Unsupervised Learning for Persian WordNet Construction

Mortaza Montazery
NLP Lab, School of ECE, College of
Engineering, University of Tehran,
Tehran, Iran
Mortaza.gh@gmail.com

Heshaam Faili
NLP Lab, School of ECE, College of
Engineering, University of Tehran,
Tehran, Iran
hfaili@ut.ac.ir

Abstract

In this paper we introduce an unsupervised learning approach for WordNet construction. The whole construction method is an Expectation Maximization (EM) approach which uses Princeton WordNet 3.0 (PWN) and a corpus as the data source for unsupervised learning. The proposed method can be used to construct WordNet in any language. Links between PWN synsets and target language words are extracted using a bilingual dictionary. For each of these links a parameter is defined that shows probability of selecting PWN synset for target language word in corpus. Model parameters are adjusted in an iterative fashion. In our experiments on Persian language, by selecting 10% of highly probable links trained by the EM method, a Persian WordNet was obtained that covered 7,109 out of 11,076 distinct words and 9,427 distinct PWN synsets with a precision of more than 86%.

1 Introduction

One of the most important challenges with respect to Natural Language Processing is the existence of ambiguity in different levels of natural language. Word sense ambiguity is one of these ambiguities. One solution for dealing with these problems is to generate knowledge repositories where human knowledge about natural language can be encoded. WordNet is a rich repository of knowledge about words that has been constructed to deal with word sense ambiguity problem.

The first WordNet was constructed for English language in Princeton University under direction of George A. Miller (Fellbaum, 1998). English words in four categories noun, verb, adjective and adverb have been grouped into sets of cognitive synonyms that are called synset. By proving of usefulness of Princeton WordNet (PWN), construction of WordNet for other languages has been considered. Two great efforts in constructing WordNet for other languages are EuroWordNet (Vossen, 1999) and BalkaNet (Tufis¸, Cristea, & Stamou, 2004). The former deals with European's languages such as English, Dutch, German, French, Spanish, Italian, Czech and Estonian. The second one deals with languages from Balkan zone such as Romanian, Bulgarian, Turkish, Slovenian, Greek and Serbian.

Manual construction of WordNet is a time consuming task and requires linguistic knowledge. The estimation of the average time for building a lexical entry depends on the polysemy of the words in the synsets, on the available lexical resources and definitely on the WordNet building tools. Thus automated approaches for WordNet construction or enrichment have been proposed to facilitate faster, cheaper and easier development. In this way several automatic methods have been proposed for constructing WordNet for Asian languages such as Japanese, Arabic, Thai and Persian that use PWN and other existing lexical resources.

In (Shamsfard M. , 2008) a semi-automated method has been proposed for developing a Persian lexical ontology called FarsNet. About 1,500 verbs and 1,500 nouns have been gathered manually to make WorldNet's core. Then two heuristics and a Word Sense Disambiguation (WSD) method have been used to find the most likely related Persian synsets. According to the first heuristic, a Persian word has only one synset if it is translated to a single English word that has only one sense in PWN. In this case no ambiguity exists for the Persian word whose one of synsets will be equivalent to that of English word. In other cases, second heuristic is used: if two translations of a Persian word have only one common synset then this common synset is linked to the

Proceedings of Recent Advances in Natural Language Processing, pages 302–308,
Hissar, Bulgaria, 12-14 September 2011.

Persian word. The existence of a single common synset implies the existence of a single common sense between the two words and therefore their Persian translations shall be connected to this synset (Shamsfard M. , 2008). For words whose English translations have more than one synset and the second heuristic could not find the appropriate synset, a WSD method has been used to select the correct synset. For each candidate synset, a score is calculated using the measure of semantic similarity and synset gloss words. Manual evaluation of the proposed automatic method in this research shows 70% correctness and covers about 6,500 entries on PWN.

In (Montazery & Faili, 2010), an automatic method for Persian WordNet construction based on PWN has been introduced. The proposed approach uses two monolingual corpora for English and Persian and a bilingual dictionary in order to make a mapping between PWN synsets and Persian words. In this paper, Persian words have been linked to PWN synsets in two different ways. Some links were selected directly by using some heuristics that recognize these links as unambiguous. Another type of links is ambiguous, in which a scoring method is used for selecting the appropriate synset. In order to select an appropriate PWN synset for ambiguous links, a score for each candidate synset of a given Persian word is calculated and a synset with maximum score is selected as a link to the Persian word. The manual evaluation on selected links on 500 randomly selected Persian words shows about 76.4% quality respect to precision measure. By augmenting the Persian WordNet with the unambiguous words, the total accuracy of automatically extracted Persian WordNet becomes 82.6%.

