Automatic Alignment of Persian and English Lexical Resources: A Structural-Linguistic Approach

Rahim Dehkharghani and Mehrnoush Shamsfard

Natural Language Processing Laboratory, Shahid Beheshti University, Tehran, Iran

Abstract. Cross-lingual mapping of linguistic resources such as corpora, ontologies, lexicons and thesauri is very important for developing cross-lingual (CL) applications such as machine translation, CL information retrieval and question answering. Developing mapping techniques for lexical ontologies of different languages is not only important for inter-lingual tasks but also can be implied to build lexical ontologies for a new language based on existing ones. In this paper we propose a two-phase approach for mapping a Persian lexical resource to Princeton's WordNet. In the first phase, Persian words are mapped to WordNet synsets using some heuristic improved linguistic approaches. In the second phase, the previous mappings are evaluated (accepted or rejected) according to the structural similarities of WordNet and Persian thesaurus. Although we applied it to Persian, our proposed approach, SBU methodology is language independent. As there is no lexical ontology for Persian, our approach helps in building one for this language too.

1 Introduction

WordNet is a rich computational linguistic resource for Natural Language Processing (NLP) used in machine translation, internet searches, document classification, information retrieval, question answering and many web applications. It is useful to use existing lexical resources for constructing WordNet for a certain language. English WordNet (Princeton's WordNet) has been employed for automatic creation of WordNets for many other languages. On the other hand, cross-lingual mapping between various WordNets of various languages enables many cross-lingual applications like the ones mentioned above.

In this paper, we propose a new approach for mapping Persian words' senses to WordNet synsets. This approach includes two phases. In the first phase, words of source language (Persian) are mapped to WordNet synsets. The associations of first phase are not reliable; therefore, we should get help from other sources. In the second phase, extracted mappings are accepted or rejected according to the hierarchy of English WordNet and Persian thesaurus.

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 conclusions and future works are discussed.

Dehkharghani R. and Shamsfard M. (2009).
Automatic Alignment of Persian and English Lexical Resources: A Structural-Linguistic Approach .
In *Proceedings of the 6th International Workshop on Natural Language Processing and Cognitive Science* , pages 55-66
DOI: 10.5220/0002175000550066

2 Related Works

Daude and colleagues [1] presented a new and robust approach for mapping multi-lingual 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 hypernymy and hyponymy relations in hierarchies. The following year, the same group [2] applied their work on mapping of nominal part of WordNet 1.5 to WordNet 1.6 with a high precision.

Lee and colleagues [3] presented automatic construction of Korean WordNet from existing lexical resources in 2000. Six automatic WSD (Word Sense Disambiguation) techniques were used for mapping Korean words collected from bilingual MRD (Machine Readable Dictionary) to English WordNet synsets. These techniques use the synonymy relations, IS-A relations and glosses of synsets in WordNet. They used Machine Learning methods to combine these six techniques.

Mihaltz and Proszeky [4] presented the results of creating the nominal database of Hungarian WordNet. They presented 9 different automatic methods, developed for mapping Hungarian nouns to WordNet 1.6 synsets. Those methods are extracted from bilingual and monolingual dictionaries and Princeton's WordNet. Synonymy and hypernymy relations and cooccurrence words were used for this mapping.

Rmanand and colleagues [5] presented observations on structural properties of WordNets of three languages: English, Hindi, and Marathi. They reported their work on mapping English, Hindi and Marathi synsets. The translations of Hindi words and equivalent synsets of each translation in English WordNet were obtained. Then the similarity of two synsets in Hindi and English hierarchy are computed by a formula considering the level of Hindi and English synset. They took advantage of synonymy and hypernymy relations.

Rodriguez and colleagues [6] focused on the semi-automatic extension of Arabic WordNet (AWN) using lexical and morphological rules and applying Bayesian inference. The AWN project was finished in 2008 but this group constructed a little part of it semi-automatically. In this research, a novel approach to extending AWN is presented whereby a Bayesian Network is automatically built from the graph and then the net is used as an inference mechanism for scoring the set of candidate associations between Arabic verbs and WordNet synsets .

