Journal of Computer and Communications, 2016, 4, 109-124
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2016.44010

Improving the Collocation Extraction Method Using an Untagged Corpus for Persian Word Sense Disambiguation

Noushin Riahi, Fatemeh Sedghi

Computer Engineering Department, Alzahra University, Tehran, Iran
Email: nriahi@alzahra.ac.ir, fatemeh.sedghi@gmail.com

Received 12 March 2016; accepted 19 April 2016; published 22 April 2016

Abstract

Word sense disambiguation is used in many natural language processing fields. One of the ways of disambiguation is the use of decision list algorithm which is a supervised method. Supervised methods are considered as the most accurate machine learning algorithms but they are strongly influenced by knowledge acquisition bottleneck which means that their efficiency depends on the size of the tagged training set, in which their preparation is difficult, time-consuming and costly. The proposed method in this article improves the efficiency of this algorithm where there is a small tagged training set. This method uses a statistical method for collocation extraction from a big untagged corpus. Thus, the more important collocations which are the features used for creation of learning hypotheses will be identified. Weighting the features improves the efficiency and accuracy of a decision list algorithm which has been trained with a small training corpus.

Keywords

Collocation Extraction, Word Sense Disambiguation, Untagged Corpus, Decision List

1. Introduction

There are some words in every language with multiple meanings and different applications that their meaning is determined based on the context in which they are placed. That is these words are vague words. Context can be a sentence or phrase. Disambiguation of the meaning of these words (WSD: Word Sense Disambiguation) is one of the research areas in the field of natural language processing and is used in Information Retrieval (IR), Machine Translation (MT), information extraction and documents classification.

Ambiguous words are divided into two categories in terms of distinction level meaning. This phenomenon is

How to cite this paper: Riahi, N. and Sedghi, F. (2016) Improving the Collocation Extraction Method Using an Untagged Corpus for Persian Word Sense Disambiguation. *Journal of Computer and Communications*, **4**, 109-124.
http://dx.doi.org/10.4236/jcc.2016.44010

called granularity. Various meanings of words have low distinction level and are called fine-grained. For example, it should be specified in machine translation that the word “discussion” must be translated to which of its equivalent in Persian according to its context. The meanings of homographs have high different level or are coarse-grained. For example, what does the word “شیر” (shir) mean in a sentence (shiras a dairy product which is milk, shir as a tool which is faucet or shir as an animal which is lion)? Most applications in the real world are dealing with coarse-grained level [1].

In the 1990s when machine learning approaches were raised, a great improvement in the area of disambiguation of the meaning of words was obtained. In this decade, supervised algorithms with optimal accuracy were provided which still have the best accuracy. Since the accuracy of these algorithms is generally related to manually tagged training data, knowledge acquisition bottleneck could be occurred in case of ambiguous words with no corresponding big tagged data or in terms of languages with no available semantic tagged corpus. There isn’t any large enough training corpus to cover the entire ambiguous words to train a supervised algorithm, even in languages such as English which was among the first target languages for making big manually labeled corpus. The ability of making such corpus is only a hypothesis because making such training data is time consuming and costly [2].

On the other hand, unsupervised algorithms do not need semantic tagged corpuses and therefore do not face knowledge acquisition bottleneck problem. However they do not have proper accuracy. The creation of a WSD system is not a goal in itself but they are needed as a tool to improve the efficiency of other practical applications such as information retrieval and machine translation. Therefore the accuracy of such systems can affect the whole system accuracy. Also, the system should not have knowledge acquisition bottleneck problem in order to be able to provide adequate coverage on all the ambiguous words in a Language.

A lot of research has been done to overcome the problem in the recent years. Methods such as semi-supervised training which uses corpus with and without tag at the same time or methods which used other linguistic tools such as dictionaries, thesaurus and ontology in the corpus are from this type.

Small tagged corpus with all ambiguous words coverage in a language is faster and less costly than a large one. The proposed method tries to upgrade the decision list algorithm which is a supervised algorithm with a relatively small tagged corpus and a large untagged one, so that the accuracy of supervised algorithm trained with a small tagged corpus gets close to corresponding supervised algorithm trained with larger untagged one.

2. Related Works

Collocations extraction usually takes place based on Association Measures (AM) usage on big corpuses. AM uses statistical data of words in corpus in order to identify collocations [3]. One of the best known AM has been suggested based on information theory which is known as Point-wise Mutual Information (PMI) or Association ratio. If we define the probability of collocation of two words of w_1 and w_2 at the distance of D from each other as the following:

$$p(w_1, w_2) = \frac{f^D(w_1, w_2)}{D \cdot N} \tag{1}$$

In which N is number of words in the corpus and f^D is frequency of collocation of two words in a widow with length of D in which the possibility decreases by increasing the length of window, because the possibility of randomness of this collocation increases. Now, if $P(w_i)$ represents the probability of separate occurrence of word w_i, then Association Measure is calculated as follows:

$$PMI = \log_2 \frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)} \tag{2}$$

This measure considers the ration of dependency of two words to their independence. Other versions of this measure have also been suggested. For example, [4] mentioned the reverse bias problem of this measure compared to frequency and proposed a method to reform it. This method is named Mutual Dependency (MD):

$$MD(w_1, w_2) = \log_2 \frac{P^2(w_1, w_2)}{P(w_1) \cdot P(w_2)} \tag{3}$$

Then they defined another measure named Mutual Dependency With frequency logarithm bias with the explanation that having self-frequency bias (not reverse) is useful in small amounts in statistical factors:

$$LFMD(w_1, w_2) = MD(w_1, w_2) + \log_2 P(w_1, w_2) \tag{4}$$

[5] introduced a measure called Pearson's X^2-test as follows:

$$X^2 = \frac{\left(f_{w_1w_2} - \frac{f_{w_1} \cdot f_{w_2}}{N}\right)^2}{\frac{f_{w_1} \cdot f_{w_2}}{N}} + \frac{\left(f_{w_1\overline{w_2}} - \frac{f_{w_1} \cdot f_{\overline{w_2}}}{N}\right)^2}{\frac{f_{w_1} \cdot f_{\overline{w_2}}}{N}} + \frac{\left(f_{\overline{w_1}w_2} - \frac{f_{\overline{w_1}} \cdot f_{w_2}}{N}\right)^2}{\frac{f_{\overline{w_1}} \cdot f_{w_2}}{N}} + \frac{\left(f_{\overline{w_1}\overline{w_2}} - \frac{f_{\overline{w_1}} \cdot f_{\overline{w_2}}}{N}\right)^2}{\frac{f_{\overline{w_1}} \cdot f_{\overline{w_2}}}{N}} \tag{5}$$

In which $f_{w_1w_2}$ is the real collocation frequency and $\frac{f_{w_1} \cdot f_{w_2}}{N}$ is the expected collocation frequency. This measure was in fact a suggestion for solving a problem named null hypothesis, according to this hypothesis simultaneous occurrence of two words together is not always indicative of their dependence but this collocation has taken place because of the chance and accident (For example the combination of "of the" and "in the"). X^2 measure has the ability to detect null collocation in this way that if its value is above threshold level, then it is the reason for the occurrence of null hypothesis.