The automated approaches for WordNet construction vary according to the resources that are available for a particular language (Fišer, 2008). In (Fišer, 2008) multilingual parallel corpora have been used for the construction of Slovene WordNet. Their experiments were conducted on two different corpora. The first corpus contains five languages (English, Czech, Romanian, Bulgarian and Slovene), 100,000 words per language and it has already been sentence-aligned and tagged. The second corpus is the biggest parallel corpus of its size in 21 languages (about 10 million words per language) and it is paragraph-aligned but is not tagged, lemmatized, sentence or word-aligned. Both corpora have been sentence and word-aligned. Word-alignments have been used to create bilingual lexicons. For noise reduction purpose in the lexicon, only 1:1 links between words of the same part of speech have been taken into account and all alignments occurring only once have been discarded. Multilingual lexicon and already existing WordNet for each language have been used in order to construct Slovene WordNet. For English, PWN has been used while for Czech, Romanian and Bulgarian WordNets from the BalkaNet project have been used. For each lexicon entry synset ids from each WordNet are extracted and, if there is an overlap of synset ids across all languages, then it is assumed that the words in question all describe the concept marked with this id. Finally, the concept is extended to the Slovene part of the multilingual lexicon entry and the synset id common to all the languages is assigned to it (Fišer, 2008). Fišer (2008) also has extended her proposed method to include multi-word expression in generated Slovene WordNet.

There have been some other efforts to create a WordNet for Persian language (Shamsfard, et al., 2010; Mansoory & Bijankhan, 2008; Rouhizadeh, Shamsfard, & Yarmohammadi, 2008; Famian, 2007); but there exists no Persian WordNet yet that covers all Persian words in dictionary and is comparable with PWN.

In this paper, a fully automated language-independent unsupervised ML-based method for constructing a large-scale WordNet for any language is proposed. The method just needs some available resources such as PWN, machine readable dictionaries and monolingual corpus to train ontology for a target language. The approach implements an Expectation/Maximization (EM) algorithm which iteratively estimates the probability of selecting a candidate synset for a given target language word. Although the whole method is language-independent and it just works with the mentioned resources, we tested it on Persian language to retrieve a large-scale Persian WordNet.

The rest of the paper is organized as follows. Section 2 presents our method for constructing Persian WordNet automatically. Experimental results and evaluations of the proposed method are explained in section 3. Finally conclusion and future works are presented in section 4.

2 Persian WordNet Construction Method

The process is started by making an initial WordNet that consists of words in Persian language and the links between them and PWN synsets. Each Persian word may have several English translations and each English translation may also have several PWN synsets. Candidate synsets of a given Persian word are the union of all PWN synsets of its English translations. We think that each candidate synset of a given Persian word may be one of its probable senses. Our proposed method tries to estimate this probability. If a candidate synset represents a correct sense of Persian word, we expect the occurrence of this sense in a Persian corpus which contains that word.

For each Persian word *w* and each PWN synset *t*, $\theta_{w,t}$ is considered as probability of selecting PWN synset *t* for Persian word *w*. That is:

$$\forall w,t :\ \theta_{w,t} \in [0,1] \tag{1}$$

$$\forall w :\ \sum_{t} \theta_{w,t} = 1 \tag{2}$$

In order to estimate these parameters we can divide the number of times that a Persian word *w* occurs with PWN synset *t* in a Persian tagged corpus to the number of times that a Persian word *w* appears in that Persian tagged corpus. However, this simple method needs a Persian sense tagged corpus. Because, there is no such corpus, we use an EM method to estimate the probability of selecting a PWN synset for each Persian word of corpus. The idea is as follows: first we make a Persian WordNet with an initial value for the mentioned parameters, then for each word occurred in a Persian corpus the probability of selecting its senses is estimated using current value of parameters and words in context. Probabilities calculated in this step are used to update the parameters of the model.

