYP/P. Online: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

Wilkinson, M, (2006). Compiling Corpora for Use as Translation Resources, Translation Journal, Vol. 10, No. 1.

Yarowsky, D. (1992). Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proceedings of 15th International Conference on Computational Linguistics, pp.454–60.