Measure introduced in [6] is also another measure which can detect null hypothesis and it is called logarithmic probability rate and is defined as follows:

$$LLR(w_1, w_2) = -2\left(f_{w_1w_2} \cdot \log \frac{\frac{f_{w_1w_2}}{N}}{\frac{f_{w_1} \cdot f_{w_2}}{N}} + f_{w_1\overline{w_2}} \cdot \log \frac{\frac{f_{w_1\overline{w_2}}}{N}}{\frac{f_{w_1} \cdot f_{\overline{w_2}}}{N}} + f_{\overline{w_1}w_2} \cdot \log \frac{\frac{f_{\overline{w_1}w_2}}{N}}{\frac{f_{\overline{w_1}} \cdot f_{w_2}}{N}} + f_{\overline{w_1}\overline{w_2}} \cdot \log \frac{\frac{f_{\overline{w_1}\overline{w_2}}}{N}}{\frac{f_{\overline{w_1}} \cdot f_{\overline{w_2}}}{N}}\right) \tag{6}$$

Also [7] has purposed a fuzzy measure by considering this concept that most of the association measures are based on the ration of collocation number to unique occurrence number of the words and the fact that high levels of this ratio are vague and imprecise due to dependence to the size of corpus and occurrence of other words. There are many different measures 82 of which have been evaluated in [8]. In addition to this, Instead of using association measures directly, this has considered the issue of collocation extraction as a classification issue for the first time and has considered these 82 measures as training features to train classifier trainings. In a same method, [9] has used other features for classifier training by considering three association measures of PMI, X^2 and DICE. [10] used a different method using the idea of aligning words in equivalent sentences in parallel bilingual corpuses which has been raised in the field of machine translation in the [11] and has raised the algorithm of aligning words in monolingual corpus and extracted collocation by using it. This method extracts collocations better especially when they occur with longer distance compared to methods that only use the Association Measures. [12] has used a corpus for improving the collocation extraction where sentences are in a meaning dependency graph. Thus, collocation with have been repeated enough in one semantic relation have created a sample collocation bank in from of noun + verb. Then, for better coverage of collocations existing in one language, collocation in this bank which had a nominal role of morphological have been generalized by a semantic dictionary in order to cover words which are in their semantic category. [13] has considered the issue of the effect of corpus size on threshold of separating collocations form candidate compounds after comparison of Association Measures and has provided a method for automatic extraction of collocations which are independent of the corpus size using outlier data identification in Statistics.

3. Decision List Algorithm

Decision list algorithm was proposed for disambiguation of homograph by [14]. This supervised algorithm uses a semantic tagged corpus to simultaneously perform automated feature extraction and feature classification. Features are words about homograph word. In this algorithm, initially a list of all of the neighbor words of homograph word in form of one word before, one word after, two words before, two words after or a window to the size of $(-k, +k)$ which means k words before homograph and k words after it featured in the corpus are collected regardless of the specific meaning of the homograph word. A separate window can be considered for words which have be featured in $(-k, +k)$ window for greater accuracy which is usually a context in comparison with homograph word which the maximum occurrence has taken place compared to it. Then, for each of the

possible meanings of homograph word, for each of the neighboring words that have the same features, one possibility is calculated labeled using corpus as follows:

$$\text{log likelihood ratio} = \text{Abs}\left(\text{Log}\left(\frac{P(\text{Sense}_1 \mid \text{Collocation}_i)}{P(\text{Sense}_2 \mid \text{Collocation}_i)}\right)\right) \quad (7)$$

Each probability represents the relationship between each feature with one of the possible meanings of the homograph word. Then this probability is sorted from largest to smallest in a decision list and possibilities which are lower than a certain threshold are excluded. When a new test sample arrives, words in the decision list are searched one by one from top to bottom according to their corresponding window about the new word until an item is found, the search stops in this case and found feature class will be assigned to the new test sample.

Decision List algorithm is a collocation-based algorithm. This means that features are local and only words themselves are about target word. Hence, they do not need pre-processing to determine the grammatical tags and are applicable for languages in which accurate grammatical tagging is not available.

4. Collocation Extraction

As mentioned, decision list algorithm extracts and classifies desirable features by receiving tagged corpus and use of probabilities which it calculates for neighbor words. These features are a form of collocation words with the target word. Collocation words are defined as follows: words which have simultaneous occurrence frequency in text or speech are greater than being considered as accident.

Since the decision list depends on tagged training corpus for extraction of these collocations and this tagged corpus cannot be prepared in a large volume for all homograph words existing in a language, so some of the collocations existing for homograph in small training corpus are not identifiable.

Collocations are in different forms and their number of involved words and method of their combination are very different. Some collocations are rigid and some are flexible. For example, a flexible collocation such as “making” and “decision” can be seen in different forms such as “to make a decision”, “decision has been taken”, “a very important decision was taken”, etc. on the other hand a collocation such as “General Motors” can be seen in one from and it is a rigid collocation. It is clear that the meaning of collocation in some tasks such as WSD is wider and more flexible than the definitions which are in the field of collocation extraction because WSD methods usually show the fact that a special word could be in a collocation with one of the meanings of an ambiguous word and there is no mandatory in having separate and special meaning in a pair collocation such as “General Motors”. For example, the collocation related between “اشراف” (nobles) and “پادشاه” (king). “King” is a word which help to determine the meaning of ambiguous word of “nobles” while these two words are not in one collocation based on some definitions for collocation with have been mentioned above. [15] method has been used in this research from different methods of collocation extraction. One of the positive features of this method which have not been used in other extraction methods is considering the fact that collocations are not same in terms of flexible or rigid or some words that are separated by other words are considered as one collocation if they have certain order of compared to each other. In other words, even if other methods have the ability to simultaneously extract both types of flexible and rigid collocation, they cannot separate this two from each other or determine the distance and position of two collocation words with are separated by other words. For example, PMI is a method which can also extract collocation words with are with separated by spaces in addition to rigid collocations but cannot determine an output list with degree of flexibility of these collocations or the position of their occurrence relative to each other. As we will see later, this method determines degree of flexibility and hardness of candidate collocations and ranks them relative to each other in addition to ranking them.

5. Smadja Method

[15] extracted collocations in lexicography tool called Xtract. His method extracts collocations using statistical information of words in the corpus which are next to each other or are in the corpus with a short distance between them.

Initially, vocabulary relations of pair words are retrieved using only the statistical information. This phase is comparable with the work of [3] which evaluates a particular collocation between pairs of words. Based on [3] words can appear in any order relative to each other and can be in distance of one to several words or separated from each other. However, the statistics measures used by the Smadja provide more information and allow the

output to be more careful. These explicit statistical measures use an untagged semantic corpus to retrieve pair wise word relations which depend on each other in case on existing. These bi-gram words are retrieved if they have occurrence frequency higher than a certain threshold and according to the definition of hard collocations with was previously mentioned, words used in these bi-gram words are relatively hard.