The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values (Bilmes, 1998). Consider a sequence of Persian word w_1^n with length *n* and its corresponding sense tag sequence t_1^n. Assuming the independence between each pair of (w_i, t_i) we have:

$$\begin{aligned} P(w_1^n, t_1^n|\Theta) &= \prod_i P(w_i, t_i|\Theta) \\ &= \prod_{(w,t)\in(w_1^n,t_1^n)} P(w,t|\theta)^{n(w,t)} \\ &= \prod_{(w,t)\in(w_1^n,t_1^n)} {\theta_{w,t}}^{n(w,t)} \end{aligned} \tag{3}$$

Where Θ is the set of all parameters $\theta_{w,t}$ and $n(w,t)$ represents the number of times that word *w* appears with sense tag *t* in word-sense tag sequence (w_1^n, t_1^n). Log-likelihood function $L(\Theta)$ is defined as below:

$$\begin{aligned} L(\Theta) &= \log P(w_1^n, t_1^n|\Theta) \\ &= \sum_{(w,t)\in(w_1^n,t_1^n)} n(w,t) * \log \theta_{w,t} \end{aligned} \tag{4}$$

Because there is no such sense tagged corpus, we assume these tags to be hidden variables and the surface words to be observations. The EM algorithm first finds the expected value of the log-likelihood function with respect to the unknown data T_1^n given the observed data w_1^n and the current parameter values. This expected value is shown with $Q\left(\Theta, \Theta^{j-1}\right)$ and is calculated as follows:

$$\begin{aligned} Q\left(\Theta, \Theta^{j-1}\right) &= E\left(L(\Theta)\middle|w_1^n, \Theta^{j-1}\right) \\ &= \sum_{T_1^n} L(\Theta) * P(T_1^n|\, w_1^n, \Theta^{j-1}) \end{aligned} \tag{5}$$

Where Θ^{j-1}stands for the current parameters value that we use to evaluate the expectation and Θ is the new parameters value that we optimize to increase Q. The second step (the M-step) of the EM algorithm is used to maximize the expectation value which was computed in the first step. That is, we find:

$$\Theta^j = argmax_{\ \Theta} \left(Q\left(\Theta, \Theta^{j-1}\right)\right) \tag{6}$$

In order to maximize $Q(\Theta, \Theta^{j-1})$ subject to constraint has shown in formula (2), we introduce the Lagrange multiplier λ and to find the expression for $\theta_{w,t}$, we should to solve the following equation:

$$\frac{\partial}{\partial \theta_{w,t}} \left[Q(\Theta, \Theta^{j-1}) - \lambda(\sum_{t'} \theta_{w,t'} - 1) \right] = 0 \tag{7}$$

Whit solving differential equation (7), we obtain the new value of parameters as follows:

$$\theta_{w,t}^{j} = \frac{\sum_{T_1^n\ s.t.\ t\in T_1^n} \left(n(w,t) * P\left(T_1^n\middle|w_1^n, \Theta^{j-1}\right)\right)}{\sum_{t'} \sum_{T_1^n\ s.t.\ t'\in T_1^n} \left(n(w,t') * P(T_1^n|w_1^n, \Theta^{j-1})\right)} \tag{8}$$

However, in order to calculate new estimation of parameters, according to the formula (8) we must iterate over all possible sense tagged sequences T_1^n for Persian word sequence w_1^n. But the number of such sense tagged sequences is exponential with respect to the length of sequence. In this step we assume that the probability of assigning a sense tag t for word w_i is dependent only on w_i and other surrounding words in the sequence and is independent from the sense tags of other neighboring words. By this assumption, we simplify formula (8) as follows:

$$\theta^{j}_{w,t} = \frac{\sum^{n}_{\substack{j=1 \\ w_{j=w}, t_j=t}} P\left(t_j = t \middle| w_1^n, \theta^{j-1}\right)}{n(w)} \qquad (9)$$

The formula (9) implies that the probability of assigning sense tag t to word w is equal to average of conditional probability $P\left(t\middle|w_1^n, \Theta^{j-1}\right)$ $P\left(t\middle|w_1^n, \theta^{j-1}\right)$ over different occurrences of w in w_1^n. For applying the formula, a method to estimate the mentioned conditional probability is required. This method can be regarded as a WSD method which will be described in section 2.2.

