Extracting Lexico-conceptual knowledge for Developing Persian WordNet

Mehrnoush Shamsfard, Hakimeh Fadaei, Elham Fekri
NLP Research Laboratory, Shahid Beheshti University
Tehran, Iran
E-mail: m-shams@sbu.ac.ir, ha.fadaee@mail.sbu.ac.ir, elham_fekri@yahoo.com

Abstract

Semantic lexicons and lexical ontologies are some major resources in natural language processing. Developing such resources are time consuming tasks for which some automatic methods are proposed.
This paper describes some methods used in semi-automatic development of FarsNet; a lexical ontology for the Persian language. FarsNet includes the Persian WordNet with more than 10000 synsets of nouns, verbs and adjectives.
In this paper we discuss extraction of lexico-conceptual relations such as synonymy, antonymy, hyperonymy, hyponymy, meronymy, holonymy and other lexical or conceptual relations between words and concepts (synsets) from Persian resources. Relations are extracted from different resources like web, corpora, Wikipedia, Wiktionary, dictionaries and WordNet. In the system presented in this paper a variety of approaches are applied in the task of relation extraction to extract ladled or unlabeled relations. They exploit the texts, structures, hyperlinks and statistics of web documents as well as the relations of English WordNet and entries of mono and bi-lingual dictionaries.

1. Introduction

WordNet is an electronic lexical database originally designed for English and replicated in several other languages. WordNet covers words from four POS categories: nouns, verbs, adjectives, and adverbs. It organizes words into sets of cognitively synonymous words, called synonym sets or synsets. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. Actually, each synset represents a distinct concept, which can be expressed by its members in a range of different contexts. Synsets are interrelated by means of lexical (word-to-word) and conceptual-semantic (synset-to-synset) relations. Nowadays WordNet is developed for more than 40 languages around the world.

This paper describes the relation extraction part in the semi automatic construction of FarsNet 1.0 (Shamsfard, 2008); the first WordNet for the Persian Language. Persian, also known as Farsi, is a member of the Iranian group of the Indo-Iranian sub-family of the Indo-European languages. It is the official language of Iran, Afghanistan and Tajikistan with more than 100 millions speakers.

Many works has been done in the field of relations learning during past years. They usually use either corpora or web documents as input text. Systems which extract relations from corpora like (Reinberger & Spyns, 2005; Ciaramita et al., 2008) usually use one or a combination of pattern based, linguistic and statistical approaches. In systems which use web documents specially Wikipedia articles as input text such as (Ruiz-Casado et al., 2008; Sanchez & Moreno, 2006), structure based methods are employed in addition to the above approaches.

2. Resources

Several resources are used in the relations extraction modules from which we mention Persian electronic resources in this section.

- **Corpora**
  - **Persian Linguistic Database (PLDB)**[1], (Assi, 1997) is an on-line database for the contemporary (Modern) Persian. The database contains more than 50 million words of all varieties of the Modern Persian language in the form of running texts. A small portion of texts are annotated with grammatical, pronunciation and lemmatization tags.
  - **Peykareh** (Bijankhan, 2004): is a collection gathered form Ettela'at and Hamshahri newspapers of the years 1999 and 2000, dissertations, books, magazines and weblogs. Written and spoken texts were collected randomly from 68 different subjects in order to cover varieties of lexical and grammatical structures. The version of Peykareh (also known as Bijankhan corpus) which we use contains about 10 millions manually tagged words with a tag set that contains 109 Persian POS tags.
  - Web: We have also used web in general and Wikipedia pages in specific to extract the relations.
- **Bilingual Dictionary**

Aryanpur (2005) English-Persian electronic Dictionary containing more than 200000 entries, is used in automatic expansion of FarsNet.

- **Lexicon**

Zaya lexicon (Eslami, et al., 2004) contains about 80000 Persian words and phrases with their POS tags and phonetic information.

3. Relation types

We divide relations into two categories: lexical and conceptual. Lexical relations refer to the relations between lexemes of a language. These relations act in lexical level more than conceptual level. Synonymy, antonymy and granularity (grading) are among these relations. These relations are usually language specific and initially, we do not expect them to be transferred

[1] http://www.pldb.ihcs.ac.ir

within different languages.
On the other hand, Conceptual relations are those who relate concepts which are usually shown by synsets in WordNets. Taxonomic and non-taxonomic relations such as hyperonymy/ hyponymy, meronymy/ holonymy, and cause/entailment are some examples of this category. We expect that the conceptual relations should be near-similar in different languages.
In this system both of the above two types of relations are extracted and learned from various resources including raw or tagged texts of available Persian corpora, Wikipedia articles and other documents on the web.