Then the output of the first step goes in parallel to the next step. The first stage output pairs of words are used in the second step to create n-gram collocations. This step analyzes all sentences including pair words and distribution of words and POSs (Part-of-speech) of position surrendering these bi-gram words. It keeps words (or POS of words) which capture of a position with possibility higher than specific threshold. For example, bi-gram words of "average-industrial" create a bigger collocation of "the Dow Jones industrial average" because these words can always be seen in more difficult nominal terms in corpus. The corpus must have POS tags before this step and the third step.

Finally, retrieved bi-gram words in the first step will be filtered in the third step by combining the results of a parser and statistical methods. In this step, Xtract adds syntactic information to retrieved collocations of the first step and filters out inappropriate information. For example, if a bi-gram word contains a noun and a verb, this step recognizes it as a bi-gram subject-verb or a verb-object word and if they do not have such relation, they will be rejected.

6. The Proposed Method

The idea of the proposed method is that the frequency of some of the collocations is not enough to receive appropriate probability and be placed in the upper tier of the decision list used by training corpus due to small size but at the same time they are in a collocation in one of the meanings of homograph. Such collocations can improve efficiency if they are detected and are strengthened in the decision list with added weight. For example, consider this sentence: "کارشناسان عقیده دارند، قدرت اقتصادی اروپای شرقی از نفس افتاده است که به سبب انجام ندادن اصلاحات اقتصادی در لهستان است" (*Experts believe the Eastern Europe economic power has fallen sharply, because the economic reforms in Poland has not been doing*) a decision list with is from a not so great trained corpus accepts the proposed class of "است" collocation which has the most occurrence along with "نفس" among other collocations of the window and is in the upper part of the list. This collocation suggests the wrong class due to being in many sentences such as "یکی از معایب این کار بالا رفتن اعتماد به نفس کاذب است" (*One of the disadvantages of this work is to rise of Self Confidence, so much*). But the collocation of "فتاده(افتادن)" is in the lower tiers of the decision list due to small size of corpus and not due to not being in the collocation. Now if we had previously identified "افتادن" as a better collection than "است", we would be able to add a weight to its probability so that it can be higher in the decision list and identify the correct class. Identification of such collocations is possible by a big untagged corpus.

The first step of Xtract has been used in this method for identification of collocation words whether attached or adjacent to each other or spaced a few words from each other (which one of these pairs is the target homograph word). Smadja has called first step as extraction of important bi-gram words. According to (Smadja, 1993), there are strong evidence that most of the juxtaposition lexical relations are between words which are separated from each other by maximum 5 words. In other words, most of the lexical relations that are involved a word such as w can be retrieved by testing the neighbor of w which occurs in quintuple neighbor window (−5 and +5 around w).

Only statistical methods have been used in this step so that the related bi-gram words are identified. These methods are based on the assumption that if two word are collocation, then:

First: these two words must be observed with each other with a high frequency in a way that being observed together is beyond chance and accident.

Second: these two words must be relatively observed hard (uncompromising) together.

The word's distribution in the sentence has been analyzed by considering these two hypotheses and the used filter has been placed based on these hypotheses.

Initially a list of w_i words with data and information of collocation frequency of w and w_i is provided in which w is the homograph and w_i is candidate word for being in collocation with w. this list only contains frequency with collocation of w and w_i which has been frequency divided based on the position of it occurrence compared to w (the possible distance between two collocation words). **Figure 1** shows occurrence frequency of words

1. Extraction of collocations of an ambiguous word using three filters of Smadja method from a big untagged corpus (evaluating the improvement of Smadja method's filters in five different process)
2. Training a ordinary decision list using a small tagged corpus
3. Training a special decision list using the output of the first step with a tagged corpus
4. Weighting the special decision list for having priority when tagging a test sample

Figure 1. The proposed method steps.

occurred in window of (−5, +5) according to its place of occurrence in each of the window locations. Fri is the appearance frequency of w_i along with w in the corpus and p_j in which j is between −5 and +5 and nonzero is appearance frequency of w_i along with w which are spaced by j words. p_j shows histogram of appearance frequency of w_i along with w in the given position.

Then, more important bi-gram words will be extracted from the list by statistic measures describing connection strength of the words and amount of hardness of this connection. The first measure is power or strength:

$$\text{strength} = \frac{freq_i - \overline{f}}{\sigma} \tag{8}$$

In which $freq_i$ is simultaneous collocation frequency of w_i in window of (−5, +5), $\overline{f}$ is the mean frequency of all w is of w and σ is the standard deviation around $\overline{f}$. In fact, the strength shows the number of standard deviation higher than the mean of frequency of w and w_i bi-gram words.

The next measure is spread:

$$\text{Spread} = \frac{\sum_{j=1}^{10}\left(p_i^j - \overline{p_i}\right)^2}{10} \tag{9}$$

In which p_i^j occurrence frequency of the word w_i in position of j compared to w in which j can be numbers 1 to 10 based on window of (−5, +5) and $\overline{p_i}$ is the mean of p^js of a w_i. In fact, spread determines a variance of occurrence of a w_i in a window around w and histogram figure of p_i^j. If a spread is small, then histogram tends to smoothing which means w_i can be used equally in each position around w. instead, if the spread is large, histogram tends to having peak which means w_i is used only in one or several special positions.

Three following filters have been defined for filtering inappropriate w_is as well as optimal window for appropriate w_is:

$$\text{strength} = \frac{freq - \overline{f}}{\sigma} \geq k_0 \tag{10}$$

This proviso helps to remove pairs of words which do not have enough frequency. This proviso determines that appearance frequency of w_i in neighborhood of w must have at least one standard deviation higher than the mean which means that occurrence frequency with the target word must be higher compared to total other candidate words in the corpus. This thresholding eliminates a large number of lexical relations in most of the statistical distributions. For example, "دفاع" (defense) will be removed in **Table 1**.

$$\text{spread} \geq u_0 \tag{11}$$

This proviso will eliminate w_is which their distribution histogram in window around w is smoother and with fewer peaks than a certain limit. In fact, it accepts tougher and more uncompromising bi-gram words. The assumption here is that if two words are frequently used together in a syntactic structure, then we will have a feature pattern of collocation. This means that they will be seen in all positions and statuses with one equal chance. For example, "این" (this) will be removed in **Table 1**.

$$p_i^j \geq \overline{p_i} + \left(k_1 \times \sqrt{\text{spread}_i}\right) \tag{12}$$

This proviso is in a different way compared to two previous provisos. First two provisos eliminated w_is completely but this proviso is applied on w_is which has met the previous two provisos and has been identified as appropriate bi-gram words and eliminates the improper position of the window (+5, −5). The first and second

Table 1. The occurrence frequency of candidate words in collocation with "اشراف" in 10 positions around it in the corpus.