Farreres [7-10] proposed a two-phase methodology for bilingual mapping of ontologies (Spanish thesaurus to English WordNet as a case study). This work is among the most comprehensive in the area of bilingual ontology mapping. The result of this work was the automatic construction of nominal part of Spanish WordNet. Furthermore, it was used in semi-automatic construction of Arabic WordNet [6]. Therefore we chose it as a base and made some improvements on it. These two methodologies (Farrares' and SBU) are compared with each other in sub-section 3.3. Farreres' methodology is structured as a sequence of two processes. The aim of the first phase that is based on a work in 1997 [11], is mapping Spanish words (SWs) to WordNet synsets (WNSs) using 17 similarity methods. To compose these similarity methods, Farreres used the Logistic Regression model. He obtained the formula which whose input was an association between SW and a WNS and the output is correctness probability of that association. The second phase takes advantage of Spanish thesaurus and WordNet hierarchies to accept or reject associations produced

in the first phase. The mapping in the second phase is sense to synset instead of word to synset.

Current work is the next generation of the work introduced at [12]. In our previous work we mapped Persian words to English synsets in a different way using a heuristic improved translation and dictionary based method.

3 Our Suggested Approach

Our goal is finding the most appropriate English synset(s) to be mapped to Persian thesaurus nodes. The suggested approach is language independent. It can be applied to any pair of languages and we used Persian-English pair as a case study.

This approach takes advantage of some existing resources in the source (Persian) and target language (English). Essential resources are bilingual Persian-English and English-Persian dictionaries, monolingual Persian-Persian dictionary, Persian thesaurus and English WordNet. We used Aryanpour dictionary [13] (including 252864 entries) as Persian-English and English-Persian dictionary, Sokhan dictionary [14] (incl. about 116 thousand entries) as Persian-Persian dictionary and WordNet 2.1.

The schema of two phases of the approach is shown in Fig. 1.

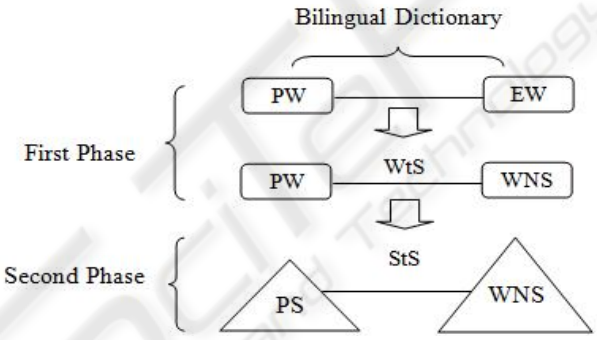

Fig. 1. Processes of first and second phases.

In this figure and the rest of the paper, PW stands for Persian Word, EW for English Word, PS for Persian Sense, WNS for WordNet Synset, WtS for Word-to-Synset association and StS for Sense-to-Synset association. As shown in Fig. 1, the first phase includes finding English translations of each PW using a bilingual dictionary and also equivalent synsets of each translation. In the second phase, Persian senses, instead of Persian words, should be mapped to WordNet synsets.

3.1 First Phase: Mapping Persian Words (PW) to WordNet Synsets (WNS)

Translations of each PW, should be found in a bilingual dictionary. Also for each English translation (EW) of PW, its synsets in WordNet should be obtained. There are many candidate synsets (WNSs) for each PW in WordNet the majority of which is not

appropriate for PW. The bilingual dictionary provides 4.93 English translations on average for each Persian noun. These English translations correspond to 3.21 WordNet synsets on average. So we should specify truth probability of associations between PW and WNSs using some similarity methods.

Similarity Methods. Similarity methods are divided into three main groups regarding the kind of knowledge sources involved in the process: Classification methods, Structural methods and Conceptual Distance methods.

Classification methods include two subgroups, namely, Monosemous and polysemous.

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.

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.

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

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

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

Intersection Method: If English words share at least one common synset in WordNet, the probability of associating Persian word to common synsets increases.

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

Ancestor Method: If some synsets are ancestors of another synset, the probability of associating the Persian word to hyponym synset increases.

Child Method: If some synsets are descendants of another synset, the probability of associating Persian word to hypernym synset increases.

Conceptual Distance Methods are based on semantic closeness of synsets in WordNet. There are many formulas computing conceptual distance (CD) between two concepts (word or synset). For example, it is defined in [15] as the length of the shortest path between two concepts in a hierarchy [10]. We used the equation 1 [3, 16] for computing semantic similarity between two concepts.

$$sim(s,t) = \frac{2 * depth(LCA(s,t))}{depth(s) + depth(t)}. \quad (1)$$

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 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. CD methods are divided into four:

CD1 Method: This method uses the words which have the *related-to* relation with PW. If some synsets of PW are semantically closer to some synsets of *related-to* words, probability of associating Persian word to its closer synsets increases. The *related-to* words are extracted from Fararooy thesaurus[1] [17]. For example, the word آموزگار (*amoozgar*, instructor) is related to استاد (*ostad*, master).

