

Towards Semi Automatic Construction of a Lexical Ontology for Persian

Mehrnoush Shamsfard

NLP Research Laboratory,
Faculty of Electrical & Computer Engineering,
Shahid Beheshti University, Tehran, Iran
E-mail: m-shams@sbu.ac.ir

Abstract

Lexical ontologies and semantic lexicons are important resources in natural language processing. They are used in various tasks and applications, especially where semantic processing is evolved such as question answering, machine translation, text understanding, information retrieval and extraction, content management, text summarization, knowledge acquisition and semantic search engines. Although there are a number of semantic lexicons for English and some other languages, Persian lacks such a complete resource to be used in NLP works. In this paper we introduce an ongoing project on developing a lexical ontology for Persian called FarsNet. We exploited a hybrid semi-automatic approach to acquire lexical and conceptual knowledge from resources such as WordNet, bilingual dictionaries, mono-lingual corpora and morpho-syntactic and semantic templates. FarsNet is an ontology whose elements are lexicalized in Persian. It provides links between various types of words (cross POS relations) and also between words and their corresponding concepts in other ontologies (cross ontologies relations). FarsNet aggregates the power of WordNet on nouns, the power of FrameNet on verbs and the wide range of conceptual relations from ontology community

1. Introduction

In recent years, there has been an increasing interest in semantic processing of natural languages. Some of the essential resources to make this kind of process possible are semantic lexicons and ontologies. Lexicon contains knowledge about words and phrases as the building blocks of language and ontology contains knowledge about concepts as the building blocks of human conceptualization (the world model) (Shamsfard & Barforoush, 2003). Lexical ontologies or NL-ontologies are ontologies whose nodes are lexical units of a language. Moving from lexicons toward ontologies by representing the meaning of words by their relations to other words, results in semantic lexicons and lexical ontologies.

One of the most popular semantic lexicons for English is WordNet. Princeton WordNet (Fellbaum, 1998), is widely used in NLP research works. It covers English language and has been first developed by Miller in a hand-crafted way. Many other lexical ontologies (such as EuroWordNet, BalkaNet, ...) have been created based on Princeton wordnet for other languages such as Dutch, Italian, Spanish, German, French, Czech and Estonian. Although there are such semantic, lexical resources for English and some other languages, some languages such as Persian (Farsi) lack such a semantic resource for use in NLP works.

Persian is an Indo-European language, the official language of three countries (Iran, Afghanistan, and a part of Tajikistan), and it is also spoken in more than six countries.

There have been some efforts to create a wordnet for Persian language (Famian & Aghajaney, 2007; Keyvan, et al., 2007) but no available products have been announced yet. The only available lexical resources for Persian are some lexicons containing phonological and syntactic knowledge of words (such as (Eslami, 2006)).

On the other hand, the major problems with wordnet are:

(1) It has very restricted relations and does not allow defining arbitrary new ones.

(2) It has weak semantic knowledge on verbs. There is no information about verb arguments and their conceptual properties in WordNet.

(3) It does not support cross-POS relationships

In this paper we introduce an effort to develop a lexical ontology called FarsNet for Persian language which overcomes the above shortcomings. We exploit a semi automatic approach to acquire lexical and ontological knowledge from available resources and build the lexical ontology. FarsNet is a bilingual lexical ontology which not only represents the meaning of Persian words and phrases, but also links them to their corresponding concepts in other ontologies such as WordNet, Cyc and Sumo. FarsNet aggregates the power of WordNet on nouns, the power of FrameNet on verbs and a wide range of conceptual relations from ontology community.