W	w_i	Freq	p_{-5}	p_{-4}	p_{-3}	p_{-2}	p_{-1}	p_{+1}	p_{+2}	p_{+3}	p_{+4}	p_{+5}
اشراف	بخش	12	2	1	4	0	1	0	0	0	4	0
اشراف	شهر	19	2	0	4	1	1	1	2	7	1	0
اشراف	اروپایی	12	0	2	1	1	0	6	1	1	0	0
اشراف	داشتن	493	26	21	16	13	10	227	87	40	38	15
اشراف	این	290	34	24	33	46	15	7	29	48	37	17
اشراف	زادگان	25	0	0	0	0	0	24	0	1	0	0
اشراف	در	478	67	64	68	56	3	45	46	48	62	19
اشراف	علمی	42	2	1	2	3	0	30	0	3	1	0
اشراف	کنترل	8	0	0	1	4	0	0	2	0	0	1
اشراف	از	419	47	28	55	56	80	13	27	45	59	9
اشراف	طبقه	25	3	1	5	1	13	1	0	1	0	0
اشراف	اعیان	32	0	2	1	27	0	0	0	1	0	1
اشراف	دفاع	3	0	0	1	1	0	1	0	0	0	0
اشراف	کامل	146	1	3	2	1	0	129	1	6	2	1

provisos delete output rows of the first step, but the third proviso selects the column from the remaining rows. One or more positions may be considered for each bi-gram words which are corresponding with the histogram peaks and so the result is selected in several p_i^j. For example, $p + 1$ column remains for "اشراف" (nobility or aware) and other positions will be removed.

u_0, k_0 and k_1 thresholds should be determined by tests and depend on the use made of the collocations. Generally, the lower threshold will accept more data and has higher recall and lower accuracy. Smadja which has used Xtract for automatic creation of over a ten-million-word corpus vocabulary has considered threshold values for k_0, u_0 and k_1 respectively 1 and 10 and 1. Our suggested method also uses three above measures for extraction of bi-gram words with space and without space which the values of each one will be described in the following.

Another point that should be considered is the importance of the distance that the extracted collocation has with the homographs. Decision list applies the rules of ±1, ±2 and ±K words to this aim in order to identify words which are only in collocation with homograph in case of occurrence with space and a particular position compared to homograph which means a word before and a word after, two words before and two words after and at the end words in a window with radius of K around the word. Two ±1 and ±2 rules are for harder collocations while collocations extracted by ±K rule are considered as soft collocations in a way that all collocations obtained from this rule are searched for in the window of ±K.

The idea that comes to mind in this regard is that can the use of that position (compared to the homograph) which has been the maximum occurrence of the collocation word as the accurate size of the window enhance the system efficiency? To evaluate it in **Table 2** for 6 homographs, we compared the decision list which considers the maximum frequency for ±k rule with window's common mode with fixed-length of 5. Evaluation measure has been F-Measure and evaluation has been carried out between two 1200 and 500 training corpuses using 5-fold-cross-validation method.

We can see that the utility of this window is not the same for the different homographs. Most utility has been for the word "اعمال" (impose or acts). While the efficiency for the word "گرم" (hot or gram) has fallen. The reason for this could be that the collocations of the homograph word "گرم" are softer and this type of windowing limits their scope of the search of reduces their efficiency. However, reduced efficiency percentage for words which have been affected by this method is lower than the percentage of increased efficiency for words that have benefited.

A decision list which only uses collocations suggested by Smadja Method is provided using training corpus

Table 2. Comparison of F-measure of the decision list in the window with size 5 and window with maximum frequency.

homograph	Training corpus with size 1200						Training corpus with size 500					
Windowing method	نفس	گرم	حسن	دور	اشراف	اعمال	نفس	گرم	حسن	دور	اشراف	اعمال
Window with size 5	0.8914	0.9268	0.9322	0.9072	0.8680	0.8118	0.8512	0.8738	0.9171	0.8606	0.8197	0.8063
Window with size max frequency	0.8936	0.9254	0.9311	0.9212	0.8565	0.8643	0.8501	0.8266	0.9205	0.9075	0.8118	0.8451

after the extraction of collocation. It is clear that this list does not need to use ±1, ±2 and ±*k* rules because size of the window that surrounding the homograph for learning the class of each collocation is determined. Such classification which we call it "special decision list" has a higher accuracy than ordinary decision lists but it has lower recall, because it has learned the learning model with fewer and more accurate collocations.

To overcome the problem of declining recall in special decision list, an ordinary decision list learned on the training corpus. Then the special decision list is used for tagging in tagging step for test samples and if the suggested special decision list did not have any tag, the class suggested by the ordinary decision list is used, or using another method the determined degree of validity of the class is multiplied by a weight and is compared with validity of class determined by the ordinary list and the more valid class is determined.

Evaluating Different Tested Processes to Improve Collocation Extraction Using Xtract

For more compliance of the Smadja Method with the goal of word sense disambiguation and improving the performance of decision list, second and third filters of Smadja method have been focused on and the proposed method created changes in them. Since the results of these changes will be evaluated in the next section, these changes will be classified and introduced in this section. The first six processes will be done in the first step of **Figure 1** and are directly involved in extraction of determination of position of occurrence and collocations of a homograph. The seventh process will be done in the fourth step of **Figure 1** and examines the method of effectiveness of these collocations on the output algorithm of the decision list. In these processes, an attempt is made to consider the fact that the raw corpus which is used for the extraction of collocations may or may not have been POS tagged and therefore improvements have been suggested in both cases.

This method extracted and weighted pairs of word with different processes. Comparison of the results of these various processes has been mentioned in experiments section of the article. Each of these processes has been explained in continuation of this section:

The first process: this process does not make any changes to the first step of Smadja. Initially we formed a matrix for each homograph which is similar to **Table 1** in which each row is a candidate collocation for homograph and columns are frequency of occurrence and number of occurrence in each decuple position of (+5, −5) window around the homograph. This matrix is also used in other processes. Two measures of strength and spread and three mentioned filters have been used in this process. Due to differences of size in our experiments and also different definition and goal that the proposed method has for collocation extraction, it is clear that thresholds must change in this task.

First, the bi-gram words that strength measure is small for them will be ineffective on their own due to not being in the decision list at all. Thus out strength measure threshold which is k0 can be lower. In other words, reducing strength measure does not significantly affect the accuracy but it increases the recall. Also as mentioned in the Smadja article, collocations in the ambiguity of words can be softer. Smadja required hard collocations based on the definition it had for collocation in order to automatically create vocabulary.

Hence two words that were repeatedly mentioned together in the text, such as "اشراف" (nobility) and "پادشاهان" (kings) just due to being in a same field has a high strength with a low spread measure because they can be seen anywhere relative to each other in the sentence, so Xtract were considered high spread threshold. But in WSD such collocations are good and lower spread threshold can be considered.

