Word Sense Disambiguation Using Target Language Corpus in a Machine Translation System

Tayebeh Mosavi Miangah and Ali Delavar Khalafi
Shahre Kord University, Iran

Abstract
This article studies different aspects of a new approach to word sense disambiguation using statistical information gained from a monolingual corpus of the target language. Here, the source language is English and the target is Persian, and the disambiguation method can be directly applied in the system of English-to-Persian machine translation for solving lexical ambiguity problems in this system. Unlike other disambiguation approaches, using corpora for handling the problem, which use the Most Likelihood Model in their statistical works, this article proposes the Random Numbers Model. We believe that this model is more reasonable from the scientific point of view and find that it offers the most precise and accurate results. This method has been tested for a selected set of English texts containing multiple-meaning words with respect to Persian language and the results are encouraging.

1 Introduction
Ambiguity is the most serious problem in natural language processing systems, and machine translation systems suffer from this problem to a high degree. The problem of ambiguity in translating texts by machine is different from human translation. The human mind is a complex machine which can choose the suitable target equivalent(s) of any source language forms, sometimes without even being aware of irrelevant alternatives, based on the understanding of the context. It can also automatically consider a group of words, rather than one individual word, in order to understand the meaning of a sentence, even if the words of the group are not relevant. For a machine to do this, a huge amount of data would be needed as input and the output may still not be error free.

In recent years, the application of statistical approaches and the study of statistics-based methods in natural language processing as well as in machine translation have been rapidly increasing. Statistical
linguistics basically relies on the study of frequencies of various linguistic units, including word-forms, lexemes, morphemes, letters, etc., in a sample corpus on the basis of which to calculate their probability, so that various linguistic problems such as ambiguity can be solved. Statistics-based approaches rule out the need for computational mechanisms and extensive linguistic knowledge for solving linguistic problems. Hence, the computational cost of a statistics-based approach is much lower than a knowledge-based or a rule-based approach (Su and Chang, 1990). A statistics-based system needs a large database or corpus to guarantee its reliability. Nevertheless, with the availability of many tagged and untagged text corpora, acquiring linguistic knowledge from a large sample corpus is no longer an impossible task (Garside et al., 1987).

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, using statistical information about different equivalents of English ambiguous words in the target language, Persian. By statistical information we mean to calculate the occurrence or co-occurrence frequencies of the ambiguous word equivalents in the target language and select the most probable equivalent for every ambiguous word using a statistical model.

Several statistics-based methods for word sense disambiguation have been developed using large tagged or untagged corpora. Brown et al. (1991) propose a word sense disambiguation which can be applied in machine translation systems. They collect data from syntactically related words in the local context of the ambiguous word. The acquisition of statistical data relies on the availability of a word-aligned bilingual corpus. This method overcomes the limitation of the trigram language model which only captures local phenomena and it is incorporated into the statistical machine translation model described in Brown et al. (1990). Each occurrence of an ambiguous word should be labeled with a sense that will increase the mutual information between the member of a connection. A connection is a pair of aligned words in the bilingual corpus. An instance of a word is assigned a sense by asking a question about the context in which the word appears. The question is constructed to have high mutual information with the translation of that instance in another language (Brown et al., 1991).

The system was tested by translating 100 randomly selected Hansard sentences, each of 10 words or less in length. Forty-five of the resultant translations were judged as acceptable as compared with thirty-seven acceptable translations produced by the same system running without sense-disambiguation questions.
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 \( \frac{Pr(w|cat)}{Pr(w)} \), 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.

Dagan and Itai’s system uses only words in specific syntactic relations. So, it cannot gain information from other words in the local context. In some cases where there are several relations involving
a word, incorrect decisions can be made by choosing the wrong relation, the relation whose frequencies point toward the incorrect choice as more important, because there are more instances of it in the corpus. It might be useful to weight relations so that close relations, such as those between a noun and its object are considered more important (Marshall, 1998).