2. Introducing FarsNet

FarsNet consists of two main parts: a semantic lexicon and a lexical ontology. Each entry in the semantic lexicon contains natural language descriptions, phonological, morphological, syntactic and semantic knowledge about a lexeme. The lexemes can participate in relations with other lexemes in the same lexicon or to entries of other lexicons and ontologies, in the ontology part. Here, the semantic lexicon is serving as a lexical index to the ontology. The ontology part contains not only the standard relations defined in WordNet but also some additional conceptual ones. FarsNet is able to add new relations for its words or concepts. We have developed an interface for FarsNet from which one can add, remove or change the entries. From this interface the user can define new relations or use the existing ones and relate words by them. It can relate words from different syntactic types

together (e.g. nouns to adjectives and verbs). It can also relate a word to its corresponding concept in an existing ontology. This makes the interoperability between various resources and various languages easier.

In addition there are some specific features for specific POS tags too. For instance, we have defined a new relation for adjectives which shows their selectional restrictions, the category of nouns who can accept this word as a modifier. For example 'khoshmazah' (delicious) usually is used for edibles while 'hajim' (voluminous, huge) is used for physical entities.

On the other hand FarsNet covers the relations introduced for verbs in wordnet and also adds the number, names and conceptual characteristics of the arguments of each verb (its selectional restrictions) in a similar way to FrameNet. For each verb, FarsNet contains the type and semantic category of its arguments (Shamsfard & SadrMousavi, 2007). For example the verb 'khordan' (to eat) belongs to a verb class which needs an agent and a theme and can have an instrument. It should be defined that the theme of this verb should be edible, its agent should be an animated thing, and the size of its instrument is small (usually smaller than a mouth) and it may be one of spoon, fork, knife,

These features help NLP systems to extract thematic roles, disambiguate syntactic parsing, chunk, represent the sentence meaning and acquire knowledge from texts.

3. Semi-automatic knowledge acquisition for FarsNet

We use an incremental approach to build FarsNet; developing a kernel and extending it in a semi automatic way. The acquisition approach consists of the following main steps:

- 1- Providing initial resources,
- 2- Developing an initial lexicon based on wordnet and performing WSD,
- 3- Extracting new knowledge from available resources,
- 4- Evaluation and refinement

We have the following resources available and use them to develop FarsNet.

- WordNet
- a (syntactic) Lexicon (Eslami, 2006) containing more than 50,000 entries with their POS tags,
- a bilingual (English- Persian) dictionary.
- Persian POS tagged corpora
- a morphological analyzer for Persian (Shamsfard, et al., 2007)

In the following subsections the next steps will be described.

3.1. Developing an initial lexicon based on WordNet

To develop an initial lexicon we exploited three separate approaches in parallel:

- (a) Manually gathering a small lexicon.
- (b) Automatic creation of a small kernel containing just the base concepts
- (c) Automatic creation of an initial big lexicon containing

almost anything covered by the bilingual dictionary

Manually creation of a small lexicon (option (a) above) using available resources and linguistic knowledge of team members was done for more than 1500 verbs (Rouhizadeh, et al., 2008) and 1500 nouns (Shamsfard, et al., 2007). Verbs were selected according to verbs occurring in BalkaNet base concepts and most frequent verbs of a Persian corpus. Nouns were selected sequentially from a Persian dictionary.

For (b) we started from English base concepts and translated them to Persian, but for (c) we moved in two directions, from English to Persian and from Persian to English separately to compare their results.

To move from English, for each English synset, first we translate all the words in the synset using an electronic bilingual dictionary. Then we should arrange the Persian synsets by exploiting some heuristics and WSD (word sense disambiguation) methods. It is obvious that each synset has some English words and each word may have several senses and each sense may have several translations to Persian. So creating Persian synsets from English ones is not a straight forward task and each Persian word may be connected to a group of synsets in WordNet. Therefore it is important to identify the right sense(s) of English word, the right translation of it and putting the right sense of translated word in the corresponding synset. This task in other WordNets has been done using a common upper ontology, e.g. ILLI (interlingual index) for EuroWordNet and SUMO for ArabicWordNet. At this step we exploited both heuristics and disambiguation methods to find the appropriate synsets. At next phases we will connect our concepts to other ontologies as well.

