Mapping Persian Words to WordNet Synsets

Rahim Dehkharghani[1], Mehrnoush Shamsfard[2]

NLP Research Laboratory, Faculty of Electrical and Computer Engineering,
Shahid Behehshti University, Tehran, Iran.

***Abstract* —Lexical ontologies are one of the main resources for developing natural language processing and semantic web applications. Mapping lexical ontologies of different languages is very important for inter-lingual tasks. On the other hand mapping approaches can be implied to build lexical ontologies for a new language based on pre-existing resources of other languages. In this paper we propose a semantic approach for mapping Persian words to Princeton WordNet Synsets. As there is no lexical ontology for Persian, our approach helps not only in building one for this language but also enables semantic web applications on Persian documents. To do the mapping, we calculate the similarity of Persian words and English synsets using their features such as super-classes and subclasses, domain and related words. Our approach is an improvement of an existing one applying in a new domain, which increases the recall noticeably.**

Keywords— Lexical Ontology, Semantic Lexicon, Princeton WordNet, Automatic Mapping.

I. INTRODUCTION

ONTOLOGY is defined as a formal, explicit specifications of a shared conceptualization [1]. In fact, an ontology assembles a shared lexicon for researchers of a specific domain indicating the concepts, relations and rules of domain. Lexical ontologies are ontologies whose concepts are lexicalized in a specific language and has special linguistic relations. Lexical ontologies sometimes called as semantic lexicons are among major conceptual-linguistic resources which are needed in many natural language processing applications especially where semantic processing is focused. Having such resources enables many semantic web and Natural Language Processing (NLP) applications.

One of the most famous semantic lexicons which has been the base for many others is WordNet. WordNet is a lexical ontology based on theories of psycho-linguistics about mental lexicon. WordNet designing was started under supervision of Professor G. A. Miller in the cognitive science laboratory of Princeton University in 1986 and its first version was presented in 1991.

WordNet is a rich computational linguistic resource for Natural Language Processing (NLP) used in Machine Translation, Internet Searches, Document Classification, Information Retrieval, and many web applications. After presenting English WordNet (Princeton), similar resources have been developed for more than 40 live languages all around the world. One of the main approaches to build a wordnet for a new language is using pre-existing lexical resources of other languages. English WordNet (Princeton WordNet) can help this process as an important lexical resource.

Persian language is the official language of Iran, Tajikistan and Afghanistan. This language with the Indo-Aryan languages constitutes the Indo-Iranian group within the Satem branch of the Indo-European family. The lack of linguistic resources such as lexical ontologies, semantic lexicons, electronic complete Persian thesauri, parallel corpora and even complete computational bilingual dictionaries have been some of the problems encountered in developing Persian NLP systems and spreading semantic web applications.

In this paper we offer an improved methodology for mapping Persian words to English WordNet synsets. To do the mapping, we calculate the similarity of Persian words and English synsets using their features such as super-classes and subclasses, domain and related words. Our approach is an improvement of an existing one [2] applying in a new domain, which increases the recall noticeably. The main resources we exploit for the mapping are an English-Persian dictionary [3] (including 252864 entries), a Persian-Persian dictionary [4] (incl. about 116 thousand entries) and a Persian thesaurus [5] (incl. about 10 thousand entries).

This paper is organized as follows: In Section 2, previous related works are described. Section 3 introduces our suggested approach and Section 4 presents some experimental results .Finally in Section 5 some conclusions and future works are discussed.

II. RELATED WORK

A Spanish research group [6] presented a new and robust approach for linking already existing lexical/semantic hierarchies. They applied a constraint satisfaction algorithm (relaxation labeling) to select the best match for a node of hierarchy among all the candidate nodes in the other side. They took advantage of hyperonymy and hyponymy relations in hierarchies. The following year, the same group [7] applied their work on mapping of nominal part of WordNet 1.5 to WordNet 1.6 with a high precision.

A Korean group [8] presented automatic construction of

[1] r.dehkharghani@mail.sbu.ac.ir
[2] m-shams@sbu.ac.ir

Korean WordNet from pre-existing lexical resources in 2000. Six automatic WSD (Word Sense Disambiguation) techniques were used for linking Korean words collected from bilingual MRD (Machine Readable Dictionary) to English WordNet synsets. They used Machine Learning methods to combine these six techniques.