Justeson and Katz (1995) describe a linguistically principled approach to disambiguation. Their system only uses syntactically or semantically relevant clues. Although they use statistical methods for data analysis in training, the disambiguation stage does not use statistical techniques. Their system disambiguates adjectives using only nouns that are modified by the adjectives. For this work, they need a training corpus, which can be generated by selecting sentences from a large corpus in which antonym pairs co-occur and parsing these to ensure the adjectives refer to instances of the same noun. Hence, they need a parser and an anaphoric resolution to determine the underlying noun that has been modified, as well as a morphological analyzer for nouns. Justeson and Katz manually simulated these three requirements and performed all testing on five of the most frequent ambiguous adjectives in English: ‘hard’, ‘light’, ‘old’, ‘right’, and ‘short’ on sets of hundred randomly selected sentences from the corpus that contained the adjectives (Justeson and Katz, 1995).

The accuracy of their system in testing all nouns reached 97%. These results are desirable only for the experiment in which adjectives are commonly used with their antonyms. However, this is not feasible practically, since some adjectives occur less frequently with their antonyms. Moreover, there are some adjectives which can be differently accompanied by the same noun for which this method cannot be helpful in disambiguating them.

Ng and Lee (1996) use exemplars, collections of feature values, which represent instances of the given word from the training corpus, and then compare the word with the current instance of the word using a distance metric. The features chosen for their informative nature are as follows: singular/plural; POS tags of the current word; three words on either side; support for verbs, which have a different verbal morphological feature; a verb–object syntactic feature for nouns; and nine local collection features. These features are calculated for each instance of w in the sense-tagged training data. The results are stored as exemplars of their senses. By calculating the same features vector for the current word and comparing against all the examples of that word, the given word is disambiguated choosing the closest matching exemplar (Ng and Lee, 1996).

Results which were calculated on a task including 121 nouns and 70 verbs, using fine-grained sense distinctions from WordNet show a 58% accuracy on a test set from the Brown Corpus and a 75.2% accuracy on a test set from the WSJ Corpus.

The method presented by Brown et al. requires a bilingual word-aligned corpus, which is costly to build. This is one of the disadvantages
of this method, which makes difficult the applicability of the method to other pairs of languages.

The method presented by this article is along the lines of the work by Dagan and Itai (1994), who use a target language model to disambiguate word translations with some kinds of manipulations in order to conform to the properties of Persian as the target language. We consider the co-occurrences of the multiple-meaning words in a monolingual corpus of the target language, namely Persian. By calculating the frequencies of these words in the corpus, we can select the most probable sense for these multiple-meaning words. However, instead of considering syntactic tuples in the target language corpus, we consider only co-occurrences of certain words in that corpus without having a syntactic analysis for corpus. In this method there is no need to analyze either the source or the target language corpus from the syntactic point of view. The only task of our algorithm, for gaining the required statistical information, is determining the nearest noun, pronoun, adjective, or verb to our ambiguous word, whether it is a noun, a verb, an adjective, or an adverb. Table 1 describes the conventions in detail. However, when applying this method for the comparison of English and Persian, only a small portion of ambiguous words in English can be correctly translated into Persian. For some others even the use of syntactic tuples and syntactic relations between the ambiguous word and other words in the corpus has not yielded satisfactory results. The most appropriate and convenient way to disambiguate multiple-meaning words seems to specify the domain or field of texts to which such words belong. This is discussed in detail in the next section.

