Automatic Persian WordNet Construction

Mortaza Montazery
School of Electrical and Computer Engineering
College Engineering, University of Tehran
Mortaza.gh@gmail.com

Heshaam Faili
School of Electrical and Computer Engineering
College Engineering, University of Tehran
hfaili@ut.ac.ir

Abstract

In this paper, an automatic method for Persian WordNet construction based on Princeton WordNet 2.1 (PWN) is introduced. The proposed approach uses Persian and English corpora as well as a bilingual dictionary in order to make a mapping between PWN synsets and Persian words. Our method calculates a score for each candidate synset of a given Persian word and for each of its translation, it selects the synset with maximum score as a link to the Persian word. The manual evaluation on selected links proposed by our method on 500 randomly selected Persian words, shows about 76.4% quality respect to precision measure. By augmenting the Persian WordNet with the un-ambiguous words, the total accuracy of automatically extracted Persian WordNet is about 82.6% which outperforms the previously semi-automated generated Persian WordNet by about 12.6%.

1 Introduction

In Natural Language Processing (NLP) wide coverage lexical databases are used in different area such as information retrieval and cross-language information retrieval. WordNet is an example for a lexical database that groups words into sets of synonyms and categorizes them in four categories: noun, verb, adjective and adverb and records various relations between synonym sets. A broad overview of the different PWN applications such as "Machine Translation", "Information Retrieval", "Document Classification", "Query Answering" and "Conceptual Identification" have been presented in (Morato et al., 2004).

PWN was created and maintained since 1990s. After this WordNet for other languages have been under development and new projects start every year. PWN database contains about 150000 words organized in over 115000 synsets.

Manual construction of WordNet is a time consuming task and requires linguistic knowledge. A number of automatic methods were proposed for constructing WordNet for other languages that use PWN and other existing lexical resources. In order to help the development of WordNets for other languages rather than English, especially for European one, a project named EuroWordNet was found (Vossen, 1999), in which a number of automatic methods for construction of such databases were proposed (Farreres et al., 1998).

There have been some other efforts to create a WordNet for Persian language (Famian, 2007; Rouhizadeh et al., 2008; Shamsfard, 2008) but there exists no Persian WordNet yet that covers all Persian words in dictionary and comparable with PWN. These projects have tried to construct Persian WordNet in the manually or semi automatic manner. In (Shamsfard, 2008) a semi automatic method is proposed in which for each Persian word a number of PWN synsets are suggested by system in order to be supervised by a human annotator to select a relevant synset. Based on experiments mentioned by Shamsfard (2008), the proposed WordNet extracted automatically by the system, retrieved about 70% accuracy.

In this paper a fully automatic method for constructing a large-scale Persian WordNet from available resource such as PWN, MRDs and corpora has been proposed. Our approach uses different word similarity metrics like mutual information and WordNet similarity to map Persian words to appropriate PWN synsets.

2 Related Works

In the related field of automatic and semi automatic WordNet construction, several efforts
have been made. In (Shamsfard, 2008) a semi-
automatic method has been used for developing a
lexical ontology called FarsNet for Persian
language. About 1500 verbs and 1500 nouns have
been gathered manually to make WorldNet's
core. Then some heuristics and Word Sense
Disambiguation (WSD) methods 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’s be translated to a single
English word. In this case no ambiguity exists for
the Persian word whose one of synsets will be
equivalent with that of English word. In other
cases, second heuristic is used: if two translations
of a Persian word have only one common synset
then for the Persian word this common synset is
selected. 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 (Shamsfard, 2008). For
words whose English translations have more than
one synset and second heuristic cannot find the
appropriate synset, WSD methods have been
used to select 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 6500 Entries on WordNet.

In (Sathapornrungkij and Pluemtipitwiwiriyawej,
2005) a semi-automatic approach has been
described to construct the Thai WordNet lexical
database from WordNet and LEXiTTON
machine readable dictionaries. Thai WordNet
synsets have been derived from the PWN. The
candidate links between Thai words and synsets
have been derived from semantic links which are
obtained from WordNet and the translation links
which are obtained from LEXiTTON. In order to
derive links between Thai words and PWN
synsets, 13 criteria have been used which are
categorized into three groups: monosemic,
polysemic and structural criteria. Monosemic
criteria focus on an English word which has only
one meaning. Such English word has one synset
in PWN. Polysemic criteria focus on an English
word which has multiple meaning. Such English
word has multiple synset in PWN. Structural
criteria focus on the structural relations among
synsets with respect to WordNet 1.7. In order to
verify links that constructed using these 13
criteria, stratified sampling technique has been
applied and for each criterion 400 links have been
verified manually. The results of verification
show that the best criterion has 92% correctness
and the lowest correctness is equal 49.25%.