The second process: the thing which was done in the first process for reducing spread threshold is not useful for all candidate collocations that have different morphological tags. Basically, the words which are considered as collocation in a WSD system are not verbs and letters (article, prepositions, pronouns, etc.). In other words, a verb which is mentioned usually in all steps of homograph forms bi-gram words with lower possibility compared

to a verb which has occurrence peak from one to three words after homograph. For example, in this sentence: " امروز میبینیم اعتماد و اتکاي به نفس داشته و قدم بعدي را در خلق ابتکار هاي بزرگتر به پیش ببرند. " (*Today we see that they have Self Confidence and can creating larger initiatives*) the verb "دیدن" is not a pair with the word "نفس" but the verb "داشتن" is a bi-gram words. In general, "داشتن" has three peaks exactly in +1 and +2, and +3 windows. It is also about letters, in a way that collocation of a letter with homograph is only created in case of it appearing as completely hard or inflexible with it. For example, this sentence " با ورزش دانشآموز علاوه براعتماد به نفس خویشتنداري مثبت را نیز فرا میگیرد" (*Students can learn Self Confidence and continence with exercise*) only collocation of "به" which is located exactly in position of -1 compared to "نفس" creates a bi-gram words. In the same sentence, letters "بر", "را" and "نیز" which are frequently used in each corpus and with every ratio with both meanings of "نفس" are not appropriate words. The word "به" itself should be searched for only in −1 window that this proviso is applicable by the third filter.

Thus this process has been considered for two types of morphological verbs and higher and more stringent threshold words and other lower threshold forms in spread measures but the tests have shown the filter is necessarily required for other morphological types because the peak of spread measure of the word is a sign of its desirability as a bi-gram word.

The third and fourth processes: the method of comparing spread measure in this way that if w_i word has a low repeat in a corpus, even if all the times it is mentioned in the window around w is a fixed position compared to it, it will not receive a large spread. For example, consider two given wm and w_n words which have frequency of distribution similar to **Table 3**.

Then we have for each: $\text{Spread}_m = 20.25$, $\text{Spread}_n = 30.29$.

We see that although w_m can clearly be a bi-gram good words but due to lack of its repetition in the corpus (which leads to a lack of collocation frequency with the homograph), it has lower spread compared to w_n. Although we determine the value of spread threshold to the extent that we do not lose these type of words but we will cause the lack of filter for probably useless words like w_n that are present just because the high frequency of occurrence in the body and not because of having a good spread. Especially if these words are letters which have a high frequency of occurrence.

A solution is that we consider the threshold of spread measure related to collocation frequency of the word which means that a word with a high collocation frequency must face more astringency when being filtered for the histogram spread (second filter). Thus, we changed the second filter in the third process as follows:

$$\text{spread}_i > u_0 \times freq_i, \tag{13}$$

and in the forth process as follows:

$$\text{spread}_i > u_0 \times \left(freq_i \wedge 2 \right), \tag{14}$$

changed and compared their results. It is clear that there is more astringency on the high-frequency words in the forth process compared to the third process.

The fifth process: inflectional form of words is not considered in third and fourth processes and filters are same for all words. Thus, the second filter is determined based on POS tag of the word in the fifth process. By testing different types and different thresholds, the most appropriate obtained filter is as follows: Equation (14) for verbs and prepositions and equation 13 for other words.

This means that this will be applied harder on verbs and prepositions of spread measure filter compared to higher collocation frequency.

The sixth process: this process evaluated the third filter; this means the determination of method for window of each word in bi-gram words. As previously mentioned:

$$p_i^j \geq \overline{p_i} + \left(k_1 \times \sqrt{\text{spread}_i} \right) \tag{15}$$

Table 3. Frequency of distribution of w_m and w_n sample words in 10 positions around the ambiguous word.

	$freq_i$	f^{-5}	f^{-4}	f^{-3}	f^{-2}	f^{-1}	f^{+1}	f^{+2}	f^{+3}	f^{+4}	f^{+5}
w_m	15	0	0	0	0	0	15	0	0	0	0
w_n	291	20	34	32	19	33	35	29	25	31	33

This filter makes windows smaller for bi-gram words which have been mentioned in previous step. This filter removes positions in (−5, +5) window in which word does not have peak based on spread measure of each word. For example consider this sentence:"رسانهها تایید کردند که او در نشست دیروز اشراف و تسلط کافی را داشته است" (*Some news had confirmed that he had enough mastery in yesterday's meeting*) this filter has already been determined that collocation of "داشتن" must be searched in (0, +5) window around the homograph and it being used out of (0, +5) positions especially before the homograph is not able to allocate class. This feature is helpful in decision list, because as evaluated at the beginning of this section, if maximum measure for each word compared to homograph is used for determination of its window, this method can provide several appropriate positions instead of one position for limit range of the window and expend the window by replacing this method with the maximum occurrence. For example, it can determine a wider (0, +5) window according to peaks of spread histogram instead of limited window of (0, +3).

In the sixth process in addition comparing to testing different thresholds for constant k1 value, Maximum mode (MAX) which is used in the ordinary decision list and has a position in the list which has the highest occurrence frequency has also be compared. It is clear that higher values of k1 will remove more positions and more astringency will be applied.

The seventh process: this process is done on fourth step of **Figure 1**. This means that this will return to the decision list and will test weighting processes. The degree of validity of the proposed class (the logarithm of the probability of the collocation determining proposed class) of special decision list can be multiplied by a weight and then it can be compared with the degree of validity of the proposed class of the ordinary decision list so that the one with higher validity determines the final class. Three different modes have been test for weighting:

1) No weighting be done and instead the class proposed by special decision list always be used and the ordinary decision list be used for tagging only when this list did not have any suggestions in which the ordinary decision list have higher covering measure due to not selecting collocations and thus, if the special decision list does not find a collocation in the field of test sample, the ordinary list may be able to find a collocation.

2) Weighting with constant numbers. Arguably the best weight for all homographs is not unique. Some of them have the best efficiency with lower weightings and some with higher weightings. It can be said that the words which have better spread measure will have the maximum efficiency in higher weighting of the special decision list.

3) Making the assigned weight to correspond the extracted bi-gram words with calculated spread for it:

$$a = \frac{\text{spread}_i - \overline{\text{spread}}}{\sigma} \tag{16}$$

and

$$\text{Weight} = w + a \tag{17}$$

In which w is a constant number and spread_i is spread of i bi-gram words which have been extracted in one process and $\overline{\text{spread}}$ is the mean spread of all extracted pairs of words in that process and σ is its variance. A variable is calculated for each extracted bi-gram words and acts as a weight during classification for its bi-gram words and adds a weight to constant weight of w. here a word with higher peak according to the mean and expansion variance of other pairs of words with receive higher weight.