4. Relation extraction

In this section we describe different approaches we used to extract different types of relations.

4.1 Lexical approach

Morpho-lexical features of words are good sources of information for extracting relations to be inserted in the ontology. Among these relations we can mention antonymy which is a lexical relation. Antonymy relations could be extracted by applying morphological rules which are language dependant. For this purpose we consider the morphological rules which build antonym adjectives. These rules are formed by adding some negations affixes to positive stems.
In many languages there are some affixes for antonym building. For English language we can mention 'un', 'in' and 'im' as negation prefixes and 'less' as a negation suffix (e.g in 'countable' vs. 'uncountable' and 'complete' vs. 'incomplete'). We have such affixes in Persian (just in the form of prefixes) as well and 'bi' and 'na' are among them, like in the words *nadorost* (incorrect) and *bifayede* (useless) which are antonyms for the words *dorost* (correct) and *bafayede* (useful). We apply such rules on adjectives to extract antonymy relations.
It worth mentioning that all negation prefixes cannot be attached to all adjectives so we can't simply add prefixes to adjectives to make antonym adjectives. Therefore to find antonyms for a given adjective and to add the relation to ontology we first collect all negating prefixes and add them to the given adjectives. Then for each created word, if it can be found in the lexicon or its frequency of occurrence is more than a threshold in a corpus then it should be accepted as an acceptable word and be related to the candidate adjective by 'antonym' relation.

4.2 Pattern based approach

Pattern based approaches are exploited to extract both taxonomic and non-taxonomic, lexical or conceptual relations from Persian texts.
To extract taxonomic relations we define a set of 36 patterns containing the adaptation of Hearst patterns (Hearst, 1992) for Persian and some other new patterns. This approach also uses some patterns for a number of well known non-taxonomic relations such as "Part of", "Has part", "Member of" and "Synonymy". Some of the patterns used in this system are shown in table 1.

Patterns	**Relation**
TW is a X.	Hypernymy
TW is considered as X	Hypernymy
TW is known as X	Hypernymy
TW is called X	Hypernymy
TW is a part of X	Part of
TW includes X	Has part
TW means X	Definition
NP0 such as NP1, NP2, … (and | or) NPn	Hypernymy
Such NP0 as NP1, NP2, … (and | or) NPn	Hypernymy

Table 1: Translation of some patterns for extracting relations

These patterns are searched in the input resource to find their occurrences from which new conceptual relations could be extracted. Since the extracted patterns are not so frequent in corpora, we decided to use Wikipedia articles as input text. These articles are high informative and some occurrence of our patterns are usually found in the first section of them.
It should be mentioned that searching the patterns need some text processing tools (e.g. chunker) to find the constituents of sentences. While there is no efficient chunker for Persian, we did some post-processing to eliminate incorrectly extracted relations. This phase includes eliminating the stop words, applying some heuristics such as matching the head of the first noun phrase in the sentence with the head of the extracted TW in copular sentences, eliminating prepositional phrases for taxonomic relations, replacing long phrases with their heads and so on. Some of the relations extracted by pattern based approach are indicated in table 2.

Isa (pen, tool)
Isa (Japan, country)
Isa (onion, plant)
Isa (pain, symptom)
Has (Greece, history)
Synonym (thought, idea)
Has part (personality, specificity)

Table 2: Translations of some relations extracted by pattern based approach

Pattern based approach could also be used for finding the type of unlabeled relations. As it will be described in following sections, some approached may only extract related terms i.e. they present as their results a set of related pairs for which the type of relation is not identified. One of the ways which can be used to determine the type of unlabeled relations is to use pattern based approach. In this case the two related words are placed at the placeholders marked as TW and X in patterns of table 1. Then the newly generated patterns are searched over the input resource to see if we can find any occurrences of them. If we can find a reasonable number of occurrences of a pattern, its type is returned as the type of unlabeled

relation.