We use some heuristics to find the corresponding synsets fast. If a word is known to be the English equivalent of a Persian word according to the dictionary, the Persian word should at least be connected to one of the synsets that include the English word as a member. There will obviously be no ambiguities if the English word has only one sense and so appears at only one synset. In this case its translations will be added to that synset too.

In other cases, to find the appropriate Persian synset for an English one, we consider word pairs in the English synset. For each word in this pair we list all synsets they appear in. If those two words appear together only in the current synset, their common Persian translations would be connected to that synset. The existence of a single common synset in fact implies the existence of a single common sense between the two words and therefore their Persian translations shall be connected to this synset. In cases which there are more than one sense (one synset) for the English word we apply a disambiguation method to find the appropriate one. The method is described in the next subsection.

3.1.1 Word sense disambiguation

Our disambiguation procedure uses other English translations of Persian word PW (that are named EWs

later in this text) as context words. Similar to the Lesk algorithm (Lesk, 1986), our method uses the EW gloss, However the measures of the semantic similarity is chosen differently.

For every sense of EW, a score is calculated using its gloss and the context words: for every word that appears in the gloss a score is assigned and the score of choosing a sense is the sum of scores of its gloss words divided by the total number of them. In order to score the gloss words according to the context words we use the scoring algorithm that was introduced in (Pedersen et al., 2005) which is described in figure 1.

The algorithm computes a score for all senses of a word that appears in the gloss (called target word). The score of the target word is defined to be the maximum of the scores of its senses.

```
foreach sense sti of target word wt
{
  set scorei = 0
  foreach word wj in window of context
  {
    skip to next word if j == t
    foreach sense sjk of wj
    {
      Temp_score[j] = relatedness(sti,
      sjk)
    }
    winning score = highest score in array
    temp_score[]
    if (winning score > threshold)
      set scorei = scorei + winning score
  }
}
return scorei, such that scorei ≥ scorej ,
forall j, 1 ≤ j ≤ n,
n =number of words in sentence
```

Fig. 1. Scoring Pseudo Code (Pedersen et al. 2005)

In the scoring algorithm, a function named “relatedness” is being used for the calculation of semantic similarity between two concepts. There are many proposals for the measurement of semantic similarity between two concepts, we use the one proposed by Resnik (1998) which is based on shortest path length and takes into account the distance from each of the two concepts from the root as well as the shortest path length from their most specific common parent to the root.

A shared parent of two concepts is known as a subsumer. The least common subsumer (LCS) of two concepts is the one that does not have any children that are also the subsumer of two concepts. In other words, the LCS of two concepts is the most specific subsumer of them. This measure finds the distance to the root of LSC. The distance of the LCS is then divided by the sum of the distances of the individual concepts to the root. The measure is formulated as follows:

$$\text{sim}(c1, c2) = 2 * \text{depth}(\text{lcs}(c1, c2)) / (\text{depth}(c1) + \text{depth}(c2))$$

Where depth is the distance from the concept node to the root of the hierarchy.

3.2. Extracting new knowledge from available resources

After creating the initial lexicon, extra words will be gathered from a tagged corpus, and assign to a synset as mentioned before.

Another part of ontology learning in FarsNet is dedicated to finding some relations from corpora exploiting lexico-syntactic and semantic patterns (Shamsfard, 2007). Some of these templates are noun phrase templates which are defined to discover relations between different parts of a noun phrase. They are used to extract hyponymy, meronymy attribute-value and possession relations. They include adaptations of Hearst’ patterns (Hearst, 1992) for Persian, the exception template, the modification template and others to extract relations between various parts of a noun phrase (head and modifiers). At this phase we have used the templates to extract hyponymy relations. As an example we can mention the exception template as follows:

The Exception template:

... {all | every} NP0 except NP1 {(and | ,) NPi}* ... (i>1), implies that (sub-class NP_i NP₀) (for all i □ 1)

Our current main problem now, which causes the major part of the errors is the structural ambiguity of noun phrases in the above templates. For example in the following phrase:

... کاربرد انواع کتابهای مرجع مثل لغتنامه ها، دائره المعارف ها و نمايه ها...
(the application various types of reference books such as dictionaries, encyclopedias, indices, ...)