7. Experiments

Corpus used in the experiments is Hamshahri corpus which is in Persian and is a set of Hamshahri Newspaper's text between 1996 and 2007 and there is no tagging or rooting on it [16].

We considered six "اعمال" (impose or acts), "دور" (round or far), "حسن" (Hasan or goodness), "گرم" (hot or gram), "اشراف" (nobility or aware) and "نفس" (breath or self) homograph words for the experiment. Initially, we extracted all of the occurrences of these words with five words before and five words after them. Semantic tag was given to each homograph for 1200 occurrence and one semantic tagging corpus was created. Then 5000 occurrences in each homograph were inflectionally tagged and rooted along with surrounding words in order to form a corpus without semantic tagging. Rooting was just limited to changing verbs into infinitive and plural nouns to singular and grammar tagging for dividing to three classes of verbs, letters and others (nouns, adjectives, etc.).

We used the standard method of 5-fold-cross-validation for evaluation. Extraction of features was carried out in the same 5000 samples without semantic tag. Evaluation measure is also F-measure which is a combined measure of accuracy and recall. The reason for selecting this measure is that basically, accuracy and recall alone cannot demonstrate the efficiency and if we aim to improve one of these measures in determination of thresholds, then we will directed to a way where the other is decreased. Thus the best measure to determine the threshold is the F-measure.

It is necessary to mention that tests done to achieve the thresholds in each process showed that better results can be achieved if we find an optimal threshold for each homograph word. But since this act reduces the efficiency of the method in reality, we used it for general threshold.

Table 4 has summarized the results for six processes from seven described processes. Thresholds of each process have been mentioned in the table which have been obtained using different experiments and the results of the most appropriate threshold have been shown, except for the sixth process which has been mentioned in the rows, other processes' results have been separated in columns. k_0 threshold in Equation (10) is zero everywhere. Weighting the extracted bi-gram words in the results of this table have been carried out according to the first mode of seventh process.

By comparing the results for different k1s in Equation (12), one general threshold for each six words can be determined. First we examined $k_1 = 2.5$. This threshold for a window with size 5 is too high so that no occurrence passed through it in practice. Lack of changes compared to the ordinary decision list indicates this fact. Although $k_1 = 2.5$ is the best threshold for "اعمال"(impose or acts), but it cannot be considered as a general and common threshold for all six words.

But to select an appropriate value among four modes $k_1 = 1$, $k_1 = 1.5$, $k_1 = 2$ and Max, the numbers of times in which they have achieved the maximum values in each process can be considered. **Table 5** shows the number of maximum F-measures for each process for different k_1s.

The results obtained from **Table 5** shows that the threshold of $k_1 = 2$ is generally more appropriate because generally, threshold of 2 has 17 case of maximum F-measure in each process while this value has been respectively 6, 7 and 0 for 1, 1.5 and Max thresholds. Also except from process 4 which has not shown a special improvement to on certain k_1, other processes had the maximum peak in $k_1 = 2$, thus we carried out the experiments of seventh process with this threshold.

Table 4 also highlights other points. Generally, process 2 should not achieve lower efficiency compared to process 1. Also the efficiency of process 5 should not be lower than processes 3 and 4 and these processes are expected to have equal efficiency in worst condition where thresholds are not separated in terms of grammatical tag, but it should be noted that considered u0 thresholds are common and general for all six words and were selected in a way that they make the result of all six word optimal. Now if we consider $k_1 = 2$ we will observe that a word such as "اعمال" had led to loss of efficiency in 1, 2 and 3 processes. The considered optimal threshold for this word had more difference compared to general threshold and this was more intense in the second process in a way that the second process has achieved lower results compared to the first process despite grammar isolation for threshold determination. This is also applied for "حسن" in 4 and 5 processes and for "اشراف" in 1 and 2 processes. If there has been the ability for separate determination of optimal threshold for each word then the results for separation of grammar has always been improving and grammatical separation of threshold also reflects that the results will be better. One reason for the better results in processes 3, 4 and 5 is optimal thresholds of different words being closer in these processes in a way that determination of general threshold is less harmful to the optimal threshold results for each word.

Experiments have also been done for seventh process and improvement percentage in each form of weighting has been shown in **Table 6**. The number of maximum F-measure for each form of weighting in each process has been shown in **Table 7**. In cases where the maximum value has been repeated in two different modes, these have been considered in **Table 7** for both forms.

Evaluating **Table 7** shows that one form of weighting is not appropriate for all processes. Form 2 with weight equal to 2.5 has generally been the best form in 1, 2 and 3 processes, but form 3 of weighting performed better in 4 and 5 processes, because making the weight assigned to a collocation based on the measure that defines its usefulness—spread measure—should be a more reasonable weighting than what we have in form 2. Because a similar weight is considered for all occurrences in form 2 without considering special utility measure, but since our utility measure is spread and according to what was already clear, the spread measure has been reclaimed in 4 and 5 processes, confidence in this measure has only led to improved results in processes 4 and 5.

Table 4. Values of Calculated F-measure for First Six Processes with Weighing in Mode 1 for Different k_1 Thresholds.

Method and thersholding for windowing	Ordinary decision list	Process 1 $u_0 = 3$	Process 2 (verbs $u_0 = 500$ letters $u_0 = 700$ others $u_0 = 3$)	Process 3 $u_0 = 0.2$	Process 4 $u_0 = 0.02$	Process 5 (verbs $u_0 = 0.01$ letters $u_0 = 0.02$ others $u_0 = 0.2$)
Homograph						
اعمال (impose or acts)	0.8643					
$k_1 = 1$		0.8411	0.8312	0.8429	0.8677	0.8640
$k_1 = 1.5$		0.8526	0.8432	0.8526	0.8694	0.8601
$k_1 = 2$		0.8525	0.8474	0.8508	0.8710	0.8660
$k_1 = 2.5$		0.8643	0.8627	0.8694	0.8710	0.8769
$k_1 = 3$		0.8643	0.8643	0.8643	0.8643	0.8643
Max		0.8283	0.8116	0.8333	0.8682	0.8568
دور (round or far)	0.9212					
$k_1 = 1$		0.9016	0.9107	0.9041	0.9359	0.9309
$k_1 = 1.5$		0.9150	0.9195	0.9158	0.9338	0.9359
$k_1 = 2$		0.9250	0.9288	0.9259	0.9355	0.9380
$k_1 = 2.5$		0.9212	0.9212	0.9212	0.9212	0.9212
Max		0.9042	0.9124	0.9100	0.9338	0.9326
حسن (Hasan or goodness)	0.9311					
$k_1 = 1$		0.9219	0.9303	0.9265	0.9397	0.9336
$k_1 = 1.5$		0.9285	0.9277	0.9284	0.9379	0.9339
$k_1 = 2$		0.9304	0.9363	0.9329	0.9379	0.9363
$k_1 = 2.5$		0.9311	0.9311	0.9311	0.9311	0.9311
Max		0.9266	0.9279	0.9261	0.9385	0.9336
گرم (hot or gram)	0.9254					
$k_1 = 1$		0.9146	0.9235	0.9181	0.9199	0.9384
$k_1 = 1.5$		0.9158	0.9239	0.9166	0.9259	0.9367
$k_1 = 2$		0.9270	0.9378	0.9305	0.9242	0.9447
$k_1 = 2.5$		0.9254	0.9254	0.9254	0.9254	0.9254
Max		0.9155	0.9320	0.9137	0.9256	0.9418
اشراف (nobility or aware)	0.8565					
$k_1 = 1$		0.8602	0.8474	0.8759	0.8556	0.8780
$k_1 = 1.5$		0.8613	0.8540	0.8682	0.8548	0.8788
$k_1 = 2$		0.8614	0.8457	0.8623	0.8576	0.8723
$k_1 = 2.5$		0.8565	0.8565	0.8565	0.8565	0.8565
Max		0.8571	0.8436	0.8623	0.8485	0.8709
نفس (breath or self)	0.8936					
$k_1 = 1$		0.9037	0.9118	0.9045	0.9084	0.9187
$k_1 = 1.5$		0.9038	0.9100	0.9021	0.9085	0.9181
$k_1 = 2$		0.9034	0.9071	0.9026	0.9051	0.9093
$k_1 = 2.5$		0.8936	0.8936	0.8936	0.8936	0.8936
Max		0.8985	0.9026	0.8960	0.9076	0.9152