4.3 Structure based approach

Structures are another source for extracting relations which are used in many systems (hazman, et al., 2008; Agichtein, 2003). These structures include XML and HTML tags, tables, and hyperlinks. These structures could be found in any type of text. In Wikipedia pages Structures such as tables, bullets and hyperlinks are among these structures. In this system we use Wikipedia structures to extract conceptual relations.

For example in many Wikipedia documents we can find some information given via bullets. This information usually shows some taxonomic relations. In these cases the title of the section which only contains bulleted text is considered as the domain of the relation and each bullet forms the range of the relation.

Hyperlinks are other sources of information. In the whole document each important word which has an article in Wikipedia is linked to its related article. These linked words, mainly the ones locating in the first section of the text, are usually related to the title of the document especially if their correspondent articles have link to the original article. We use this fact to extract some taxonomic and non-taxonomic relations i.e. for a given Persian word we retrieve its Wikipedia article and find all hyperlinked words located in the first section of the article. These words are related to the given word. To make sure that the hyperlinked words are related enough to the given Persian words we examine their related Wikipedia article to see if they have link to the article of the given word or not. If such a link exists the word is returned as related word to the input word.

The third Wikipedia structure used in this system is 'disambiguation pages'. While searching a polysemous word in Wikipedia, if there are separate articles for each meaning of the word, Wikipedia brings a disambiguation page as the search result. In this page some or all of the meanings of the word are presented, usually with a brief explanation, in front of them. These explanations could be either a phrase or just a word indicating the parent of the word and they lead us to extraction of taxonomic of 'has domain' relations.

Some of the relations extracted by applying structure based method are shown in table 3.

Isa(car accident, event)
Isa (hypertension, disease)
Isa (Municipality, administrative division)
Isa (watch , device)
Isa (valve , device)
Related to (Instruction, School)
Related to (Calculus, Math)
Related to (Life, Death)
Related to (Child, Son)

Table 3: Some relations extracted by structure based method

4.4 Statistical approach

Statistical methods are widely used in extracting relations in many systems. In this system we use this approach for finding two types of relations: (1) general co-occurrence (2) class of noun phrases to be modified with an adjective. To extract the first type (general co-occurrence relations), for each pair of words within the 500 most frequent nouns of Persian we searched a 100,000 word subset of Bijankhan corpus to find in how many sentences these two words co-occur. If this number is above a certain threshold, these two words are considered as co-occurrents.

For the second type, we use the noun part of FarsNet. For each adjective, we extract all nouns for which the adjective has been appeared as a modifier in the corpus. Then a graph is built in which the leaves are these nouns and non leaf nodes are all noun synsets. Every noun is connected to all synsets in which one of its senses occurs. Finally with a search on graph we find the node which is connected to more leaves with least distance as the best noun synset (class) which can be modified by this adjective.

4.5 Ontology based approach

In this approach an existing English ontology like WordNet is used to extract taxonomic or non-taxonomic relations for Persian. This method consists of the following steps:

1- Mapping the given Persian word to a WordNet synset
2- Retrieving related synsets
3- Translating the related synsets to Persian
4- Forming new relations

In first step system finds the closest WordNet synset to the given Persian word. To perform this task first the English equivalents of the input Persian word is extracted from the bilingual dictionary. Then the system retrieves all the WordNet synsets in which any of the English equivalents appear. These synsets form our candidate synsets and the target synsets is selected from them. System picks the synset covering more English equivalents as target synset. After finding the target WordNet synset the system extract its related synsets in step 2. Regarding the type of the relations we are looking for, the related synsets are retrieved i.e. if we are looking for taxonomic relations, hypernym synsets are extracted.

In third step, this synset(s) is translated to Persian. In this step system uses a bilingual dictionary, Wiktionary and Wikipedia to translate the English words to Persian. The English words in a synset are looked up in these three resources and systems votes among them to choose the most suitable one. In the final step the new relations are created between the given Persian word and the Persian words resulted in step 3. Some of the hypernymy relations extracted by this approach are shown in table 4. Other types of relations could be extracted as well.

Isa(Newspaper, Press)
Isa (Book, Publication)
Isa (Country, Political unit)
Isa (Instruction, Activity)
Isa (Sir, Title)
Isa (Face, External body part)

Table 3: Translations of some relations extracted by ontology based method

It is worth mentioning that this method is suitable for less resourced languages because it uses the resources of other languages to extract conceptual relations.