Another group [9] presented observations on structural properties of WordNets of three languages: English, Hindi, and Marathi. They reported their work on linking English, Hindi and Marathi synsets. They proposed a formula for computing the similarities of nodes in two hierarchies.

Farreres [2] proposed a two-phase methodology for mapping Spanish thesaurus to English WordNet. His methodology is structured as a sequence of two processes. The aim of the first process that is based on a work in 1997 [10], is mapping of Spanish words to WordNet synsets. The second process takes advantage of hierarchies to accept or reject associations produced in the first phase.

One of the ways of constructing a WordNet for a certain language (source language) starts by mapping a thesaurus of source language to English (destination language) WordNet. This approach includes two processes. In the first process, words of source language are mapped to WordNet synsets. In the second process, these mappings are accepted or rejected according to the hierarchy of English WordNet and source language thesaurus

In Our work we have improved the first phase of Farreres' work-the most complete work due to 2007- and applied it on Persian language. We will show that our improvements will increment the recall noticeably while saves or also makes the precision a little bit better.

III. Suggested Approach

In the previous section a brief history of related works was presented. Since our approach is an improvement to Farreres' methodology, in this section we explain the first process of his work in parallel with our approach (called SBU methodology) and show the similarities and differences. Our goal is finding the most appropriate synset(s) for mapping Persian words to them. The suggested approach is language independent. It can be applied to any language and we used Persian language as a case study.

This approach takes advantage of some preexisting resources in the source language (Persian) and target language (English). Essential resources are bilingual Persian-English and English-Persian dictionaries, monolingual Persian-Persian dictionary, and English WordNet. We used Aryanpour dictionary as Persian-English and English-Persian dictionary, the Sokhan dictionary as Persian-Persian dictionary and WordNet 2.1.

At the start, for a Persian word PW, we should find its translations in a bilingual dictionary. For English translations (EW) of PW, we find its synsets in WordNet (WNS). As is shown in the Fig. 1, for each PW there are many candidate synsets in WordNet (WNS), the majority of which is not appropriate for PW. So we should specify truth probability of associations between PW and WNSs.

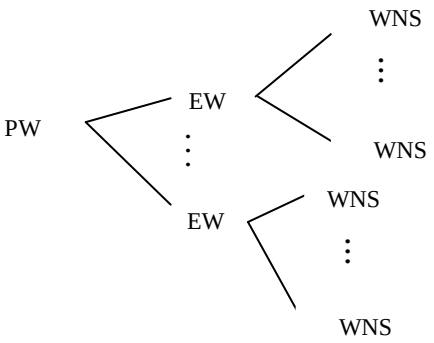

Fig. 1. Candidate WordNet synsets for a Persian word

A. Similarity Methods

According to Farreres' classification, similarity factors between PW and WNS are divided into four main groups regarding the kind of knowledge sources involved in the process: Class methods, Structural methods, Conceptual Distance methods and Hybrid methods.

Classification Methods

These methods classified Persian words in eight categories depending on its English translations (EWs) and their WordNet synsets (WNSs) for each EW. These methods are divided into two main groups, namely, Monosemous and Polysemous. Our approach is the same as Farreres' methodology in Classification methods.

a- Monosemous Group.

English words in this group have only one synset in WordNet. Four Monosemous methods are described below:

Mono1 (1:1): A Persian word has only one English translation. Also the English word has Persian word as its unique translation (Fig. 2).

Fig. 2. Mono1 method

Mono2 (1:N, N>1): A Persian word has more than one English translation. Also each English word has the Persian word as its unique translation (Fig. 3).

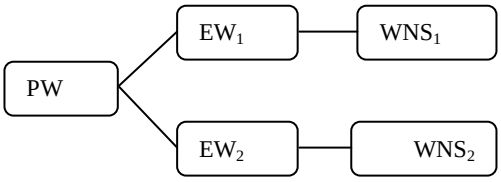

Fig. 3. Mono2 method

Mono3 (N:1, N>1): Several Persain words have the same translation EW. The English word EW has several translations to Persian. (Fig. 4).