2 Linguistic model

Our model has been implemented within the framework of Marchuk’s theory (1998) of machine translation, which is called ‘the theory of
translation equivalencies’ or ‘translational correspondencies’. It is based on the assumption that translation per se (as opposed to the interpretation of the source text context) may be and should be performed using only the means offered by the systems of the languages involved (Marchuk, 1988; Miram, 1998). In order to carry out the experiment, first of all we need a machine-readable bilingual dictionary of English to Persian to be able to distinguish all possible translations of each word, especially those of the ambiguous words which we will focus on in this study. For determining the correct equivalent of each ambiguous word in certain sentences in the source language, namely English, our algorithm searches for each of its alternative translations (within a single part of speech) in the monolingual corpus of the target language, namely Persian, and calculates the frequencies of occurrences along with their nearest linguistic units referred to in Table 1. Then, with the help of a Random Number Model, the most probable alternative for every English multiple-meaning word is selected as the most appropriate Persian equivalent for that word.

Consider, for example the following English sentence extracted from the textbook Psychology Applied to Teaching.

Tentative analysis of the behavior has been provided an acceptable perception of learning process by which we can overcome many problems of the primary students (Biehler, 1974).

In the above sentence, each of the underlined parts has more than one equivalent in Persian, although in English they may not be known as ambiguous words. In what follows we illustrate the Persian translation of this English sentence including all alternatives for each of these ambiguous words.

Tajziyeye azmayeshiye raftar darke ghabeleghabuli az farayande yadgiri/amuzesh/danesh be dast dadeh ast/tahiyeh kardeh ast, ke be vasileye an ma mitavanim bar besyari az moshkelate daneshamuzane/danesgjuyane ebetaaie avaliye ghalabeh konim.

The verb ‘provide’ has two equivalents in Persian. For selecting the most suitable one we should compare its co-occurrence frequency with its complement, which is the nearest noun or pronoun by which it follows or precedes, here, ‘perception’. Identifying the nearest noun or pronoun to the verb is not a difficult task due to the POS tagger availability for the text. Referring to our Persian corpus, we extract the following co-occurrences for the word combination ‘to provide perception’: be dast dadane dark, 14 times, and tahiyeh kardane dark, 0 times. Naturally we prefer the first-expression for the best suitable equivalent for the verb ‘provide’. To find the most appropriate equivalent for the word ‘learning’ in Persian we should calculate the frequency of its alternative co-occurrences with the nearest noun (here, ‘process’) in a monolingual corpus of its related field (here, psychology and learning). Referring to our monolingual Persian corpus in this field, we find that the noun phrase farayande yadgiri appears 240 times.
in our corpus, while the noun phrase farayande amuzesh appears 20 times, and farayande danesh does not occur at all. Using the statistical model we prefer yadgiri to amuzesh as the more appropriate translation for the word ‘learning’.\(^1\) The English noun phrase ‘primary students’ also has four alternative translations in Persian: daneshamuzane ebtedaii, daneshamuzane avaliye, daneshjuyane ebtedaii, and daneshjuyane avaliye. In our Persian corpus we find that the first combination appears 150 times, the second 15 times, the third 0 times, and finally the last one 8 times. Using our statistical model we can choose the alternative daneshamuzane ebtedaii as the best translation equivalent of the English noun phrase ‘primary students’.

When we consider all different aspects of this algorithm we see that it cannot cope with all types of multiple-meanings in different domains, if we only rely upon a single monolingual corpus. This means that one word or a combination of words in one domain may appear more frequently than in another. Hence, to achieve more precise and satisfactory results, it is better to extract frequencies of ambiguous words from a target language corpus, which covers the domain comparable to the source language. Following this procedure, frequencies gained from the target language corpus will illustrate the actual frequencies, with which we can work to disambiguate many multiple-meaning words of the source language, in particular special terms of that kind of text. For instance, our previous example ‘learning process’ has been extracted from an English psychology text, and statistical data has also been gained from a psychology corpus of the Persian language. Our assumption is that searching for different equivalents of every ambiguous word and the co-occurrences of these words in a general corpus seems an impossible task, since the frequencies of ambiguous words and their co-occurrences appearance in a general corpus is not sufficiently high to be helpful for calculating the probability of their occurrences.\(^2\) For example, the English word ‘old’ may appear thousands of times in a general corpus accompanied by different nouns, but the noun phrase ‘old Persian’ in which the adjective ‘old’ has different Persian equivalents from the ‘old’ appearing in noun phrases ‘old friend’ or ‘old shoes’ may appear only a few times. Thus, to find the most suitable equivalent for the ambiguous word ‘old’,\(^3\) it is more logical to calculate its frequencies of co-occurring with the word ‘Persian’ in a philology corpus or some similar fields in which this phrase appears more frequently.