Table 5. The number of maximum F-measures for each process for different k_1s.

	Process 1	Process 2	Process 3	Process 4	Process 5	total
$k_1 = 1$	0	1	2	2	1	6
$k_1 = 1.5$	2	1	1	2	1	7
$k_1 = 2$	4	4	3	2	4	17
Max	0	0	0	0	0	0

Table 6. Percentage of iimprovement obtained from different processes compared to the ordinary decision list for 1200 and 500 corpuses.

	Training corpus with size 1200					Training corpus with size 500				
Method and thersholding for windowing	Process 1 $u_0 = 3$	Process 2 (verbs $u_0 = 500$ letters $u_0 = 700$ others $u_0 = 3$)	Process 3 $u_0 = 0.2$	Process 4 $u_0 = 0.02$	Process 5 (verbs $u_0 = 0.01$ letters $u_0 = 0.02$ others $u_0 = 0.2$)	Process 1 $u_0 = 3$	Process 2 (verbs $u_0 = 500$ letters $u_0 = 7000$ others $u_0 = 3$)	Process 3 $u_0 = 0.2$	Process 4 $u_0 = 0.02$	Process 5 (verbs $u_0 = 0.01$ letters $u_0 = 0.02$ others $u_0 = 0.2$)
Homograph										
اعمال										
Form 1	−1.18%	−1.69%	−1.35%	0.67%	0.17%	1.63%	1.83%	1.63%	2.45%	2.24%
Form 2	−0.93%	−1.69%	−1.01%	0.25%	0.25%	1.43%	1.63%	1.43%	2.24%	2.24%
Form 3	−1.1%	−1.6%	−1.27%	**0.84%**	**0.09%**	1.63%	1.83%	1.63%	**2.65%**	**2.04%**
دور										
Form 1	0.38%	0.76%	0.47%	1.43%	1.68%	−1.92%	−1.51%	−1.31%	2.15%	1.14%
Form 2	0.72%	1.01%	0.97%	1.26%	1.59%	−1.51%	−1.31%	−0.7%	1.94%	1.13%
Form 3	0.55%	0.84%	0.63%	**1.68%**	**1.85%**	−1.92%	−1.51%	−1.31%	**2.35%**	**1.14%**
حسن										
Form 1	−0.07%	0.52%	0.18%	0.68%	0.52%	0.1%	0.11%	−0.1%	0.61%	0.11%
Form 2	−0.16%	0.44%	0.01%	0.6%	0.44%	0.31%	0.11%	0.1%	0.61%	−0.09%
Form 3	−0.07%	0.52%	0.09%	**0.68%**	**0.43%**	0.1%	0.11%	0.1%	**0.61%**	**0.52%**
گرم										
Form 1	0.16%	1.24%	0.51%	−0.12%	1.93%	8.84%	10.15%	8.94%	8.55%	10.97%
Form 2	0.43%	1.33%	0.6%	0.31%	1.76%	9.05%	9.34%	8.94%	8.15%	9.35%
Form 3	0.16%	1.24%	0.51%	**0.49%**	**2.18%**	8.84%	10.15%	8.94%	**8.37%**	**10.97%**
اشراف										
Form 1	0.49%	−1.08%	0.58%	0.11%	1.58%	0.73%	−0.46%	1.01%	1.84%	2.65%
Form 2	0.58%	0.79%	0.58%	0.74%	1.58%	0.73%	1.01%	1.01%	1.62%	2.44%
Form 3	0.49%	−1.05	0.76%	**1.07%**	**1.79%**	0.73%	−0.46%	1.01%	**2.25%**	**2.86%**
نفس										
Form 1	0.98%	1.35%	0.9%	1.15%	1.57%	1.15%	1.14%	1.15%	2.05%	3%
Form 2	1.07%	1.35%	1.23%	1.32%	1.74%	0.73%	1.15%	0.73%	1.84%	2.37%
Form 3	0.98%	1.18%	1.06%	**1.15%**	**1.74%**	1.15%	1.14%	1.15%	**2.05%**	**3%**

In general it can be said that when a big POS tagged corpus is available, the fifth process and otherwise the fourth process is more appropriate solution.

Experiments Related to Change Corpus Size

The next experiment tests the effect of improving the weighting of extracted collocations from a different size of a semantic tagged corpus. Results of improvement related to the corpus with 500 training sample have been shown in **Table 6**. Generally, **Table 6** shows that if there is the possibility to tag POS on the untagged semantic corpus which is used for extraction of collocations, process 5 in the third form and if there is no possibility of tagging POS then process 4 in the third form will have the best results improvement. Two different sizes for training corpus will be obtained from comparison of results improvement percentage in which the purposed method in conditions will have more tangible impact when semantic tagged corpus is smaller.

In another experiment the size of corpus is considered to be even smaller in order to evaluate the effect of purposed method on this small size. **Table 8** shows the results obtained for corpus with 50 samples. The size of

Table 7. The number of maximum f-measure for each form of weighting in each process.

	Process 1	Process 2	Process 3	Process 4	Process 5
Form 1	1	2	1	1	1
Form 2 with weight 2.5	4	5	4	1	2
Form 2 with weight 6	2	2	2	1	1
Form 3with weight 0.5	1	1	1	5	4

Table 8. Improvement percentage obtained from different processes compared to the ordinary decision list for corpus with 50 samples.