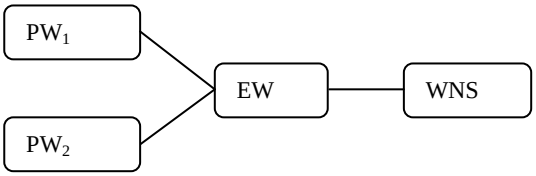

Fig. 4. Mono3 method

Mono4 (M:N, M,N>1): Several Persian words have different translations. English words also have several translations to Persian (Fig. 5). Note that there is at least two Persian words having several common English words.

b- Polysemous Group

English words in this group have several synsets in WordNet. Polysemous methods are like the Monosemous ones. We do not expand them for avoiding repetition.

Structural Methods

These methods are based on the comparison of the taxonomic relations between WordNet synsets. Four methods constituting structural methods are as follows:

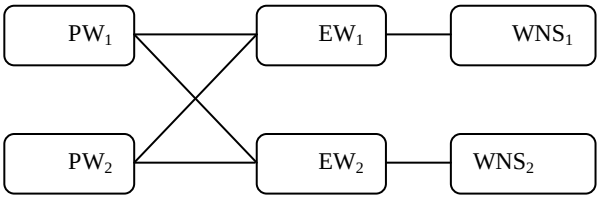

Fig. 5. Mono4 method

a- Intersection Method:

If English words share at least one common synset in WordNet, the probability of associating Persian word to common synsets increases (Fig. 6).

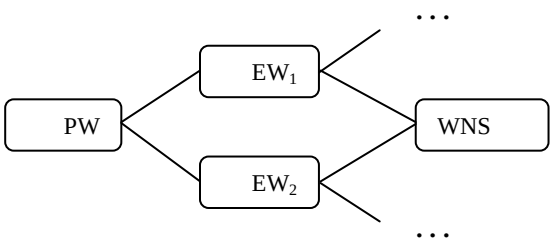

Fig. 6. Intersection method

b- Brother Method:

If some synsets of English words are brothers (they have common father), the probability of associating Persian word to brother synsets increases (Fig. 7).

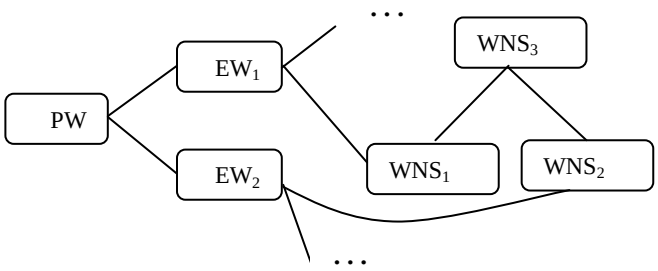

Fig. 7. Brother method

c- Ancestor Method:

If some synsets are ancestors of another synset, the probability of associating the Persian word to hyponym synset increases (Fig. 8).

d- Child Method:

If some synsets are descendants of another synset, the probability of associating Persian word to hyperonym synset increases (Fig. 8).

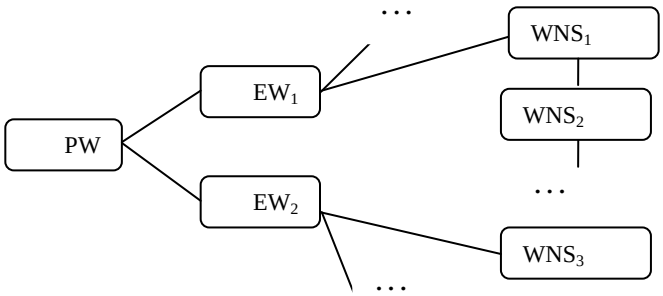

Fig. 8. Ancestor and Children method

Some differences of our approach (SBU) and Farreres' methodology lie in the structural methods. Farreres divided structural methods into Intersection, Brother, Father and Distant Methods. Intersection and Brother are the same as above. Father method is based on immediate hyperonym and Distant method is based on non-immediate hyperonyms. We merged two methods Father and Distant as Ancestor method. We applied Child method in a different way from Father and Distant methods, while in the Farreres' methodology they are not detached. Severance of Ancestor and Child methods causes to lead associations into hyperonym synsets with general meanings or hyponym synsets with specific meanings. This leading is done by means of training phase in machine learning techniques (explained below in Composition of Methods subsection). The mapping system learns which hyperonym or hyponym associations are more important than others in training phase. Then it applies this collected information to automatic computing of correctness probability of each association.