### 3 Statistical model

There are several statistical models which can resolve the ambiguity problem introduced in this study including the Hidden Markov Model, Bayes law (Charniak, 1993), the Most Likelihood Model (Dagan and Itai, 1994) and the Random Numbers Model (Shannon, 1975) to name a few. The Random Number Model has been used to solve linguistic problems in this article for the first time. Here, we use Random

---

1 It can be said that we are concerned with the two Persian equivalents of the English word ‘learning’, namely, yadgiri and amuzesh which are sometimes synonyms and sometimes antonyms in Persian.

2 Dagan and Itai were able to use a general corpus in their experiment (Dagan and Itai, 1994).

3 It is ambiguous from the standpoint of its translation into Persian.
Numbers, which is a device for simulation, because it is more reasonable from the scientific point of view. When we select a target language equivalent with the Most Likelihood Model (the highest probability) in the target corpus as the correct choice, the probability of choosing the alternative equivalents with the lower probability will be practically zero, while in a proportion of cases it should be possible to select them.

To determine the appropriate sense of a certain word, first of all alternative combinations for the given word and the frequency of each of them in the target language are to be extracted from the monolingual corpus using the algorithm designed for this purpose. Suppose our word has \( n \) different senses \( tw_1, \ldots, tw_n \). First, we get the frequency of each of these alternative senses \( f_1, \ldots, f_n \), and then we calculate their probabilities from the following probability function:

\[
P(tw_i) = \frac{f_i}{\sum_{j=1}^{n} f_j} \quad i = 1, \ldots, n.
\]

where \( P(tw_i) \) stands for the occurrence probability of the \( i \)th sense of the word, \( f_i \) stands for the frequency of \( i \)th sense of the word, and \( \sum_{j=1}^{n} f_j \) stands for the sum of total frequencies of different senses of a word, where \( i \) varies from 1 to \( n \). Then a related table is constructed (Table 2).

Using the empirical distribution table and some goodness of fit tests, we can find the best statistical distribution for the observed sample (Phillips, 1972). For large samples (\( n \geq 100 \)) the Chi-square test is very useful; however, for samples with a quantity smaller than 10 (\( n \leq 10 \)) it seems that the Cramer–Von Mises test (Phillips, 1972) is more appropriate than others. And nearly all of our samples in this study are of this sort, namely, smaller than 10 different senses for each ambiguous word. We choose the Cramer–Von Mises test to find the best statistical distribution for our samples. The statistical distribution of the observed sample was as follows:

\[
X \sim f_X(tw)
\]

Using random numbers produced by this distribution, one of the \( n \) senses of a given word can be selected (Jansson, 1966). The algorithm of this model can be displayed as follows:

1. Start
2. Produce a random number with Uniform Distribution
3. Produce a random number of the observed community \( X \sim f_X(tw) \)

Table 2: Empirical distribution table

<table>
<thead>
<tr>
<th>( i )</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(tw_i) )</td>
<td>( P_1 )</td>
<td>( P_2 )</td>
<td>\ldots</td>
<td>( P_n )</td>
</tr>
</tbody>
</table>

4 Curve fitting to find statistical distribution can be obtained using some statistical packages with a high degree of precision.
5 Naturally, if a simulation model is computerized, we should have some means to be able to (1) get random numbers with Uniform Distribution and (2) produce random numbers with desirable characteristics using these random numbers.
(4) Determine the word sense by the random number produced in stage 3.