In PWN, there is a gloss for each synset that
can be used in automatic WordNet construction.
In (Kaji and Watanabe, 2006) this information
has been used for automatic construction of
Japanese WordNet. Given an English synset, it
calculates the score for each of its Japanese
translation candidates according to the gloss
appended to the synset. A pair of words is called
associated if mutual information between them be
larger than a threshold. The score is defined as
the sum of correlations between the translation
candidate and the associated words appearing in
the gloss. Whereas availability of bilingual
corpora is limited, for calculating pair wise
correlation between the Japanese translations of
an English word and its associated words an
iteratively approach has been proposed that
calculate this correlation without using bilingual
corpora.

In (Lee et al., 2000) a set of automatic WSD
techniques have been described for linking
Korean words collected from a bilingual MRD to
PWN synsets. For a given synset, 6 individual
heuristic scores are calculated and then a decision
tree is used to combine these scores to classify
the synset as linking or discarding. In order to
make the decision tree, a set of synsets have been
labeled manually as linking or discarding and
corresponding heuristic scores have been
calculated and then used for training data set. To
evaluate the accuracy of proposed method the
candidate synsets of 3260 senses of Korean
words have been classified manually as linking or
discarding. This test set has been used to
calculate precision of each heuristic. The results
of experiments show that the precision of all
heuristics is better than random mapping and the
best heuristic have 75.21% precision. The
combination of heuristics using decision tree
shows 93.59% precision.

3 Automatic Persian WordNet Con-
struction
Each Persian word can have several English
translations and each English translation has also
several PWN synsets. For a given Persian word, a bilingual dictionary is used to extract English equivalent words, and then a set of candidate synset is generated using PWN that contains all synsets of English translations of Persian word. As in (Shamsfard, 2008), if the English translation of a given Persian word has only one synset in PWN, then the Persian word is linked to this PWN synset directly, or if for a candidate synset at least two English translations belong to it, then Persian word is linked to this PWN synset.

In other cases, a score is calculated for each remaining candidate synset and the synset with maximum score is selected as an appropriate synset of the Persian word. Note that after selecting a synset, all synsets that share English words are removed from candidate synsets.

The following resources have been used in the process of score calculation:

- PWN: synset words, synset definition and hyponymy relations have been used.
- Bilingual dictionary (Persian – English)
- Raw Persian text corpus for extracting related words of a given Persian word
- Raw English text corpus for extracting mutual information between English words

Text corpora have been used to extract the related words of any given word. To do this, Mutual Information (MI) metric between any words in corpus and given Persian word are calculated and n-best words with higher MI values are selected. Mutual Information of pair x and x’ is defined as follows:

\[
MI(x, x') = \frac{n(x, x')}{n(x) * n(x')} * N
\]  

(1)

In formula 1, n(x, x’) is co-occurrence frequency of x and x’ in corpus. This frequency is calculated using a window with specific size. n(x) is the frequency of word x in corpus and N is the number of unique words in corpus.

So, in order to select the most related words for a given Persian word, an additional step is considered. For each Persian word w, other related Persian words with highest mutual information are selected and considered as a set R = {r1, r2, ..., rn}. Then for each Persian word rj a similar process is used and a set of words is extracted that is called Rj. If Rj contains the word w, then rj is selected as the related word for w and otherwise discarded.

After extracting the related words of the given Persian word, a Persian to English dictionary has been used to find equivalent translation of each related word. These words are referred as Related Translation Set (RTS). In scoring algorithm words that appear in gloss of each synset and words that appear in hyponym synset are called Gloss Words (GW). These words are considered as related words to the candidate synset and distinguish each synset from other.

Now for each candidate synset of a given Persian word a score is calculated that is based on the idea that two related words in the two-side languages share the same words in the correlation set. That is, if Persian word w relates to English synset e, then other co-related Persian words r1, r2, ..., rn which have gotten the best MI respect to w, should be related to the same synset e again.