Conceptual Distance Methods

These methods are based on semantic closeness of synsets

in WordNet. There are many formulas computing conceptual distance between two concepts (word or synset). For example, it is defined in [12] as the length of the shortest path between two concepts in a hierarchy [2]. We used the equation 1 [11] for computing semantic similarity.

$$(1) \qquad sim(s,t) = \frac{2*dept\ h(LCA(s,t))}{dept\ h(s) + dept\ h(t)}.$$

in which s and t are the synsets; sim(s, t) is semantic similarity of s and t; depth(x) is depth of synset x regarding the root of WordNet hierarchy (the node "entity" for nouns); and finally LCA(s, t) is the Least Common Ancestor of synsets s and t. LCA(s, t) is an ancestor of s and t which is the deepest one in the WordNet hierarchy.

Two implications of equation 1 are (a) deeper synsets have higher semantic similarity together than the shallow ones and (b) shorter path between s and t causes higher semantic similarity. Farreres divided this group into three methods:

1) CD1 Method

This method uses co-occurrent words of Persian word. Following [13] two words are co-occurring in a dictionary if they appear in the same definition [2]. If some synsets of PW are semantically closer to some synsets of co-occurring words, probability of associating Persian word to its closer synsets increases.

2) CD2 Method

This method uses genus word(s) of Persian word. In fact, genus is one of hypernyms of PW. PW is a kind of genus word. If some synsets of PW are semantically closer to some synsets of genus words, probability of associating Persian word to its closer synsets increases. For example, Sokhan dictionary defines the Persian word آواز- avaz (song) as: صدایی که ...- sedayi ke … (the sound that …) . So the term صدا - seda (sound) is genus of آواز- avaz (song) and آواز- avaz (song) is a kind of صدا - seda (sound).

3) CD3 Method

This method is based on the semantic similarity of candidate synsets of Persian word. If some synsets of PW are semantically closer to all other candidate synsets, probability of associating Persian word to its closer synsets increases.

We considered these three methods in our approach but with two minor modifications. As the first difference, we utilized the words having "related-to" relation with PW instead of co-occurrence relation. We used Fararooy Thesaurus [5] for extracting "related-to" words of PW. Because co-occurrent words could not help us so much disambiguate the PW to find the best association. For example, as for the Persian word استاد – ostad (master), one of co-occurring words is محترم – mohtaram (respectable) because the term استاد محترم – ostade mohtaram (respectable master) is repeated many times in documents and dictionaries. But semantic similarity of these two words is very low. We used the words ماهر – maher (skillful) and آموزگار– amoozgar (instructor) extracted from Fararooy thesaurus which have "related-to" relation with the Persian word استاد – ostad (master). They have remarkable similarity with the main Persian word and could help disambiguate the meaning of استاد – ostad (master) more precisely.

The second modification is about CD3. We will exemplify to explain the modification. In the Farreres' methodology if two synsets have a brother relation together, value of both brother and CD3 methods becomes 1 for these two synsets, indicating that these synsets are brother and have high semantic similarity (low conceptual distance) since brother relation cause high semantic similarity.

This assignment makes create dependency between methods, while the methods must be independent from each other. According to statistical method Logistic Regression for estimating coefficients (importance) of each method (explained in subsection 3.3), this dependency prevents exact estimation of coefficients.

For this reason we got advantage of CD3 method only for synsets that do not have Brother, Ancestor and Child relations with other synsets. The last improvement of CD methods is using gloss and examples of synsets to achieve more similarities. If English translations of genus word(s) and "related-to" words (and semantic label explained in hybrid methods) occur in glosses or examples of some synsets of PW, the probability of associating Persian word to those synsets increases.

Hybrid Methods

In this group, two methods, namely, Variant and Field are presented without relation to other methods.

1) Variant Method

This method seeks WordNet synsets whose words share the same translations in English-Persian dictionary. In the other words, if two or more words of a synset have only one translation for the same Persian word, probability of associating Persian word with that synset increases.

2) Field Method

It uses semantic label(s) of Persian word. This label indicates the domain of Persian word PW and PW is a member of that domain. If some synsets of PW are semantically closer to some synsets of semantic label(s), the probability of associating Persian word with its closer synsets increases.

For example, Sokhan dictionary defines the Persian word اردک – ordak (duck) as: (پرنده ای که) جانوری) ... – (janevari) parandeyi ke …((animal) a bird that …) , then the term جانوری – janevari (animal) is the semantic label of اردک – ordak (duck).