(5) End

To prevent undesirable effects of the random numbers on the ultimate results, it is necessary to repeat the selection of random numbers and choose one of the alternative senses from the corpus repeatedly (practically, at least for 30 random numbers) until the word sense with the highest repetition is chosen. To demonstrate how our algorithm and the Random Numbers Model work, first of all we construct Table 3 according to the counts gained from our sample English sentence and its corresponding Persian translation.

Then, we calculate the probability of the alternative translations of the ambiguous source co-occurrence 'primary students' from the following formula (Table 4):

\[
P(tw_i) = \frac{f_i}{\sum_{j=1}^{n} f_j} \quad i = 1, \ldots, n.
\]

At this stage, using the above empirical distribution table and goodness of fit tests we can calculate the distribution of geometric probability with the parameter \( P = 0.848039 \)

\[
X \sim f_X(tw) = \begin{cases} \frac{pq^{tw}}{1-p}, & if tw = 0, 1, 2, 3, 4, \ldots, \\ 0, & otherwise \end{cases}
\]

where \( 0 \leq p \leq 1, \ q = 1 - p \). To produce random numbers from geometric distributions with the parameter \( P = 0.848039 \), first we produce the following Bernoulli Random Variable (Sheldon, 1976)

### Table 3.
The alternative target co-occurrences for every ambiguous source word with their counts in the target language corpus

<table>
<thead>
<tr>
<th>Source Co-occurrence</th>
<th>Alternative Target Co-occurrence</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provide-perception</td>
<td>(1) be dast dadan-dark</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>(2) tahiyeh kardan-dark</td>
<td>0</td>
</tr>
<tr>
<td>Learning-process</td>
<td>(1) farayand-yadgiri</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>(2) farayand-amuzesh</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(3) farayand-danesh</td>
<td>0</td>
</tr>
<tr>
<td>Primary-students</td>
<td>(1) daneshamuzan-ebtedaie</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>(2) daneshamuzan-avaliye</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>(3) daneshjuyan-avaliye</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>(4) daneshjuyan-ebtedaie</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 4.
Empirical distribution table for co-occurrence ‘primary student’ in Persian corpus

<table>
<thead>
<tr>
<th>( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(tw_i) )</td>
<td>0.867052</td>
<td>0.086705</td>
<td>0.046243</td>
<td>0</td>
</tr>
</tbody>
</table>
using uniform random numbers $U \sim U(0,1)$, $i = 1, 2, 3, \ldots$ which can be extracted from random number table. The related table can be found in Sheldon, R. (1976) as well as at http://www.pfks.org/toolkits/tutv/random.pdf

$$X_B = \begin{cases} 1 & U \leq 0.848039, \\ 0 & U \geq 0.848039. \end{cases}$$

We continue the Bernoulli tests until we get the desirable result $X_B = 1$. The maximum number of repetitions of random numbers indicates our selected word. Table 5 has been completed using 66 random numbers and $N$ is the number of repetitions of the selected sense of the ambiguous word:

In this case we may state that $tw_1$ can be selected as the best Persian equivalent for the co-occurrence ‘primary students’.

### 4 Processing steps

This section describes the processing steps of this experiment in detail. To begin with, it is necessary to compile a bilingual lexicon from English to Persian including all possible translations of each English word into Persian. This kind of lexicon or dictionary is already available in the form of the machine-readable dictionary of English to Persian. It is well known that many English words belong to different parts of speech. In our dictionary there is one Persian equivalent for each part of speech of the given words. However, sometimes a word may have different senses in a single part of speech category. We are interested in ambiguous words of the latter type in this experiment.

After identifying the word for which we want to find the suitable target equivalent, the next stage naturally is to collect a Persian monolingual corpus in which we can find different equivalents of the mentioned word accompanied by some certain nouns, pronouns, adjectives, or verbs with different frequencies. At this stage we try to collect a separate Persian corpus for each domain and then we can refer to that special corpus of the target language with regard to the subject matter of our source text.