CD2 Method: This method uses genus word(s) of PW. In fact, genus is one of hypernyms of PW. 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)[2] 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).

CD3 Method: This method is based on the semantic similarity of candidate synsets of PW. If some synsets of PW are semantically closer to all other candidate synsets, probability of associating Persian word to its closer synsets increases.

CD4 Method: It uses semantic label(s) of Persian word. This label indicates the domain of PW. 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 اي که (جانوري) ((*janevari) parandeyi ke* …, ((animal) a bird that …) , then the term جانوري (*janevari*, animal) is the semantic label of اردک (*ordak*, duck).

Presentation of Similarities. We used the vector (PW-EW-WNS, m_1,m_2, … , m_{16}, AccOrRej) to present associations between PWs and WNSs according to English translations (EWs) of PW. $\underline{m_i}$ (i^{th} method) can be assigned values between 0 to 5 to indicate its intensity in association. The value 0 indicates that the method is not applicable to the association and the values 1 to 5 indicate the intensity of application. AccOrRej indicates accepting or rejecting the association.

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?

[1] This thesaurus provides only “related-to” relations between Persian words and has about 7500 word groups including the words that are related together

[2] First term in parenthesis is pronunciation of Persian term and other term(s) is its English translation

Coefficients of each method in final equation of probability computation should be specified. The input of our methodology is an association between PW and a synset having vector of 16 values and the output is correctness probability of that association. To achieve this goal, we took advantage of Logistic Regression model [18]. Equation 2 is the formula computing correctness probability of an association (P(ok)) using Logistic Regression.

$$p(ok) = \frac{e^{\beta_0} + e^{\sum \beta_i m_i}}{1 + e^{\beta_0} + e^{\sum \beta_i m_i}} \quad (2)$$

β_i is coefficient of i^{th} method but β_0 is a constant. The higher the value of β_i, the higher the impact of m_i on probability computation. m_i is value of i^{th} method in an association.

We used SPSS as a statistical tool for Logistic Regression. For training step, 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 to SPSS.

3.2 Second Phase: Taxonomy Matching

After the first phase ended, the candidate synsets were obtained with different probabilities for each Persian word. In this phase, the Persian senses, instead of Persian words, should be mapped to WordNet synsets exploiting taxonomy matching. To do the taxonomy matching we need hypernymy and hyponymy relations between senses. We used about 200 Persian senses of nominal part of FarsNet (an ongoing project to develop Persian WordNet) as the test data.

We used a sense-level taxonomy instead of a word-level one. For two reasons, the sense-level is more suitable in this mapping: 1) there are usually several parents for a word in word-level taxonomy (each parent for each sense) [9]; 2) also, using the sense-level taxonomy, we can specify that on the basis of which sense of PW it is associated with WNS. Therefore, the mapping in this phase is more complicated than the first phase because the sense of PW in an association should be specified.

The candidate synsets of the sense PS are summation of candidate synsets of all PWs which are the member of PS, but the common synsets in this collection, which are shared by several PWs, considered as only one.

As Fig. 1 showed the aim of the second phase is exchanging some WtSs to StSs. WtS means a weighted association between Persian word and WordNet synset [8], having a probability value as its weight while StS means Sense-to-Synset association. As mentioned before, a PW may be repeated in one or more PS(s) depending on the number of its meanings (senses). A PW is called monosemous, if it has only one sense, and it is called polysemous, if it has several senses. It is obvious that the mapping process is easier for monosemous PWs.

There are one or more candidate synsets (usually more than one) for each PS. The problem lies in finding the most appropriate WNS(s) for each PS. To evaluate the StSs, we took advantage of results obtained in the first phase and hierarchical similarities of Persian and English branches in the second phase. The parameters which are considered in evaluation of an StS are nconn, $level_p$, $level_g$, gap and prob (correctness probability of WtS). $Level_p$ is the level of Persian sense according to the base sense. $Level_e$ is the level of WordNet synset according to the base synset. *nconn* is the total number of associations between Persian senses (including base sense and its hypernyms) and WordNet synsets (including base synset and its hypernyms). A *gap* in a Persian branch is a sub-branch without associations which separates two sub-branches with associations.