	Training corpus with size 50				
Method and thersholding for windowing	Process 1 $u_0 = 3$	Process 2 (verbs $u_0 = 500$ letters $u_0 = 700$ others $u_0 = 3$)	Process 3 $u_0 = 0.2$	Process 4 $u_0 = 0.02$	Process 5 (verbs $u_0 = 0.01$ letters $u_0 = 0.02$ others $u_0 = 0.2$)
homograph					
اعمال					
Form 1	1%	1.01%	1%	0.53%	1%
Form 2	1%	1.01%	1%	0.53%	1%
Form 3	1%	1.01%	1%	0.53%	1%
دور					
Form 1	2.3%	2.55%	2.08%	2.17%	3.42%
Form 2	2.3%	2.55%	2.08%	1.95%	3.21%
Form 3	2.3%	2.55%	2.08%	2.17%	3.42%
حسن					
Form 1	2.47%	3.56%	2.4%	2.23%	3.71%
Form 2	2.02%	3.11%	1.95%	1.78%	3.27%
Form 3	2.47%	3.56%	2.4%	2.23%	3.71%
گرم					
Form 1	1.08%	0.99%	1.08%	0%	0.61%
Form 2	1.08%	0.99%	1.08%	0.69%	1.06%
Form 3	1.08%	0.99%	1.08%	0%	0.61%
اشراف					
Form 1	−0.01%	−0.21%	−0.23%	0%	1.71%
Form 2	0.62%	0.42%	0.41%	0%	1.71%
Form 3	−0.01%	0.42%	−0.23%	0%	1.71%
نفس					
Form 1	1.9%	1.9%	1.9%	1.55%	2.21%
Form 2	1.68%	1.68%	1.68%	0.65%	1.32%
Form 3	1.9%	1.9%	1.9%	1.55%	2.21%

50 for training corpus leads to smaller training decision list and it is clear that collocations learned with this corpus have a little diversity. Thus it is expected that a lot of collocation detected by collocation extraction not be mentioned in the decision list and not be able to have effectiveness. Therefore the improvement is lower than results in **Table 6**. As it is clear, the fifth process still has better results but this does not apply in the case of the fourth process especially for "اعمال", "اشراف" and "گرم". The reason for this behavior can be due to more astringency of the fourth process in the extraction of collocations which leads to providing fewer collocations. This low number has shown its adverse effect on this small corpus which has a low diversity in learning collocations. This occurred while the fifth process does not have to extract a low number of collocations due to grammatically separated words. Therefore fifth process in **Table 8** almost in all homographs has the best improvement.

Finally, it is worth recalling that untagged corpus used in these experiments included 5000 occurrences from each of the ambiguous words and better improvements are expected if they become bigger. However, by changing the size of corpus, the general optimal threshold (u_0, k_0) also can change.

8. Conclusions

This article has focused on the subject of adverse impact of small size of semantic tagged corpus to remove the ambiguity of the meaning of homograph words in supervised methods. The amount of tagged data required in supervised methods in word sense disambiguation is much more than other tasks related to the field of machine learning. This is due to the frequency of homograph words in natural languages and needs of ambiguity removal methods for training separate classifications for each homograph. Thus the proposed method tries to improve the supervised algorithm using an untagged corpus.

Since collocations to homograph words in a text are considered as the most important features used in classifications, the small size of the corpus can reduce the performance of a disambiguation method. Smadja has a statistical approach for extraction of collections from an untagged corpus. The approach in this article revised and assessed the Smadja method in different processes and weighted the collocations in a supervised algorithm decision list. This weighting has been based on the fact that collocations resulted from a big untagged corpus is more valid than collocations which have been extracted by a decision list which depend on a small tagged corpus. The results of the evaluation for six different homographs have shown an improvement in the purposed method in a way that the improvements in the different homographs and processes have been from 1 to 3 percent.

References

[1] Ide, N. and Wilks, Y. (2006) Making Sense about Sense. In: Agirre, E. and Edmonds, P., Eds., *Word Sense Disambiguation*: *Algorithms and Applications*, Springer, New York, 47-73. http://dx.doi.org/10.1007/978-1-4020-4809-8_3

[2] Edmonds, P. (2000) Designing a Task for SENSEVAL-2. Tech Note. University of Brighton, Brighton.

[3] Church, K.W. and Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, **16**, 22-29.

[4] Thanopoulos, A., Fakkotakis, N. and Kokkinakis, G. (2002) Comparative Evaluation of Collocation Extraction Metrics. *Proceedings of the* 3*rd International Conference on Language Resources and Evaluation* (*LREC* 2002), Las Palmas, May 2002, 620-625.

[5] Manning, C.D. and Schütze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

[6] Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**, 61-74.

[7] Das, B. (2012) Extracting Collocations from Bengali Text Corpus. 2*nd International Conference on Computer*, *Communication*, *Control and Information Technology*, Kuching, 25-26 February 2012, Vol. 4, 325-329.

[8] Pecina, P and Schlesinger, P. (2006) Combining Association Measures for Collocation Extraction. *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, Association for Computational Linguistics, Sydney, 651-658. http://dx.doi.org/10.3115/1273073.1273157

[9] Karan, M., Šnajder, J. and Bašić, B.D. (2012) Evaluation of Classification Algorithms and Features for Collocation Extraction in Croation. LREC, 657-662. http://www.informatik.uni-trier.de/~ley/db/conf/lrec/lrec2012.html#KaranSB12

[10] Liu, Z, Wang, H, Wu, H and Li, S. (2009) Collocation Extraction Using Monolingual Word Alignment Method. *Proceedings of the* 2009 *Conference on Empirical Methods in Natural Language Processing*, **2**, 487-495. Association for

Computational Linguistics. http://dx.doi.org/10.3115/1699571.1699575

[11] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, **19**, 263-311.

[12] Shijun, L., Yanqiu, S., Lijuan, Z. and Yu, D. (2015) Construction of Semantic Collocation Bank Based on Semantic Dependency Parsing. 29*th Pacific Asia Conference on Language* (*PACLIC* 29), Shanghai, 30 October-1 November 2015, Information and Computation*:* Posters, 232-240.

[13] Suárez, O.S., Sánchez-Berriel, I., Aguiar, J.P. and Rodríguez, V.G. (2015) Outlier Detection in Automatic Collocation Extraction. *Procedia—Social and Behavioral Sciences*, **198**, 433-441. http://dx.doi.org/10.1016/j.sbspro.2015.07.463

[14] Yarowsky, D. (1994) Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. *Proceedings of the* 32*nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 88-95. http://dx.doi.org/10.3115/981732.981745

[15] Smadja, F. (1993) Retrieving Collocations from Text: Xtract. *Computational Linguistics*, **19**, 143-177.

[16] AleAhmad, A., Amiri, H., Rahgozar, M. and Oroumchian, F. (2009) Hamshahri: A Standard Persian Text Collection. *Journal of Knowledge-Based Systems*, **22**, 382-387. http://dx.doi.org/10.1016/j.knosys.2009.05.002