To gain statistical data from the Persian corpus, we are mainly concerned with the occurrence of different alternative translations of every ambiguous source language word in the target language corpus and their co-occurrence noun, pronoun, adjective or verb. That is, if the ambiguous word is a noun we consider its modifying adjective

<table>
<thead>
<tr>
<th>$i$</th>
<th>$tw_1$</th>
<th>$tw_2$</th>
<th>$tw_3$</th>
<th>$tw_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>39</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5. Frequencies of possible senses of the co-occurrence ‘primary student’ in Persian corpus
or noun where there is one. However, if the noun occurs alone without any modifying word we consider the nearest verb either before or after that. If the ambiguous word is an adjective, naturally its noun is considered. However in some cases, an adjective may occur without any noun, so we should consider the nearest noun or pronoun to it. If the ambiguous word is a verb, whether transitive or intransitive, its nearest noun, pronoun or adjective is considered, and naturally for a transitive verb the complement, in most cases, is the nearest noun or pronoun after that.

In usual models of statistical study of linguistic data, researchers often tend to use the Most Likelihood Model for selecting the best case in linguistic problems. However, this method is not a scientific one, since it changes a probabilistic model into a deterministic one. This can be a dangerous approach, because in some cases we may encounter a linguistic unit with the highest occurrence frequency, which cannot serve as the best choice (correct sense, for instance) in a given context. In these cases the Random Number Model shows itself as a scientific as well as practical solution and chooses the best choice according to its probability distribution relying on the data in the given corpus. A similar subject has been discussed by Law and Kelton (2000) as ‘the danger of replacing a probability distribution by its mean’ referring to which may be helpful for a better understanding of the matter (Law and Kelton, 2000).

In this article, using a kind of probabilistic model (known as Random Number Model) for available data, we do not necessarily choose the case with the highest probability, but give the other cases with lower probabilities the possibility to appear as the correct choice according to their occurrence frequencies in the corpus. Although in the majority of cases choosing the correct case selected by our model and Most Likelihood Model coincide with each other, these two models may give different results for linguistic problems in which the case with the highest frequency and the case next to it in the given corpus have a small interval.

4 Conclusion

While the method proposed in this article for multiple-meanings disambiguation bears some resemblance to the experiment carried out by Dagan and Itai (1994), our method has several advantages in both linguistic and statistical models. For one thing, it uses only lexical co-occurrences both in the source language text and in the target language corpus instead of syntactic tuples. As such, we do not need any syntactic parser, whether for the source or for the target language. The only need is a simple part-of-speech tagged corpus for the target language.

To obtain data from the target language corpus, we used a domain-specific monolingual corpus for texts of every domain in the source language. In this way, the number of counts for any lexical
co-occurrence in the target language corpus will be statistically significant. Moreover, the results will be more precise and accurate. In this experiment we used random numbers and goodness of fit tests in the statistical part, while all the previous methods for word sense disambiguation used Most Likelihood Model in the statistical part of their works. We believe that using random numbers for selecting the best target equivalent for an ambiguous word of the source text in a machine translation system gives the results closer to reality and can be more precise than when selection of the most probable case is used. The precision of the proposed model, which is a working system for English–Persian machine translation in specific domains has been tested for a relatively large corpus of psychology (specific domain) in English as well as in Persian. The algorithm coped with the problem of ambiguity of 604 ambiguous words in related English text out of 764 ambiguous words, considering the related Persian text. Thus, the precision of the model has been calculated as 79%.

This approach is directly applied in the system of English-to-Persian machine translation. In this system, the problem of multi-meaning words and ambiguity resolution is one of the major questions for which up to now no answer has been found. While the experimental results are very encouraging for the comparison of English and Persian languages, the procedure may be applied to other pairs of languages as well.

References