2.1 Model Initialization

As in iterations of EM methods is guaranteed to increase the log likelihood function of observed data but there is no guarantee that the method converge to a maximum likelihood estimator (Bilmes, 1998). Depending on starting values, the EM method may converge to a local maximum of the observed data likelihood function. So, in our experiments initial value of $\theta_{w,t}$ has been initiated as follows.

FarsNet is the first published WordNet for Persian language that organized about 18,000 Persian words in about 10,000 synsets. Table 1 shows some statistics about FarsNet. For about 6,500 synsets in FarsNet equivalent synset in PWN have been identified. We have used these synsets for initializing model parameters.

	#Words	#Synsets
Noun	9,351	5,180
Adjective	3,935	2,526
Verb	4,380	2,305
Total	17,046	10,011

Table 1: Statistics of FarsNet

Suppose Persian word w has n candidate synsets such that m candidate synsets between them are equivalent with m synsets of w in FarsNet. With these assumptions $\theta_{w,t}$ is initiated as follows.

$$\theta_{w,t} = \begin{cases} \dfrac{1 + n\alpha}{n + nm\alpha}, if\ t\ is\ between\ m\ synset \\ \dfrac{1}{n + nm\alpha}, \qquad otherwise \end{cases}$$

In our experiments we used value 0.05 for parameter α.

2.2 Word Sense Disambiguation

WSD is the task of selecting the correct sense for a word in a given context. WSD methods can be classified into two types: supervised and unsupervised methods (Agirre & Edmonds, 2007). The former uses statistical information gathered from training on a corpus that has already been semantically disambiguated. Unlike supervised methods that require sense-tagged corpus, unsupervised methods just use a raw corpus and don't need any annotated data. Based on the types of used resources, unsupervised methods are classified into the following methods: raw corpus-based, dictionary-based and knowledge-based (Agirre & Edmonds, 2007).

In order to identify the sense of each word of corpus according to the initial Persian WordNet, knowledge based methods have been used. In (Agirre & Edmonds, 2007), three categories of knowledge based methods which use WordNet as their source of knowledge have been described: WordNet gloss based, conceptual density based and relative based. A gloss is a definition of synset in WordNet; WordNet gloss based approach is similar to dictionary based approach. However because our initial Persian WordNet does not have Persian gloss, this approach can not been applied to generate Persian sense-tagged corpus. Conceptual distance among the senses of a word in a context is used in conceptual density based approaches. In these approaches sense with shortest conceptual distance from words of context is selected. A conceptual distance is usually defined as the number of links between two concepts in a hierarchical lexical database such as WordNet or a thesaurus. In WordNet several relations between synsets and words are defined such as synonym, hypernym and hyponym. Relative based approaches use these relations to extract the relatives of each polysemous word from WordNet for WSD.

In our experiments a relative based WSD method similar to the one presented in (Seo, Chung, Rim, Myaeng, & Kim, 2004) has been used. In (Seo, Chung, Rim, Myaeng, & Kim, 2004) for a word in a context, a set of related words are extracted from WordNet and then the highest probable relative that can be substituted with the word in the context is chosen. In order to calculate the probability of selecting a relative, co-occurrence frequency has been used. Now consider Persian word w that occurred in the word sequence w_1^n and its sense correspond to PWN synset *t*. In our Persian WordNet there are other words that have the same PWN synset *t* in their candidate synsets. These words are synonyms of Persian word *w* with some probability that were estimated using parameter Θ. We consider a window around *w* and calculate the correlation of words linked to PWN synset *t* with words appeared in the window as a score of this sense in this context. That is:

$$Score(w,t) = \frac{\sum_{w'} \sum_{w^{"}} \theta_{w',t} * PMI(w', w^{"})}{n} \qquad (10)$$

In this formula, *w'* represents words that have *t* as their candidate synset and *n* is the number of such words and *w"* represents the words appeared in a window around *w*. This score is based on the idea that synonym words occurred in similar context and then maximum score is obtained for a sense whose linked words have highest association with the words of the context. In our experiments point-wise mutual information has been used in order to measure association between two words. Point-wise mutual information between two words *w* and *w'* is defined as follows:

$$PMI(w,w') = \log_2(\frac{P(w,w')}{P(w) * P(w')}) \qquad (11)$$

According to formula (10), we can define the probability of selecting sense tag t_i for word w_i in context w_1^n as follows:

$$P\left(t_i \middle| w_1^n, \Theta^{j-1}\right) = \frac{Score(w,t)}{\sum_{t'} Score(w,t')} \qquad (12)$$

The proposed EM method is repeated until the changes of probability of selecting a candidate synset for a Persian word becomes negligible.

