Statistical analysis of target language corpus for word sense disambiguation in a machine translation system

Tayebeh Mosavi miangah[1] and *Ali Delavar khalafi*[2]
[1]English Language Department, Shahre Kord University, Shahre Kord, Iran.
E-mail: mousavi-t @lit.sku.ac.ir
[2]Mathematics Department, Shahre Kord University, Shahre Kord, Iran.
E-mail: delavar@sci.sku.ac.ir

Abstract

This article studies different aspects of a new approach for word sense disambiguation using statistical information gains from a target language monolingual corpus. Here, the source language is English and the target one is Persian, and this disambiguation method in those aspects which gives desirable results can be directly applied in the system of English-to-Persian machine translation for solving lexical ambiguity problems in this system. Unlike the other disambiguation programs using corpora for handling the problem, which use probabilistic Model in their statistical works, this paper uses Simulation Model. We believe that this model is more reasonable from the scientific point of view with the most precise and accurate results. This method has been tested for a selected set of English texts having some multiple-meaning words in respect to Persian language and the results are encouraging.

1. Introduction

Ambiguity is the most considerable problem in natural language processing systems, among which machine translation systems suffer this problem in a high degree. The problem of ambiguity in translating texts by the machine is different from one by the human. Human is a complex machine so that he can choose the suitable target equivalent(s) of any source language forms, sometimes without becoming aware of the irrelevant alternatives, based on his understanding in the context. He can also automatically consider a group of the words, rather than individual word to understand the meaning of a sentence, even if the words of the group are not relevant. But a machine cannot think at all. Since only written texts are presented to the computer, it can not mechanically use the relevant text. However, this problem has been somehow solved by the field of discourse analysis which is not within the scope of this study.

Nowadays the application of statistical approaches and studying statistical-based methods in natural language processing as well as in machine translation has been rapidly increasing. Statistical linguistics basically relies on the studying of various linguistic units occurrences frequencies including word-forms, lexemes, morphemes, letters, etc. in a sample corpus, and solving various linguistic problems as ambiguity with reference to these certain frequencies and calculating the probability of them. Statistics-based approaches factor out the need for computational mechanisms and high linguistic knowledge for solving linguistic problems. So 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, however, with 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 lexical categories with different senses, or a lexical unit for which there are more than one sense while these different senses fall into the same lexical 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 calculating the occurrences or co-occurrences frequencies of the ambiguous word equivalents in the target language and selecting the most appropriate equivalent for every ambiguous word using a statistical model.

Several statistics-based methods for word sense disambiguation have been recently developed using a large tagged or untagged corpus. However, most of these systems are dealing with lexical ambiguity of the first type mentioned above, that is, those lexical units which belong to different lexical categories. A method for word sense disambiguation which was described by Marshall is known as CLAWS system. The main feature of this system is using a collocational probabilities matrix showing the relative likelihood of co-occurrences of all ordered units previously tagged by WORDTAG program. CLAWS uses a large proportion of the tagged Brown corpus to generate statistics of each tag frequency and also the frequency of any two tags adjacent to each other (Marshall, 1983).

Another system known as VOLSONGA has been designed to overcome the non-polynomial complexity of CLAWS. This system does not use tag triple and idioms, and by this reason it is not necessary to manually construct special lists. VOLSUNGA considers only two successive tags in each time of calculation and so reduces the algorithm from exponential complexity to linear. It only requires a tagged corpus based on which establishes its tables of probabilities (De rose, S. J. 1988) .

Considering the two methods of disambiguation mentioned above we can see that they use a previously tagged corpus to disambiguate those lexical units which have more than one lexical category. After determining the suitable category for a word by these methods, in the case that it has more than one meaning in the limits of its category, the problem of ambiguity still remains.

A rather novel method for disambiguation of multiple-meaning words presented by Dagan and Itai tries to select the most probable sense of a word using frequencies of the related word combinations in a second language corpus. In this method the word combinations fall in the limits of the syntactic tuples in the second language. However, first of all the system identifies syntactic relations between words using a source language parser and maps the alternative interpretations of these relations to the target language using a bilingual lexicon (Dagan and Itai, 1994).

Koehn and Knight (2000) have also developed a novel method in which they use only unrelated monolingual corpora and a lexicon. By estimating word translation probabilities using the EM (expectation maximization) algorithm, they extended upon target language modeling. They propose the use of syntactic relations such as subject-verb, verb-object, adjective-noun to disambiguate word translations. However, Koehn and Knight focus in their experiment on nouns to simplify their experimental setup. They combine the notion of translation probabilities with the use of context.