Based on the above notion, the score of each candidate synset S can be estimated as follow:

\[
Score(S) = \sum_{w \in RTS} \sum_{e \in EW} Sim(w, S) * MI(w, e)
\]  

(2)

The score of synset S is defined as summation on product of semantic similarity between words in RTS and synset S, and mutual information between words in RTS and words in GW. In (Pedersen et al., 2004) several methods for calculating semantic similarity based on WordNet’s structure have been presented. Some of these methods are based on path lengths between concepts and some of them are based on information content. One of these methods is named path in which for each word w and synset s is defined as inverse of shortest path length between any synset of w and s. In our experiments the measure path has been used and calculated using formula 3.

\[
Sim(w, S) = \frac{1}{\min_{s_i \in \text{synsets of } w} \text{path}(s_i, S)}
\]  

(3)

In formula 2 the words from RTS which has less similarity to synset S has little effect on the amount of score in synset.

848
4 Experiments and Evaluations

Persian WordNet constructor components are Word Translator, Related Word Extractor, Synset Extractor and Synset Selector. Persian words and their selected synsets are input and output of this system. Persian word is given as input to the Word Translator and Related Word Extractor components. In our experiment, 10 words with highest MI to the given Persian word are extracted using Related Word Extractor. For this purpose, 3000 documents of IRNA\textsuperscript{1} newspaper text corpus have been used. IRNA is a news agency published their news on different languages, mainly on Persian. In order to count the number of co-occurrences of words x and x', a window with the size of 20 words was considered. Translations of related words and candidate synsets are given to Synset Selector and appropriate synsets for the given Persian word are selected. In this step PWN is used for semantic similarity calculation and an English text corpus (USENET corpus) is used to calculate mutual information. Table 1 shows the number of words and documents in the Persian and English text corpora. About 30698 Persian words from Aryanpour\textsuperscript{2} Persian to English dictionary has been used for constructing Persian WordNet.

<table>
<thead>
<tr>
<th></th>
<th>Num of documents</th>
<th>Num of Unique Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persian</td>
<td>3000</td>
<td>32197</td>
</tr>
<tr>
<td>English</td>
<td>3000</td>
<td>32899</td>
</tr>
</tbody>
</table>

Table 1: number of documents and unique words in Persian and English corporuses

As it was mentioned in the previous section, Persian words were linked to PWN synsets in the two different ways. Some links was selected directly without calculating their score by using some heuristics. We call these links as unambiguous links. Some of these links are shown in table 2. As it shown in the table, unambiguous links are wrong in some cases. For example in the case of '<barchasb>tag', a verb synset is selected while the Persian word is noun, so the selection is judged as incorrect. If the part of speech tag information of word is used in this example the correct synset would be selected.

Another type of links are ambiguous links, in which a scoring method is used for selecting the appropriate synset. Examples of these links are shown in table 3. As it’s shown in the table, the word '<karmozd>commission' has been linked to 6th sense of word 'commission' that is wrong. In constructed Persian WordNet also word '<farman>commission' has been linked to this sense of word 'commission' but the word '<karmozd>commission' and the word '<farman>commission' have less similarity together. In this example link between '<farman>commission' and 6th sense of word 'commission' is an unambiguous link. Therefore we can avoid of selecting this synset for '<karmozd>commission' using this information.

In order to evaluate the quality of the selected links, 500 Persian words have been randomly selected and the accuracy of selected synsets has been evaluated manually. Table 4 summarizes the results of this evaluation. As it’s shown in the table, the precision of unambiguous links is about 95.8% while this precision is 76.4% for the ambiguous links. The weighted average precision of the whole links in our automatically generated Persian WordNet is 82.6%, which outperforms the only comparable semi-automated Persian WordNet which was previously presented by (Shamsfard, 2008), about 12.5%. Also, by comparing the PWN coverage rate of these Persian WordNets, it reveals that our result covered 29716 entries on PWN which it is about 4 times more than the previously generated Persian WordNet.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous links</td>
<td>95.8%</td>
</tr>
<tr>
<td>Ambiguous links</td>
<td>76.4%</td>
</tr>
<tr>
<td>All links</td>
<td>82.6%</td>
</tr>
</tbody>
</table>

Table 4: accuracy of selected links for 500 words

The experimental results reveal that in PWN there is a short gloss for some synsets which makes the calculated score for those synsets to be lower than other candidate synsets of a given Persian word. This problem can be overcome by normalizing the scores of candidate synset of a given Persian word, i.e. by dividing the score of each synset by the number of words in GW. Another solution of this problem is proposed by (Kaji and Watanabe, 2006).