Suggested Algorithm. We have presented a formula to compute the correctness probability of each StS (equations 3 and 4).

$$FinalProb = \frac{\sum_{l_p=0}^{5}\sum_{l_{e=0}}^{10}((Co_p * (6 - l_p) + Co_e * (11 - l_e) * prob) - GapValue}{417} \quad (3)$$

Where

$$GapValue = \sum_{l_p=0}^{5} \begin{cases} Co_g * (6 - l_p) & if\ there\ is\ a\ gap\ in\ l_p \\ 0 & if\ there\ is\ no\ gap\ in\ l_p \end{cases} \quad (4)$$

l_p and l_e are the levels of PS and WNS respectively. Co_p and Co_e are the coefficients of Persian and English branches indicating the importance (influence) of each branch in the value of FinalProp. Prob is the correctness probability (p(ok)) of an association obtained in the first phase. Since the gaps cause the reduction of the correctness probability of StS, we reduced the value of FinalProb for each gap according to its level. GapValue is the summation of gaps' scores (equation 4). Because we need a value between 0 and 1, the FinalProb is divided into the value 417. The number of all possible associations between six Persian senses (one base sense and five hypernyms) and 11 synsets (one base synset and ten hypernyms) is 417. The English branch is more complete and enriched than Persian one, and considering the number of hypernyms more than 5 and 10 in the Persian and English branches respectively, does not increase the precision of mapping noticeably, therefore, we used five and ten hypernyms in Persian and English branches. Values of Co_p, Co_e, Co_g and *threshold* are computed by conducting a lot of experiments. We evaluated the associations by lots of values of mentioned parameters and we obtained the values 2.1, 0.8, 0.08 and 0.15 for Co_p, Co_e, Co_g and *threshold* as the most appropriate values.

The system should decide which association can be accepted and which should be rejected. In the second phase, given a PW, the steps of SBU methodology are as below:

- Specify the PSs of PW
- Specify the total number of WtSs from PSs to WNSs
- If PW is monosemous
  - If PW has only one WtS, it is accepted as StS for the PS which includes the PW
  - If PW has several WtSs which have the FinalProb value higher than the threshold (0.147) are accepted and the rest are rejected.
- If PW is polysemous, it is repeated in several PSs. For each WtS_i of PW in PS_i, there are equivalent WtSs in other senses (e.g. PS_j) of PW, which share WNSs with WtS_i. Then, only the WtS which have the higher FinalProb is accepted as StS. Of course, the FinalProb of accepted WtS has to be higher than threshold, otherwise it will be rejected.

3.3 Comparison of SBU and Farreres' Methodologies

The differences of first phase in two methodologies are:

- We merged two methods, Father (immediate hypernym) and Distant (non-immediate hypernym) as proposed by Farreres as Ancestor method.
- We applied Child method in a different way from Father and Distant methods, while in the Farreres' they are not detached. Severance of Ancestor and Child methods causes to lead associations into hypernym synsets with general meanings or hyponym synsets with specific meanings. This leading is done by means of training step. The mapping system learns which hypernym or hyponym associations are more important than others in training step.
- We utilized the words having "related-to" relation with PW instead of co-occurrence relation because at most cases, "related-to" words were more close to the PW rather than co-occurrent words. We used Fararooy thesaurus for extracting "related-to" words of PW.
- We got advantage of CD3 method only for synsets that do not have Brother, Ancestor and Child relations with other synsets. Having these relations 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, this dependency prevents exact estimation of coefficients. This limitation was not considered in Farreres' methodology.
- In Farreres' there is a Hybrid similarity method consisting of Variant and Field methods. Variant method is the inverse case of Intersection method in Structural group but Intersection starts from the PW to arrive at WNS, while Variant starts from WNS to arrive at PW. The dependency of these two methods is another drawback of Farreres' methodology. Therefore, we eliminated the Variant method in our approach. We also eliminated Field

Method which is the same as CD1 method but uses the semantic label(s) of SW which indicate the domain of SW.

- We used the numbers 0 to 5 instead of 0 and 1 to represent the intensity of each method. In other words, two values of 0 to 1 were replaced by six values of 0 to 5. For example, in the Intersection method, if two, three or four English translations of PW share a synset, 1, 2 or 3 are assigned to m_9 respectively, while in Farreres' methodology, there is no difference between these cases and the value of m_9 is 1 in all of them.