4.6 Clustering

Automatic adjective clustering is another method we used for relation extraction. The goal is to put adjectives that are defining different degrees of the same attribute in one cluster. For example words {hot, warm, cool, cold, chilly} describe temperature attribute with different intensity, and so they must be put into the same class. To cluster adjectives we compute dissimilarity between them. Our system employs known linguistic and statistical methods for adjective clustering. In linguistic side we use a pattern based approach and search for co-occurring adjectives in noun phrases. If two adjectives are co-occuring in an Ezafe-construction, they may not be in a cluster while if they occur in a positive or negative conjunction they probably belong to a cluster. For example, adjectives "سرد" [sard, cold] and "گرم" [garm, hot] which belong to one cluster, usually cannot be used in one Ezafe-construction ("آب سرد گرم" [äb - e sard -e garm: cold hot water]) because one thing cannot be hot and cold at the same time. While they can occur in a conjuction such as ("نه سرد و نه گرم" [na sard va na garm: neither cold nor hot]).

On statistical side we assume that similar adjectives appear with common set of nouns. Suppose that frequency of occurrence of adjective i with noun j is Fij. For each two adjective, A and B and nouns X and Y If Fax<Fay and Fbx<Fby, or, Fax>Fay and Fbx>Fby the two adjectives are concordant and otherwise they are discordant.

Similarity is define as: Similarity= Pc - Pd , where Pc is the probability of being concordant, and Pd is the probability of discordance, so it's range is between -1(dissimilar) and 1(similar).

Then we cluster adjectives according to their dissimilarity value by minimizing the following objective function by hill climbing approach.

$$\varphi(p) = \sum_{i=1}^{R} [1/|Ci| \sum_{\substack{x,y \in Ci \\ x \neq y}} d(x,y)]$$

In which R shows the number of classes, Ci shows the ith class, |Ci| is the total number of elements in ith class. d(x, y) is dissimilarity parameter calculated for adjectives x and y.

5. Results and Conclusion

In this article we described some relation extraction methods which were used to build Persian WordNet semi automatically.

The related pairs extracted in the statistical and structure based section are considered as candidate relations and are verified with pattern based section. The precision of pattern based section is 76%. Test results in structure based approach shows a precision of 55% in extracting relations from bullet structures and 74% in relation extraction via disambiguation pages. The best results of our clustering approach shows 54.5% precision, 74% recall and 60.5% F-measure to find the granularity relations.

6. References

Agichtein, E., Ho, H., Josifovski, V. and Gerhardt, J. (2003). *Extracting Relations from XML Documents*, Lecture Notes in Computer Science, Springer, pp390-401.

Aryanpur, M., Assi, M. (2005). *The Aryanpur Progressive Persian-English Dictionary*.

Assi, S. M. (1997). Farsi Linguistic Database (FLDB). *International journal of Lexicography*, V10, Euralex Newsletter.

Bijankhan, M. **(**2004). Role of language corpora in writing grammar: introducing a computer software. *Iranian Journal of Linguistics*, No. 38: pp. 38-67.

Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J. and Rojas, I. (2008). Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. *Ontology learning and Population: Bridging the Gap Between Text and Knowledge*. IOS press. chapter of book

Eslami, M., Sharifi, M., Alizadeh, S., Zandi, T., (2004). 'Persian Generative Lexicon', *1st workshop on Persian Language and Computer*, pp 6-11.

Hazman, M., El-Beltagy, S.R. & Rafea, A. (2008). Ontology learning from textual web documents, *Proceedings of the 6th International Conference on Informatics and Systems (INFOS'2008)*, NLP, (pp.113-120), Giza, Egypt.

Hearst, M. A. **(**1992). Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Reinberger, M. and Spyns, P. (2005). Unsupervised Text Mining for the Learning of DOGMA-Inspired Ontologies. *Ontology learning from text: Methods, Evaluation and Applications*. IOS press.

Ruiz-Casado, M., Alfonseca, E., Okumura M. and Castells, P. (2008). Information Extraction and Semantic Annotation of Wikipedia. *Ontology learning and Population: Bridging the Gap Between Text and Knowledge*. IOS press.

Sanchez, D. and Moreno, A. (2006). Discovering non-taxonomic relations from the We. *Lecture Notes in Computer Science*, Springer.

Shamsfard M. (2008). Developing FarsNet: A Lexical Ontology for Persian, *In proceedings of the 4th global WordNet conference.*