Now, let us analyze hybrid methods. Variant method is the inverse case of Intersection method in Structural group but Intersection starts from the Persian word to arrive at WordNet synset, while Variant starts from WordNet synset to arrive at Persian word. Here the dependency problem appeared in CD3 method is the same but more obvious than previous case. In the other word, if PW shares its translations in one synset, value of Intersection and Variant methods will be 1. Actually the Intersection and Variant methods are the

same and dependency of these two methods is another drawback of Farreres' methodology. Therefore, we eliminated the Variant method in our approach.

As for Field method, we applied it as a member of conceptual distance methods. Then the name of hybrid methods is deleted in our approach. We mention again that genus and semantic label of a Persian word is extracted from Sokhan dictionary and the "related-to" words are extracted from Fararooy thesaurus.

B. Presentation of Similarities

Farreres used the vector (SW-synID, m_1,m_2, … , m_{17}, Accept or Reject) to present associations between Spanish words and WordNet synsets. For example, the vector (SW_1- 8054099 , 000000101000110 , Reject) indicates, an association between SW_1 and synset with ID 8054099 is rejected.

m_i specifies whether the i^{th} method can be applied to this association or not. The value 1 indicates that the method is applicable and 0 indicates that it is not applicable to the association. In this example only m_7, m_9, m_{13} and m_{14} methods could be applied to this association. The value of m_9 in this example means that at least two English translations of SW_1 are located in the synset with ID 8054099. But there is no difference as to how many translations of Spanish word share the synset. This is another drawback of Farreres' methodology. It means that under the same condition, probability of associating SW with a synset that shares two translations of SW are the same as other synset that shares, say, four translations. This problem recurs in other methods except for classification methods. We solved this problem by using the values 0 to 5 instead of 0 and 1. In other words, two values of 0 to 1 were replaced by six values of 0 to 5. For example, in the Intersection method, depending on the number of English translations of Persian word that share a synset, the values 1, 2 and 3 are assigned to m9 respectively.

Table 1 compares SBU and Farreres' methodologies regarding the methods used.

As an example to compare two methodologies suppose that PW is included in poly3 method by one of its translations in Classification methods. Consequently m_7 is 1. Four words of its translations share the synset WNS_1, then the value 1 is assigned to m_9 and m_{17} in Farreres. But in the SBU, only m_9 is assigned by value 3. This synset does not have brother relation with other candidate synsets, then value of m_{10} in Farreres and SBU is 0. WNS_1 does not have an immediate hypernym among candidate synsets but two synsets of translations of PW are the second and third level hypernyms of WNS_1. As a result in Farreres, m_{11} and m_{12} get values 0 and 1 respectively and in SBU, only m_{11} gets the value 2. We considered only five levels of ancestors (and also five levels of children) for a synset. So in this case, value 2 is suitable for m_{11}.

TABLE 1. COMPARISON OF SBU AND FARRERES' METHODOLOGIES

Method groups		SBU	Farreres
Classification	m_1	Mono1	Mono1
	m_2	Mono2	Mono2
	m_3	Mono3	Mono3
	m_4	Mono4	Mono4
	m_5	Poly1	Poly1
	m_6	Poly2	Poly2
	m_7	Poly3	Poly3
	m_8	Poly4	Poly4
Structural	m_9	Intersection	Intersection
	m_{10}	Brother	Brother
	m_{11}	Ancestor	Father
	m_{12}	Children	Distant
Conceptual Distance	m_{13}	Related-to	CD1
	m_{14}	Genus (CD2)	CD2
	m_{15}	CD3	CD3 (m_{15})
	m_{16}	Field (m_{16})	
Hybrid			Variant(m_{16})
			Field(m_{17})

Two other candidate synsets of PW are immediate children of WNS_1. This relation does not change values of methods of Farreres but the value 1 is assigned to m12 in SBU. WNS_1 does not have any close semantic similarity with candidate synsets of co-occurrent words and those words that have "related-to" relation with PW, then m_{13} is 0 in both methodologies. The sum of semantic similarities of candidate synsets of PW with candidate synsets of genus word of PW is 3.83. It causes to assign the values 1 and 4 to m_{14} in Farreres and SBU methodologies respectively. There is no semantic similarity between candidate synsets of PW and its semantic label, thus the value 0 is assigned to m_{16} and m_{17} in SBU and Farreres respectively.