The method presented by this study is somehow similar to the two methods mentioned above 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. 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 certain co-occurrences words in that corpus without having a syntactic analysis for corpus. In this method there is no need to analyze the second 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, with applying this method for the pair of English and Persian languages 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 been satisfactorily successful. So, the most appropriate and convenient way to disambiguate the multiple-meaning words seems to specify the domain or field of texts to which such words belong. In the following lines we will discuss the matter in detail.

	ambiguous word	possible combinations respectively	example
if	noun (N)	1) adjective + N 2) noun + N 3) N + noun 4) verb + N 5) N + verb	sandy bank river bank bank robber He borrows from the bank. The bank works 24 hours a day.
if	adjective (Aj)	1) Aj+ noun 2) noun/pronoun +Aj	old man He is old.
if	verb (V)	1) V + noun/pronoun 2) noun/pronoun + V	He minds the baby. Do you mind the baby
if	adverb (Av)	1) verb + Av 2) Av + verb	He has not yet seen it. Yet, you should go further.

Table 1. Possible combinations (co-occurrences) of an ambiguous word with the other parts of speech in source language

1. 1. Persian background

Persian is a member of synthetic language family. It means that in Persian a new word is created by adding prefix, suffix, infix or another noun, adjective, preposition or verb to the beginning or the end of the word. In these cases the basic form of the word or verb stem usually is not broken (Mosavi miangah, 2001). Grammatical word order in Persian is shown as SOV, although a rather free word order is also possible but not grammatically acceptable. In this language every verb has two stems, present stem and past stem, and different inflectional forms of a verb is constructed using either the present or past stem.

Persian uses Arabic alphabet. Texts are written from right to left. Short vowels (a, e, o) are usually not written; only the long vowels (y, u, a:) are represented in the text.

Persian morphology is an affixal system consisting mainly of suffixes and a few prefixes. The nominal paradigm consists of a relatively small number of affixes. The verbal inflectional system is quite regular and can be obtained by the combination of prefixes, stems, inflections and auxiliaries.

The elements within a noun phrase are linked by the enclitic particle called **ezafe**. This morpheme is usually an unwritten vowel, but it could also have an orthographic realization in certain phonological environments. Adjectives follow the same morphological patterns as nouns. They can also appear with comparative and superlative morphemes. Certain adverbs, mainly manner adverbs, can behave like adjectives and can appear with all the adjectival affixes.

The inflectional system for the Persian verbs consists of simple forms and compound forms; the latter are forms that require an auxiliary verb. The simple forms are divided into two groups according to the stem they use in their formation, present or past. The citation form for the verb is the infinitive (Megerdoomian, 2000).

2. Linguistic model

Our model has been implemented within the framework of Marchuk's theory of machine translation called as "the theory of translation equivalencies" or " translational correspondencies".

It is based on an 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, and Miram, 1998).

In order to carry out the experiment, first of all we need an automatic bilingual dictionary of English to Persian to be able to distinguish all possible translations of each word especially those of the ambiguous words which are our aim in this study. For determining the correct equivalent of each ambiguous word in the certain sentences in the source language, namely, English our algorithm searches each of its alternative translations (within a single part of speech) in the monolingual corpus of the target language, namely, Persian and calculate the frequencies of their occurrence along with their nearest linguistic units referring to the table 1 separately. Then by the help of a simulation model the most probable alternative for every English multiple-meaning word are 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 above sentence the words underlined have more than one equivalents in Persian although in English they may not be known as ambiguous words. In the following lines we illustrate Persian translation of this English sentence including all alternatives for each of its ambiguous word.

> *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 ebtedaii/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 by which it follows, here, "perception". 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 case 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 see that the noun phrase *farayande yadgiri* has been appeared 240 times on our corpus, the noun *phrase farayande amuzesh* has been appeared 20 times and *farayande danesh* 0 times. Using 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 as *daneshamuzane ebtedaii, daneshamuzane avaliye, daneshjuyane ebtedaii* and *daneshjuyane avaliye*. Considering our Persian corpus we can see that the first combination has been appeared 150 times, the second 15 times, the third 0 times and finally the last one 8 times in our corpus. Using our statistical model we can choose the alternative *daneshamuzane ebtedaii* as the best equivalent translation of English noun phrase "primary students".

2. 1. Processing steps

In the following lines it has been tried to describe 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 automatic dictionary of English to Persian. As every one knows, many English words belong to different parts of speech and in our dictionary there is one Persian equivalent for each part of speech of the given words unless it has more than one equivalent in a single part of speech. The latter case is the target word for disambiguation in this experiment. Determination of the word part of speech is supposed to be carried out by a syntactic parser and some context-frame rules (Mosavi miangah, 2002). Thus, the reminding ambiguities which fall within the scope of a single part of speech have to be resolved by the present procedure.