The NPs after ‘such as’ (dictionaries, encyclopedias, indices) will be hyponyms of the NP before it (the application of ...), while they should be hyponyms of the modifier NP (reference books).

This rules works properly in some other cases for example in the following phrase:

مطبوعات جهان عرب مانند الحيات
(The newspapers of Arab world such as Alhayat)

The NP after ‘such as’ (Alhayat) will be found as the hyponym of the NP before it (the newspapers of Arab world) which is a correct choice.

We plan to use modification templates to find relations between nouns and their possible adjectives in the next phases too. The adjective modification template implies that the adjective modifier should be added to the list of possible adjectives for the head. Refining the ontology may cause some categorization of these adjectives (or heads) and relate the head (or adjective) to a superclass of the adjectives (or head).

3.3. Evaluation and refinement

The final phase of the lexicon building life cycle is evaluation and refinement. As it was mentioned before, we build each part of FarsNet using more than one approach. The evaluation procedure is done by two

methods too.

In the first method a linguistic expert reviews the automatically extracted knowledge and confirms or corrects them according to valid Resources (manual evaluation). The manual evaluation of the part of lexicon built so far shows an accuracy of about 70% in the resulting Persian lexicon.

In the second method we compare the results of various exploited methods on a common task to find the common built knowledge. For example to confirm the inclusion hierarchies, we extract hierarchical relations from text using templates in one hand and find this hierarchy according to the hyponym/hypernym relations between corresponding English synsets on the other hand. Comparing the results shows the most confident knowledge extracted by both two methods.

4. Conclusion

FarsNet project is an ongoing project in NLP research laboratory of Shahid Beheshti University. Manually developing a small lexicon as the kernel of FarsNet (Shamsfard, et al., 2007), manually translating the base concepts of wordnet into Persian, automatic finding the corresponding WordNet synsets for each entry of the

syntactic lexicon and automatic acquisition of new words and relations from the tagged corpus using template driven methods are some of performed tasks. As a result of the performed tasks we have created a base lexicon containing 37000 correspondences between Persian words and English synsets and chose the entries which have ranked above the threshold (0.7) in our WSD ranking procedure. Some results of the ranking procedure are shown in table 1. On the other hand we selected the most frequent words (which are not appeared in the last experience) from the initial lexicon (Eslami, 2006) (with frequency more than 3000 in our corpus) and attached them to the synsets containing their English translations. This results in 16800 new correspondences which a part of it is shown in table 2. The manual evaluation methods show about 70% correctness in our automatic approaches. There are some further works to complete the project such as completing the verbs knowledge base Exploiting (or linking to) FrameNet, enhancing the sense disambiguation modules in the automatic translations, improving the semantic templates to extract non-taxonomic relations from text and establishing a mapping between various ontologies.