Finally the values 0 and 1 are assigned to m_{15} in SBU and Farreres respectively. Despite the fact that WNS_1 has some semantic relations like hyperonymy and hyponymy with other candidate synsets, we consider the CD3 method just for synsets that have no close relations like Intersection, Brother, Ancestor and Child with other candidate synsets. Note that eliminating this condition causes a dependency between each structural method with CD3 method. For example, if a synset has a brother among candidate synsets, the value 1 is assigned to the Brother method, and also the value of CD3 becomes 1. Note that brother relation is a kind of close semantic similarity. This dependency is explained above in Variant method and is a drawback of Farreres' methodology. Table 2 shows comparison of vectors of the example explained.

TABLE 2. VECTORS OF EXPLAINED EXAMPLE FOR EACH METHOD

Farreres	00000010100101001
SBU	0000001030230201

C. Composition of Methods

Now some questions come into mind: Are all of methods useful? Should they be independent? How important is each of them? How can we specify their coefficients for computing final similarity?

We should specify coefficients of each method in final equation of probability computation. Then the input of our methodology is an association between PW and a synset having vector of 16 values and the output is the correctness probability of that association.

To achieve this goal, we took advantage of Logistic Regression model [14] like Farreres' methodology. Logistic Regression is a statistical method for calculating the importance coefficients of each method in the composition. A positive regression coefficient means that that method increases the probability of the outcome (association correctness), while a negative regression coefficient means that method decreases the probability of that outcome. Actually this model is used as a Machine Learning method whose training phase includes analyzing input data (the associations, their vectors and their human evaluation) and the test phase computes P(ok) that is correctness probability of an association according to its vector of methods. Equation 2 is the formula computing P(ok) using Logistic Regression.

$$(2) p(ok) = \frac{e^{\beta 0 + e^{\sum \beta_i m_i}}}{1 + e^{\beta 0 + e^{\sum \beta_i m_i}}}$$

β_i is coefficient of i^{th} method but β_0 is a constant. The higher value of β_i means the higher impact of m_i on probability computation. m_i is value of i^{th} method in the association. We used SPSS as a statistical tool for Logistic Regression.

D. Training Phase

At first, we applied our methodology on 150 Persian words. Having computed vectors of each association, about 2500 associations between Persian words and WordNet synsets were created. For regressing these associations, it was necessary to enter only some of them and their correctness probability achieved by human evaluation to SPSS. Of course the more associations given to SPSS leads to more accuracy in computation of coefficients. SPSS estimates coefficients according to correctness probabilities of given associations. For this reason we classified associations in groups having the same vector. Then about 120 groups were achieved. Groups having less than 5 vectors were eliminated because their effects in this regression were very low. For each association of each group, we accepted or rejected it. For example, the vector 0000000104400111 was accepted in 40 cases and was rejected in 10 cases, then its correctness probability by human evaluation is 40 / 50 = %80.

After computing of this probability for each vector, we entered them into SPSS. Then coefficients of each method were achieved. We repeated this regression for Farreres' methodology. Results are presented in Table 3.

TABLE 3
COEFFICIENTS OF METHODS IN EACH METHODOLOGY

β_i	SBU	Farreres
β_0	-3.505	-2.291
β_1	0	0
β_2	0	0
β_3	1.515	0.3
β_4	0	-0.301
β_5	0	22.037
β_6	0	0
β_7	0.510	-0.683
β_8	0	-0.86
β_9	1.643	1.628
β_{10}	0.639	0.503
β_{11}	0.311	0.973
β_{12}	0.974	0.302
β_{13}	0.673	0.137
β_{14}	0.408	1.054
β_{15}	-2.140	0.403
β_{16}	0.177	0
β_{17}	-	-0.315

Now we justify coefficients of our methodology. As can be seen, some methods have coefficients zero. This might have two reasons: (1) these methods occur rarely in practice, and (2) their influence on final probability is very low. Since the methods mono1, mono2, mono4, poly1 and poly2 occur rarely in practice (and also in test data), their coefficients are zero. But as for poly4, although this method is repeated a lot, it does not change final probability noticeably. Therefore its coefficient is zero as well. Values of other methods are justifiable according to their effect and importance in computing probability. For example, intersection of two words in a synset has more effect than brother relationship of synsets. Negative coefficient of m_{15} (CD3) is due to the fact that it is applied only to associations whose values of their structural methods are zero. It means that in these associations, there is no close semantic similarity with other candidate synsets; then in these cases, negative coefficient reduces correctness probability of association.