After distinguishing 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. In this stage we try to collect a separate Persian corpus for each field of knowledge and then we can refer to that special corpus of the target language considering the subject matter of our source text.

To gain statistical data from the Persian corpus we are mainly concerned with the

[1] It can be said that the two Persian equivalents of the English word "learning", namely, *yadgiri* and *amuzesh* sometimes are synonym and sometimes antonym in Persian language.

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 or noun in the case that there is one, however, if the noun occurs alone without any modifying word we consider the nearest verb either before or after that. Consider, for instance, the following sentences:

a) Give me a glass of water.
b) Frying the potatoes she hurts her hand by *hot* oil.
c) If you need so much money, you can *borrow* it from the bank.

In sentences above, the words underlined are ambiguous words from the standpoint of Persian. However, existence of the italic words in the sentence determine the suitable Persian equivalents of the ambiguous word by calculating the frequencies of the co-occurrences in the Persian monolingual corpus. It means that the word "glass" in sentence a) when acts as a noun, has several equivalents in Persian as *livan, shisheh adasi* and *aiineh* , however, when it accept the word "water" as its modifying noun in a sentence, the first Persian equivalent is appropriate for that sentence because the occurrence frequency of *livan-e ab* is much more than the frequency of *shishe-ye ab, adasi-ye ab or aiine-ye ab*. So, the algorithm selects *livan* over the other equivalents in most cases (depending on the rate of probability each of the alternatives has) for the word "glass" in sentence a). In this manner, our algorithm will be also able to find the most suitable Persian equivalent for ambiguous word "oil" and "bank" in sentences b) and c) by calculating their co-occurrences with the words "hot" and "borrow" respectively.

If the ambiguous word is an adjective, naturally its noun has to be considered. However in some cases an adjective may occur without any noun, so we should consider the nearest noun or pronoun to it. Consider, for instance, the following sentences:

a) She has fair hair.
b) This land is even.

The adjective "fair" has several equivalents in Persian as *ziba, roshan, bour, monsef*. When we refer to its noun, namely, "hair" and calculate the frequency of the co-occurrences of its Persian equivalent with all equivalents of the word "fair" in Persian, the result are very encouraging. That is, the frequency of occurrences of two words *mooye bour* is in a high degree higher than the other cases. In sentence b), although the ambiguous adjective "even" has no noun combined to it, we can go backward to the nearest noun, namely, "land" and search for their co-occurrence frequencies of the alternative equivalents of this ambiguous word in Persian. In this case we see that word *hamvar* is the most appropriate Persian translation for the ambiguous word "even", because the co-occurrence of the words *sarzamin* (land) and *hamvar* is much more frequent than the other co-occurrences in our selected Persian corpus.

If the ambiguous word is a verb, whether transitive or intransitive, its nearest noun or pronoun is considered, and naturally for a transitive verb the complement, in most cases, is the nearest noun or pronoun after that. Consider, for example, the following sentences:

a) Britain has recognized the new regime.
b) Who is minding the baby?

In sentence a) the verb "recognize" which has several Persian equivalents as *shenakhtan, be rasmiyat shenakhtan, qabul nemudan* and *qadr dani kardan* takes the second alternative due to occurrence of the noun "regime" as its complement. In this manner the verb "mind" in sentence b) which is an ambiguous word whether in Persian or in English is translated as {\it movazebat kardan}, over the other alternatives (*ahammiyat dashtan, be khater avardan, residegi kardan*, etc.) due to occurrence of the noun "baby" as its complement.

2. 2. Some limitations

When we consider all different aspects of this algorithm we see that it cannot cope with all types of multiple-meanings in different branches of knowledge by the help of a single monolingual corpus. It means that one word or combination of words in one type of corpus may appear more frequent than in the other type of corpus. So, to achieve more precise and satisfactory results it is better to use a separate type of monolingual corpus in the target language for searching ambiguous words frequencies of that type or similar types of text in the source language. Following this procedure, statistical data gained from the target language corpus will illustrate the real data by 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 Persian language. Our assumption is that searching different equivalents of every ambiguous word in a general corpus and further searching co-occurrences of these words in a single corpus seems an impossible work, since the number of times that the ambiguous words or their co-occurrences appear in a general corpus is not sufficiently large to be able to help us for calculating the probability of their occurrences although Dagan and Itai used a general corpus in their experiment (Dagan and Itai, 1994). The English word "old" may appear thousands times in a general corpus, but the noun phrase "the old Persian" in which the adjective "old" is translated into Persian different from that in, say, "the old friend" or "the old shoes" may appear a few times. Thus, to find the most suitable translation for the ambiguous word "old"[2] it is more logical to calculate its frequencies co-occurrences with the word "Persian" in a philology corpus or some similar fields in which this phrase has been appeared more frequently.