3 Experiments and Evaluation

In order to generate initial Persian WordNet as mentioned in section 2, Aryanpour [1] Persian/English dictionary has been used to find equivalent English translations of each Persian word. Also, PWN version 3.0 was used to extract candidate synsets of Persian words.

In order to implement the E-step of proposed method we should select a Persian corpus and calculate the probability of selecting each candidate synset of Persian words using formula (10). To get better WSD result, we used an available POS-tagged Persian corpus instead of raw-corpus. Using this corpus has the benefit that formula (10) is calculated only for senses of word that have the same POS tag to those identified in the corpus and also candidate synsets of Persian words can be pruned according to their POS and appeared POS of Persian word. For this purpose Bijankhan POS-tagged corpus (BijanKhan, 2004) has been considered and all unique words that fall into three categories noun, adverb and adjective have been selected to generate initial Persian WordNet. Now consider Persian word *w* with POS tag *p* in Persian corpus. We want to calculate the probability of selecting each sense of *w* regarding its context. To do this, all senses of *w* in generated Persian WordNet that have POS *p* are extracted and their probabilities are calculated using formula (10). Probabilities of selecting other senses of *w* with different POS tags are considered to be zero in this context. Whereas words in corpus appear in inflected form, extraction of candidate synsets from our Persian WordNet may not perform properly. Thus in order to deal with this problem, before beginning our iterative method we performed a shallow stemming process for Persian on corpus. This process converts nouns to its singular form.

In order to calculate PMI between each pair of Persian words, Hamshahri text corpus has been used. Hamshahri is one of the online Persian newspapers in Iran that has been published for more than 20 years and its archive has been presented to the public. In (AleAhmad, Amiri, Darrudi, Rahgozar, & Oroumchian, 2009) this archive has been used and a standard text corpus with 318,000 documents has been constructed. In order to count the number of co-occurrences of two words *w* and *w'*, a window with the size of 20 words has been considered.

In our experiments, we used 1,000 documents as training data set. All unique words in corpus fall into just three categories noun, adjective and adverb and there exist entry for each of them in bilingual dictionary were selected to generate the initial Persian WordNet. In table 2 the number of PWN synsets covered by initial Persian WordNet using words in 1,000 documents has been shown.

POS	1,000 documents
Noun	22,988
Adjective	6,121
Adverb	480
Total	29,589

Table 2: Number of PWN synsets covered in initial Persian WordNet with respect to number of documents

[1] http://www.aryanpour.com/

Table 3 shows the number of words in initial Persian WordNet and number of their related candidate synsets. This table also shows average number of occurrence of words in documents.

	1,000 documents
# Words	11,076
# candidate synsets	111,919
Average number of occurrence	110.6

Table 3: Number of words and candidate synsets and average number of occurrence with respect to number of documents

The learning process will be iterated until the maximum changes in probabilities become less than a predefined threshold. In our experiments, we set the threshold to be 0.001. After the termination of EM algorithm, a WordNet in target language and the probabilities of selecting each candidate synsets to each word are acquired. Based on the threshold value has been set before, the model is converged to its final state after 73 iterations.

In order to evaluate the accuracy of trained WordNet, we generate a test set manually that contains 1365 randomly selected links between Persian words and PWN synsets. These links are manually divided into two categories: correct and incorrect. The number of links in each category with respect to the different POS tags has been shown in table 4. The average number of initial candidate synsets of words in this test set is about 66. It means that the words in this test set have high polysemy.

POS	Correct	Incorrect	Total
Noun	452	386	838
Adjective	300	87	387
Adverb	67	73	140

Table 4: Number of correct and incorrect links in test set

In figure 1, the curve indicating the relation between the precision and recall is shown. If we select the highest 10% of probable links as final Persian WordNet, the precision about 86.7% is achieved. In this case, the Persian WordNet contains 7,109 distinct words from 11,076 words appeared in corpus and covers 9,427 distinct PWN synsets. By selecting more links, less precision is retrieved. In the case of accepting all trained links after removing links with probability zero, the lowest precision, about 66%, is achieved.