synsetno	Definition	PersianWord	Score
104143	capable of thinking and expressing yourself in a clear and...	درخشان	0.35000002384...
94815	lighted with red light as if with flames	درخشان	0.35076922178...
104616	clear and sharp and ringing	درخشان	0.33666667342...
85674	(of surfaces) make shine	درخشش	0.83000004291...
90281	glitter as if covered with spangles	درخشش	0.86111110448...
26784	the visual property of something that shines with reflecte...	درخشش	0.85181820392...
26102	ceremonial elegance and splendor	درخشش	0.80777776241...
38935	the occurrence of a small flash or spark	درخشش	0.85222220420...
38935	the occurrence of a small flash or spark	درخشندگی	0.79899996519...
95480	richly and brilliantly colorful	درخشنده	0.30799999833...
94264	very favorable or advantageous	درخشنده	0.24250000715...
49082	any celestial body visible (as a point of light) from the Ear...	درخشیدن	0.76277774572...
93152	emit light	درخشیدن	0.84538459777...
26784	the visual property of something that shines with reflecte...	درخشیدن	0.88499999046...
81174	make lighter or brighter	درخشیدن	0.57357144355...
19528	a piece of furniture holding one or more electric light bulbs	درخشیدن	0.76882356405...
26784	the visual property of something that shines with reflecte...	درخشیدن	0.88499999046...
26769	an indication of radiant light drawn around the head of a ...	درخشیدن	0.78529411554...
38935	the occurrence of a small flash or spark	درخشیدن	0.78428572416...
60041	a column of light (as from a beacon)	درخشیدن	0.81857138872...

Table (1) – part of the results of the ranking procedure

synsetno	Definition	PersianWord
0	that which is perceived or known or inferred to have its own distinct existence ...	وجود
0	that which is perceived or known or inferred to have its own distinct existence ...	وجود
1	a separate and self-contained entity	کار
2	a thing of any kind	شیچ
4	a nonexistent thing	شیچ
5	an assemblage of parts that is regarded as a single entity	بخش
5	an assemblage of parts that is regarded as a single entity	خانه
5	an assemblage of parts that is regarded as a single entity	میزان
5	an assemblage of parts that is regarded as a single entity	واحد
7	a living thing that has (or can develop) the ability to act or function independe...	سازمان
7	a living thing that has (or can develop) the ability to act or function independe...	وجود
7	a living thing that has (or can develop) the ability to act or function independe...	وجود
13	any entity that causes events to happen	جهت
14	a human being	تك
14	a human being	تن
14	a human being	نفر
14	a human being	نهاد
14	a human being	وجود

Table (2)- A part of correspondences created automatically

5. Acknowledgements

I would like to thank Miss Negar Hariri for providing some of the tests and evaluations.

6. References

- Eslami, M. (2006). The generative lexicon, In: *2nd workshop on Persian language and computer*, Tehran.
- Famian, A., Aghajaney, D. (2007). Towards Building a WordNet for Persian Adjectives, In: *3rd Global wordnet conference*.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, Mass. MIT press.
- Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*.
- Keyvan, F., Borjjan, H., Kasheff, M., Fellbaum, C. (2007). Developing PersiaNet: The Persian Wordnet, In: *3rd Global wordnet conference*.
- Lesk, M (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone, in: *Proceedings of the 5th annual international conference on Systems documentation*, ACM Press, pp. 24–26.
- Pedersen, T. Banerjee, S. Patwardhan, S. (2005). Maximizing Semantic Relatedness to Perform Word Sense Disambiguation, *Supercomputing*
- Resnik, P. (1998). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, *Journal of Artificial Intelligence Research* 11.
- Rouhizadeh, M., Shamsfard M., Yarmohammadi, M. (2008). Building a WordNet for Persian Verbs, *The Fourth Global WordNet Conference*, Hungary.
- Shamsfard, M., Barforoush, A.A. (2004). Learning Ontologies from Natural Language Texts. In: *International Journal of Human-Computer Studies*, vol. 60, pp.17-63.
- Shamsfard, M., SadrMousavi, M. (2007). A Rule-based Semantic Role Labeling Approach for Persian Sentences, In: *Second workshop on Computational Approaches to Arabic-script Languages (CAASL'2)*, Stanford, USA.
- Shamsfard, M., Mirshahvalad, A., Pourhassan, M., Rostampour, S. (2007). Developing basic analysers for Persian: combining morphology, syntax and semantic, In: *15th Iranian conference on Electrical Engineering*, Tehran.
- Shamsfard, M. (2006). Introducing Linguistic and Semantic Templates for Knowledge Extraction from Texts, In: *workshop on ontologies in text technology*, Germany