3. Statistical model

There are several statistical models to be able to resolve the ambiguity problem introduced in this study among which we can name Hidden Markov Model, Bayes law (Charniak, 1993), Probabilistic Model (Dagan and Itai, 1994) and Simulation Model (Shannon, 1975). The latter has been used for solving linguistic problems by this paper for the first time. Here, we use random number which is a device for simulation because it is more reasonable from the scientific point of view. When we select a translation by the help of Probabilistic Model (the highest probability) in the target corpus as the correct choice, the probability of choosing the alternative translations with the lower probability will be practically zero, while in a proportion of cases their choosing must be preferred.

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 target language are to be extracted from the monolingual corpus by the help of the algorithm designed for this purpose. Suppose our word has n different senses as $tw_1, ..., tw_n$. Now we get the frequency of each of these alternative senses as $f_1 ... , f_n$, and then calculate their probabilities from the following formula :

$$P(tw_i) = \frac{f_i}{\sum_{j=1}^{n} f_j} \qquad i = 1, \dots , n$$

Then we construct the related table as follows:

i	1 … n
P(twi)	P1 … P_n

Table 2. Empirical distribution table

Using empirical distribution table and some goodness of fit tests we can form the best statistical distribution for the observed sample (Phillips, 1972). For large samples ($n \geq 100$) K-square testing is very useful, however, for samples with a quantity smaller than 10 ($n \leq 10$) it seems that applying Cramer-Von Mises testing (Phillips, 1972) is more appropriate than any other one[3]. And almost all samples we concern with in this study are of this sort, namely, smaller than 10 different senses for each ambiguous word. So, here, we use Cramer-Von Mises testing to find the best statistical distribution for our samples. Suppose that the statistical distribution of the observed sample be as follows:

$$X \sim f_X (tw)$$

Now, using random numbers produced by this distribution[4] one of the n senses of the given word can be selected (Jansson, 1966). The algorithm of this model can be displayed as follows:

[2] It is ambiguous from the standpoint of its translation into Persian

[3] Curve fitting to find statistical distribution can be obtained by the help of certain statistical soft wares with a high degree of precision.

[4] 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.

```
1.start
2.Produce random number with Uniform distribution
3.Produce random number of the observed community   $X \sim f_X(tw)$
4.Determine word sense by Random number
5.end
```

To prevent undesirable effect of the random numbers on the ultimate results, it is necessary to repeat selection of random numbers and choosing one of the alternative senses from the corpus more and more (practically at least 30 times), and then that sense with the most frequency is chosen. To demonstrate how our algorithm and Simulation Model work, first of all we construct table 3 according to the counts gained from our sample English sentence and its corresponding Persian translation as follows:

source co-occurrence	alternative target co-occurrence	frequencies
provide-perception	1) *be dast dadan-dark* 2) *tahiyeh kardan-dark*	14 0
learning-process	1) *farayand-yadgiri* 2*) farayand-amuzesh* 3) *farayand-danesh*	240 20 0
primary-students	1) *daneshamuzan-ebtedaii* 2*) daneshamuzan-avaliye* 3) *daneshjuyan-avaliye* 4) *daneshjuyan-ebtedaii*	150 15 8 0

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

Now, we calculate the probability of the alternative translations of the ambiguous source co-occurrence "primary students" from the following formula:

$$P(tw_i) = \frac{f_i}{\sum_{j=1}^{n} f_j} \qquad i = 1, \dots, n$$

i	1	2	3	4
P (tw_i)	0.867052	0. 086705	0.046243	0

Table 4. Empirical distribution table

In this stage, using above empirical table and goodness of fit tests we calculate distribution of geometric probability with the parameter P = 0.867052

$$X \sim f_X(tw) = \begin{cases} pq^{tw-1} & tw = 1, 2, 3, 4, \dots , \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq p \leq 1$, $q = 1 - p$. Now, to produce random numbers from geometric distribution with the parameter P= 0. 867052, first we produce the following Bernoulli Random Variable (Sheldon, 1976) using uniform random numbers $U \sim U(0, 1)$, i = 1, 2, 3,... which can be extracted from the related statistical tables:

$$X_B = \begin{cases} 1 & U \leq 0.867052, \\ 0 & U \geq 0.867052. \end{cases}$$

In this stage we continue Bernoulli tests up to get desirable result $X_B = 1$. The number of tests indicates our selected word. The following table has been filled by the help of 66 produced random numbers in which N is the frequency of the selected sense of the ambiguous word:

i	tw_1	tw_2	tw_3	tw_4
N	54	11	1	0

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

4. Conclusion

It seems that the presented method for multiple-meanings disambiguation is similar to the experiment carried out by Dagan and Itai (1994), however, this method has several advantages whether in the side of linguistic model or in the side of statistical model. For one thing, it uses only lexical co-occurrences both in source language text and in target language corpus instead of syntactic tuples. So, we don't need any syntactic parser whether for the source or for the target language. The only need is a simple tagged corpus for the target language and a selected source language text to be analyzed.

To gain data from the target language corpus we used domain-specific monolingual corpus for every kind of text in the source language. In this way, the numbers of counts for any lexical co-occurrence in the target language corpus will reach a considerable rate to be worked with for carrying out statistical analysis of the experiment. Moreover, the results will be more precise and accurate. In this experiment we used Simulation Model, random numbers and goodness of fit in statistical part, while all the previous methods for word sense disambiguation used Probabilistic Model in their 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 the reality and more precise than when selection of the most probable case is used. Suppose we have two alternative Persian equivalents for the English word "saw". And when we search in the Persian monolingual corpus of the related field the frequencies 80 and 10 are gained for tw_1 and tw_2 respectively. We calculate the probabilities of each of these cases as 89 percent and 11 percent for tw_1 and tw_2 respectively. If we use the Probability Model, in all cases the tw_1 which is the most probable is selected and tw_2 is ignored. While we know that in 11 percents of cases we may have tw_2 as the correct equivalent for that English ambiguous word. However, consider the Simulation Model and using random number for selecting the best equivalent for our English word. Here, our algorithm which is based on Simulation Model selects tw_1 for 89 percents of cases and for the rest it selects tw_2 as a suitable target equivalent. In any case, the results gained from Simulation Model is more scientific and reasonable. The precision of the proposed model has been tested for a rather 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 704 ambiguous words considering the related Persian text. Thus, the precision of the model has been calculated as 79%.

This approach can be directly applied in the system of English-to-Persian machine translation. In this system the problem of multi-meaning and resolving ambiguities related to them is one of the major questions for which up to now there has been no answer to find. While the experimental results are very encouraging for the pair of English and Persian languages, the procedure also may be applied to the other pairs of languages as well.

References

Biehler, R. (1974). *Psychology applied to teaching*. Houghton Mifflin Company.

Dagan, I. and Itai, A. (1994). "Word sense disambiguation using a second language monolingual corpus." C*omputational Linguistics*, 10(4), 563-596.

De Rose, S.J. (1988). "Grammatical category disambiguation by statistical optimization." *Computational Linguistics*, 14(1):31-39.

Garside, R. et al. (1987). *The computational analysis of English: a corpus-based approach*. Longman Group UK Limited, London and New York.

Jansson, B. (1966). *Random numbers generators*. Almqvist & Wiksell, Stockholm.

Koehn, P. and Knight, K. (2000). "Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm." Information Science Institute, University of Southern California, Online.

Marshall, I. (1983). "Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus." *Computers and the Humanities*, 17(4):139-150.

Marchuk, Yu.N. (1988). "Machine translation in USSR." In: *Encyclopedia of Library and Information Science*, 44(9):183-194.

Miram, G. E. (1998). *Translation algorithms*. Kiev, "Twin inter".

Megerdoomian, K. (2000). *Persian computational morphology: a unification-based approach*. Computing Research Laboratory, New Mexico State University, New Mexico.

Mosavi miangah, T. (2000). "Ambiguity problem in English-Persian machine translation." In: *Problems of Language Theory and Translation Science*, Moscow Pedagogical University, 4:88-98.

Mosavi miangah, T. (2001). "Comparative analysis of noun phrase for MT (with reference to English and Persian)." In: *Problems of Language Theory and Translation Science*, Moscow Pedagogical University, 6:68-78.

Mosavi miangah, T. (2002). Problems of English-Persian machine translation. *Journal of Philology*, 3(12): 38-42.

Phillips, D.T. (1972). *Applied goodness of fit testing*. American Institute of Engineering, Atlanta, Ga.

Shannon, R.E. (1975). *Systems simulation: the art and science*. Prentice-Hall, Inc.

Sheldon, R. (1976). *A first course in probability*. Macmillan Publishing Co., Inc. New York.

Su, K. and Chang, J. (1990). "Some key issues in designing machine translation systems." *Machine Translation*, 5:265-300.