In table 5, the mean average precision (MAP) over different recall rates with respect to different POS tags is shown. The highest precision is acquired for adjective which is 89.7% while the lowest precision is for noun, which is about 61%.

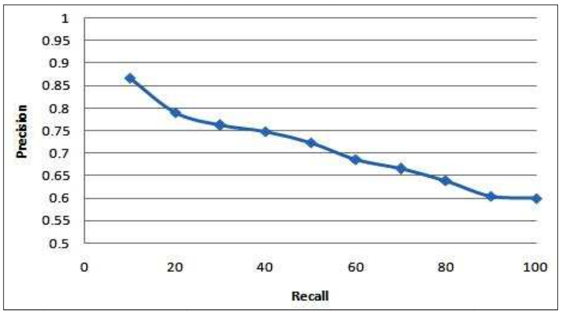

Figure 1: Recall Precision curve

	MAP
Noun	0.612
Adjective	0.897
Adverb	0.656

Table 5: Mean average precision with respect to different POS tags

4 Conclusion

In this paper we have presented a language-independent unsupervised EM method for automatically linking PWN synsets to Persian words using pre-existing lexical resources such as Persian text corpus, PWN and bilingual dictionary.

In the first step (E-step) of EM method, for each Persian word in corpus the probability of selecting each of its candidate synsets is calculated then these probabilities are used in the second step (M-step) to update probability of selecting candidate synsets of each Persian word. The final Persian WordNet is obtained by removing links those probabilities are less than a threshold or by selecting the top probable links as correct links. However the result of this method is dependent to the corpus that is used in E-step. In fact, the probability of selecting correct candidate synsets of a given Persian word that haven't appeared in corpus will be zero and these synsets will be ignored.

We guess that better results can be obtained by using more effective methods to initialize the parameter values rather than using FarsNet which may initialize some senses with higher values even if they have not even been observed in the corpus.

References

Agirre, E., & Edmonds, P. (2007). *Word Sense Disambiguation Algorithms and Applications.* Springer.

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Journal of Knowledge-Based Systems , 22* , 382-387.

BijanKhan, M. (2004). The Role of the Corpus in Writing a Grammar: An Introduction to a Software. *Iranian Journal of Linguistics, vol. 19, no. 2* .

Bilmes, J. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *ICSI*, (pp. 1-13). U.C. Berkeley.

Famian, A. A. (2007). Towards Building a WordNet for Persian Adjectives. *In Proceedings of the 3rd Global wordnet conference*, (pp. 307-309).

Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database.* Bradford Books.

Fišer, D. (2008). Using Multilingual Resources for Building SloWNet Faster. *The Fourth Global WordNet Conference*, (pp. 185-193). Szeged, Hungary.

Mansoory, N., & Bijankhan, M. (2008). The Possible Effects of Persian Light Verb Constructions on Persian WordNet. *The Fourth Global WordNet Conference*, (pp. 297-303). Szeged, Hungary.

Montazery, M., & Faili, H. (2010). Automatic Persian WordNet Construction. *the 23rd International conference on computational linguistics* (pp. 846-850). Beijing, China: Coling 2010 Organizing Committee.

Rouhizadeh, M., Shamsfard, M., & Yarmohammadi, M. A. (2008). Building a WordNet for Persian Verbs. *The Fourth Global WordNet Conference*, (pp. 406-412). Hungary.

Seo, H.-C., Chung, H., Rim, H.-C., Myaeng, S. H., & Kim, S.-H. (2004). Unsupervised word sense disambiguation using WordNet relatives. *ELSEVIER* , 253-273.

Shamsfard, M. (2008). Developing FarsNet: A Lexical Ontology for Persian. *The Fourth Global WordNet Conference*, (pp. 413-418). Szeged, Hungary.

Shamsfard, M. (2008). Towards Semi Automatic Construction of a Lexical Ontology for Persian. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).* Marrakech, Morocco.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., et al. (2010). Semi Automatic Development of FarsNet; The Persian WordNet. *5th Global WordNet Conference (GWA2010).* Mumbai, India.

Tufis¸, D., Cristea, D., & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology* .

Vossen, P. (1999). *EuroWordNet General Document.* Version 3 Final University of Amsterdam EuroWordNet LE2-4003, LE4-8328.