In the second phase, Farreres proposed a bootstrapping having these steps in which for each association the PRB denotes the pair of related branches:

- A disconnected PRB is a good indicator for incorrectness of the originating StS.
- A connected PRB is, on the other side, an indicator that the base StS would be correct, although in many cases it wasn't enough evidence; some factors (nconn, level, gaps) can be extracted to filter those connected PRB. It means that the PRB having lowest level, highest nconn and no gap, would be accepted.
- Other PRBs are pending to the next iteration because there were not enough evidences for accepting or rejecting them.

Differences of two methodologies in the second phase are as follows:

- Farreres' methodology needs the total number of nodes of a thesaurus for mapping because an StS can be accepted based on pre-accepted StSs. Our methodology does not need it but only the PRB which would be evaluated. One of the advantages of our methodology is that it can be used in languages which do not have a complete thesaurus because the suggested approach evaluates an StS independent from other ones. Of course, pre-accepted StSs help to map more exactly, but lack of them does not limit the methodology.
- The associations of hypernym senses of Spanish and English branches in Farreres' methodology are considered as 0 and 1, not as a probability value. In SBU methodology, they are considered as a value between 0 and 1 and so more precise values of final correctness probability of an StS can be computed.
- The parameter *level* in the Farreres' is the level of first connected sense from the base sense to WordNet branch but we studied all levels of associations between two branches, either on Persian ($level_p$) or WordNet level ($level_e$).
- In Farreres' methodology only the number of Spanish ancestors in the Spanish branch with some association to the WordNet branch is considered as nconn, but in SBU methodology, the total number of associations between PSs and WNSs are studied as nconn, which covers the Farreres definition of nconn, too.
- Some specific techniques that are used in SBU methodology, could not be considered in the Farreres' methodology. For example, when there are several association from a PS_i to WNS_j (their levels are higher than 0), several probabilities exist between the PS_i and WS_j. What is done in SBU methodology in these cases is use the arithmetical mean (average) of probabilities. Since in the Farreres' methodology, there is no probability in hypernym associations, no decision considered for this case.

4 Evaluation

In the first phase, the coefficients of Table 1 are obtained for two methodologies. According to those coefficients, different precisions and recalls were obtained depending on threshold values. The highest precision and recall values for SBU methodology are 0.62 and 0.7, and for the Farreres' methodology are 0.61 and 0.57 respectively, considering 0.38 as the threshold.

Table 1. Comparison of similarity methods and their coefficients in 2 methodologies.

Methods' Category	β_i	SBU		Farreres- based	
Classification	β_0	-	-3.505	-	-2.291
	β_1	Mono1	0	Mono1	0
	β_2	Mono2	0	Mono2	0
	β_3	Mono3	1.515	Mono3	0.3
	β_4	Mono4	0	Mono4	-0.301
	β_5	Poly1	0	Poly1	22.037
	β_6	Poly2	0	Poly2	0
	β_7	Poly3	0.510	Poly3	-0.683
	β_8	Poly4	0	Poly4	-0.86
Structural	β_9	Inters.	1.643	Inters.	1.628
	β_{10}	Brother	0.639	Brother	0.503
	β_{11}	Ancestor	0.311	Father	0.973
	β_{12}	Child	0.974	Distant	0.302
Conceptual Distance	β_{13}	CD1	0.673	CD1	0.137
	β_{14}	CD2	0.408	CD2	1.054
	β_{15}	CD3	-2.140	CD3 (β_{15})	0.403
	β_{16}	CD4	0.177		
Hybrid				Variant(β_{16})	0
	β_{17}	-	-	Field	-0.315

In the second phase, an algorithm and formula were presented. According to the values obtained for Co_p, Co_e, Co_g and threshold, our highest precision and recall values were 0.69 and 0.73 respectively. Since the second phase of Farreres' methodology needs a complete thesaurus to do the mapping, it is impossible to implement it on the languages like Persian. Therefore, we compared evaluation of our methodology and an adaptation of Farreres' methodology according to the parameters used by the methodologies under discussion.