IV. EVALUATION

To evaluate our work, we compared its results with Farreres'. For this comparison we set the acceptance threshold to different values and calculated the precision and recall for each threshold. Before describing the comparison results lets clear the issue by an example.

Consider the Persian word بغض – boghz (spite, hatred). The words دشمنی – doshmani (enemy) and کینه – kineh (rancor) have "related-to" relation with this Persian word obtained from Fararooy thesaurus and its genus and semantic label are احساس – ehsas(sensation) and روانشناسی –

ravanshenasi (psychology) respectively obtained from Sokhan dictionary. Results of our methodology for the word بغض – boghz (spite, hatred) are presented in Table 4.

In this table in the forth column A stands for Accept and R for reject and shows the human evaluation of this association. If we select a threshold between 0.30 and 0.40, then associations 1, 2, 4 and 7 are correctly and only association 6 is incorrectly accepted.

TABLE 4. CANDIDATE ASSOCIATIONS FOR PERSIAN WORD بـغض – *BOGHZ* (SPITE, HATRED)

EW	Synset ID	Vector	Human Eval.	Esti-mation
grudge	7446948	0001000000200310	A	0.39
spite	7448078	0000000100200310	A	0.39
spite	4787145	0000000100000210	R	0.15
hatred	7443888	0001000000030320	A	0.9
dislike	6119053	0000001000000123	R	0.27
dislike	7399432	0000001000000320	R	0.46
animus	7445512	0010000000200310	A	0.74

We employed precision and recall measures for evaluating and comparing our methodology (SBU) with Farreres' methodology. Results of applying two methodologies to Persian language are presented in Table 5 and their comparisons are presented in Fig. 9 and Fig. 10.

TABLE 5. COMPARISON OF PRECISIONS AND RECALLS OF SBU AND FARRERES' METHODOLOGIES

Threshold	Precisions		Recalls	
	SBU	Farreres	SBU	Farreres
0.25	0.53	0.53	0.91	0.67
0.30	0.58	0.58	0.77	0.62
0.35	0.60	0.60	0.72	0.60
0.36	0.61	0.61	0.71	0.60
0.37	0.61	0.61	0.71	0.57
0.38	0.62	0.61	0.70	0.57
0.39	0.62	0.62	0.68	0.57
0.40	0.63	0.62	0.67	0.56
0.45	0.64	0.66	0.62	0.47
0.50	0.66	0.68	0.58	0.42
0.60	0.69	0.72	0.53	0.29
0.70	0.72	0.76	0.51	0.27

As can be seen, for each threshold of accepting or rejecting associations, we obtained various precision and recall values. Since in the second phase of this work, the pre-produced associations will be accepted or rejected ultimately, production of associations is more important than their correctness in the first phase. In other words, high value of recall is more important than high value of precision because most of incorrect associations will be rejected further in the second phase using hierarchical structures of Persian thesaurus and WordNet. This final acceptance or rejection will take advantage of hyperonymy and hyponymy relations in the hierarchies. Then we do not have to select a decisive threshold value in this phase. Also note that the structural similarity between Persian thesaurus and English WordNet was not used in the first phase.