---

\textsuperscript{1} Islamic Republic News Agency (http://www.irna.ir)
\textsuperscript{2} http://www.aryanpour.com/
Table 2: Examples of unambiguous links

<table>
<thead>
<tr>
<th>Persian word</th>
<th>English translation</th>
<th>Selected synset</th>
<th>Gloss</th>
<th>Correct/incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>mosen aged</td>
<td>aged, elderly, old</td>
<td>aged, elderly</td>
<td>people who are old collectively</td>
<td>correct</td>
</tr>
<tr>
<td>barchasb tag</td>
<td>tag, label, mark</td>
<td>tag, label, mark</td>
<td>attach a tag or label to</td>
<td>incorrect</td>
</tr>
</tbody>
</table>

Table 3: Examples of ambiguous links

<table>
<thead>
<tr>
<th>Persian word</th>
<th>English translation</th>
<th>Selected synset</th>
<th>Gloss</th>
<th>Correct/incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>enteshar publication</td>
<td>publication</td>
<td>publication</td>
<td>the communication of something to the public; making information generally known</td>
<td>correct</td>
</tr>
<tr>
<td>karmozd commission</td>
<td>commission, charge, direction</td>
<td>a formal statement of a command or injunction to do something</td>
<td>incorrect</td>
<td></td>
</tr>
</tbody>
</table>

nabe, 2006), the gloss is given as a query to text retrieval engine and the words that appear as the answer of this query are used instead of the words of gloss. In our experiments, the first solution is chosen which retrieved the results shown in table 4.

5 Conclusion

This paper explored a method for automatically linking WordNet synsets to Persian words using pre-existing lexical resources such as Persian and English text corpora and PWN. The proposed method calculates a score for each candidate synset of a given Persian word and selects the synset with maximum score to be linked to the Persian word. This score is calculated considering related words of Persian word and words that appear in gloss of synset. A preliminary experiment shows that this method can be used to construct Persian WordNet. In the proposed method for each Persian word synsets with maximum calculated score are selected without considering other Persian words. In future work we intend to adapt our method and contribute other Persian word in order to select a synset for a given Persian word.

References


Automatic Persian WordNet Construction the 23rd International conference on computational linguistics pp. 846â€“850. fin Lindên K., Carlson. L., (2010). Thai Wordnet Construction Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP) Suntec, Singapore. Language codes linked to Lewis, M. Paul (ed.), 2009. Persian wordnet construction using supervised learning. arXiv preprint arXiv:1704.03223, 2017. [36] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation, pages 24â€“26. Cite-seer, 1986. [37] Jim Cowie, Joe Guthrie, and Louise Guthrie. Lexical disambiguation using simulated annealing. Semi Automatic Development of FarsNet; The Persian WordNet. Mehrmoush Shamsfard Computer Engineering Dept., Shahid Beheshti University, Tehran, Iran m-shams@sbu.ac.ir. Akbar Hesabi Linguistic Dept., Allameh Tabatabaiee University, Tehran, Iran a.hesabi11@yahoo.com. Abstract This paper describes the development process of FarsNet; a lexical ontology for the Persian language. FarsNet is designed to contain a Persian WordNet with about 10000 synsets in its first phase and grow to cover verbs' argument structures and their selectional restrictions in its second phase. In this paper we discuss the semi-automatic approach to create the first phase: the Persian WordNet. Introduction. WordNet (Miller 1995, Fellbaum 1998) is an. WordNet, showed most successful results during the experiments for automated construction of the Romanian WordNet. As reported by the authors, on a selected subset of synsets an error rate of only 2% has been achieved. After applying this heuristic for the construction of the Macedonian WordNet, we found out that 45% of the synsets produced by the GSD method contained exactly the same words. However, because the two methods rely on different rules for translating the synsets, each succeeds to translate different subsets of PWN. The last step of the construction of the MWN was to combine the sy... So, automatically constructing a WordNet resource from unlabeled text data is the best way for languages like Amharic which have limited resource. In this study, we propose Automatic Amharic WordNet construction using word embedding. The proposed model includes different tasks. The first task is text pre-processing which consists of commonly used text pre-processing tasks in many natural language processing applications.