With the comparisons above, our suggested formula (equation 3) transforms to equation 5 in the Farreres' methodology. Note that the GapValue of Farreres' methodology is computed by equation 4. In his methodology, definitions of level and nconn are different from those in ours, the coefficient $level_e$ is ignored and some specific points are not considered. After conducting many experiments, the values 0.60, 0.64 and 0.43 were obtained for parameters level, Co_g and threshold in Farreres'. According to the values of three parameters level, Co_g and threshold, the highest

precision and recall values in Farreres' methodology obtained by our suggested algorithm and the equation 5, were 0.63 and 0.62 while the highest values of precision and recall in our methodology were 0.69 and 0.73 respectively (Table 2).

$$FinalProb = \frac{\sum_{l_p=0}^{5}\left(Co_p * \left(6 - level_p\right)\right) - GapValue}{15} \quad (5)$$

Table 2. Comparison of Final precision and Recall of two methodologies.

	Threshold	Precision	Recall
SBU	0.21	0.69	0.73
SBU-based Farreres	0.43	0.63	0.62

5 Conclusion and Future Works

In this paper we proposed a methodology for mapping bilingual lexical ontologies based on Farreres' methodology. The source ontology is a thesaurus in a language and the destination one is English WordNet. Suggested methodology is language independent, however we used Persian thesaurus as the source ontology. This methodology is composed of two phases: a) mapping the words of source language to WordNet synsets and b) acceptance or rejection of associations between source thesaurus senses and WordNet synsets using the associations of first phase. The recall values we obtained during the first phase of our methodology were higher than those obtained in Farreres' but precision was almost the same.

In the first phase, we 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. In this phase, we obtained a formula whose input is an association between Persian word and WordNet synset and whose output is correctness probability of this association. In the second phase, we took advantage of associations between hypernym senses and hypernym synsets of a given base association. Actually, the similarities of concepts between two branches of thesaurus and WordNet hierarchies helped us decide which synset(s) are the most appropriate candidate(s) for a given sense of source language. At the end, the values 0.69 and 0.73 were obtained for precision and recall respectively.

For the future works, we will investigate if the second phase can be done applying the Bayesian Network which gives a formula to compute the final correctness probability of an StS (sense-to-synset association) according to its hypernyms associations.

References

1. Daudé, J., Padró, L., Rigau, G.: 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).

2. Daudé, J., Padró, L., Rigau, G.: Mapping WordNets using structural information. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China (2000).
3. Lee, C., Lee, G., Jung Yun, S.: 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).
4. Mihaltz. M., Prószéky. G.: Results and Evaluation of Hungarian Nominal WordNet v1.0. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, Czech Republic (2004).
5. Lebart, L.: Traitement Statistique des Données. DUNOD, Paris (1990).
6. Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Antonia Martí, M.: Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. Proceedings of the 6th Conference on Language Resources and Evaluation LREC2008. Marrakech, Morocco (2008).
7. Farreres, J., Gibert, K., Rodríguez, H.: Semiautomatic creation of taxonomies. In G. N. et al., editor, Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks', Taipei, (2002).
8. Farreres, J., Gibert, K., Rodríguez, H.: Towards binding Spanish senses to WordNet senses through taxonomy alignment. In S. et al., editor, Proceedings of the Second International WordNet Conferenc, pages 259–264, Brno. Masaryk University, (2003).
9. Farreres, J., Rodríguez, H.: Selecting the correct synset for a Spanish sense. In Proceedings of the LREC international conference, Lisbon, Portugal (2004).
10. Farreres, J.: Automatic Construction of Wide-Coverage Domain-Independent Lexico-Conceptual Ontologies. PhD Thesis, Polytechnic University of Catalonia, Barcelona (2005).
11. Atserias, J., Climent, S., Farreres, X., Rigau, G., Rodriguez, H.: 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).
12. Shamsfard M., Towards Semi Automatic Construction of a Lexical Ontology for Persian, Proceedings of the 6th Conference on Language Resources and Evaluation LREC2008. Marrakech, Morocco (2008).
13. Assi, M., Aryanpour, M.: Aryanpour English-Persian and Persian-English dictionary. http://www.aryanpour.com
14. Anvari, H.: Persian-Persian Sokhan dictionary. Sokhan Pub., Iran, Tehran (2002).
15. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1):17–30, (1989).
16. Simpson, T., Dao, T.: WordNet-based semantic similarity measurement. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx
17. Fararooy, J.: Fararooy Persian Corpus. http://www.persianthesaurus.com
18. Ramanand, J., Ukey A., Kiran Singh, B., Bhattacharyya, P.: Mapping and Structural Analysis of Multi-lingual WordNets. IEEE Data Eng. Bull. 30(1): 30-43 (2007).