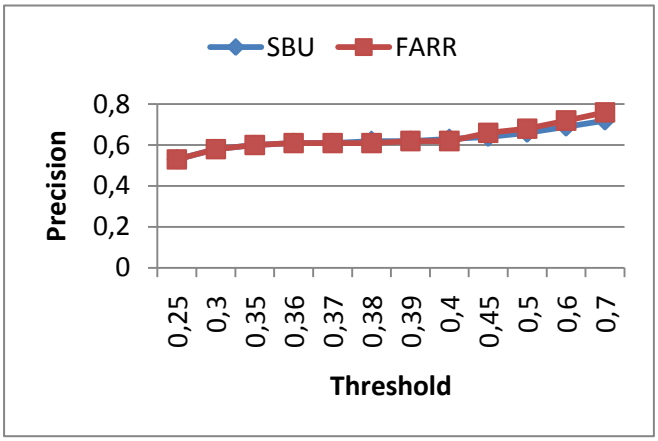

Fig. 9. Comparison of precision of two methodologies

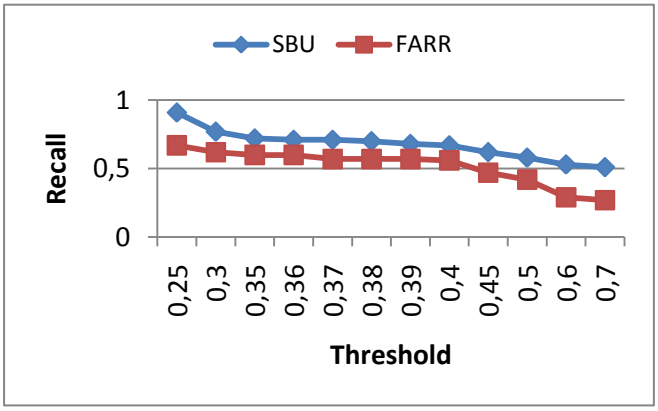

Fig. 1. Comparison of recalls of two methodologies

V. CONCLUSION AND FUTURE WORKS

In this paper we proposed an improved methodology based on Farreres' methodology for mapping Persian words to WordNet synsets. The methodology is language independent and we used Persian language as a case study. The recall values we achieved in our methodology were higher than those achieved in Farreres' methodology. An association between Persian word and every candidate synset for it was constructed. This work took advantage of 16 similarity methods indicating how similar a Persian word is to each of its candidate synset.

We obtained coefficients (importance) of each method used in computing correctness probability of each

association by Logistic Regression model. This model uses evaluated associations by human for estimating coefficient of each method. Finally we obtained a formula whose input is an association and whose output is correctness probability of this association.

After evaluating our methodology, different Precisions and recalls were obtained based on threshold values. In the future works, we will do second phase of this methodology. In the second phase, pre-produced associations will be accepted or rejected ultimately using the structural properties of synsets in two languages. In the first phase, high value of recall is more important than high value of Precision because most of incorrect associations will be rejected in the second phase using hierarchical structures of Persian thesaurus and WordNet. This ultimate acceptance or rejection, will take advantage of hypernymy and hyponymy relations in the hierarchies.

ACKNOWLEDGMENT

This work has been funded by Iran Telecommunication Research Center (ITRC) under contract no. T/500/19231.

REFERENCES

[1] T. R. Gruber, A Translation Approach to Portable Ontologies, knowledge Acquisition, 5(2), pp. 199-220, 1993.

[2] J. Farreres, Automatic Construction of Wide-Coverage Domain-Independent Lexico-Conceptual Ontologies. PhD Thesis, Polytechnic University of Catalonia, Barcelona, 2005.

[3] M. Assi, M. Aryanpour, Aryanpour English-Persian and Persian-English dictionary. http://www.aryanpour.com.

[4] H. Anvari, Persian-Persian Sokhan dictionary. Tehran: Sokhan Pub., 2002.

[5] J. Fararooy, Fararooy Persian Thesaurus. Available at http://www.persianthesaurus.com

[6] J. Daudé, L Padró, G. Rigau, Mapping multilingual hierarchies using relaxation labeling. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99), Maryland, 1999.

[7] J. Daudé, L Padró, G. Rigau, Mapping Wordnets using structural information. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China (2000)

[8] C. Lee, G. Lee, S. Jung Yun, Automatic Wordnet mapping using word sense disambiguation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 2000), Hong Kong (2000)

[9] J. Ramanand, A. Ukey, B. Kiran Singh, P. Bhattacharyya, Mapping and Structural Analysis of Multi-lingual Wordnets. IEEE Data Eng. Bull. 30(1): 30-43 (2007)

[10] J. Atserias, S. Climent, X. Farreres, G. Rigau, H. Rodriguez, Combining multiple methods for the automatic construction of multilingual Wordnets. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), Tzigov Chark, Bulgaria (1997)

[11] T. Simpson, T. Dao,: WordNet-based semantic similarity measurement. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx

[12] R. Rada, H. Mili, E. Bicknell, M. Blettner.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1):17–30, (1989)

[13] Y. Wilks, D. Fass, C. Guo, J. McDonal, T. Plate, B. Slator, Semantics and the Lexicon. chapter Providing Machine Tractable Dictionary Tools, pages 341–401. Kluwer Academic Publishers, Dordrecht (1993)

[14] L. Lebart,. Traitement Statistique des Données. DUNOD, Paris (1990)