Corpus based Semi-Automatic Extraction of Persian Compound Verbs and Their Relations

Somayeh Bagherbeygi[1], Mehrnoush Shamsfard[2]
[1]Linguistic Dept., Allameh Tabatabaiee University, Tehran, Iran.
[2]Electrical and Computer Engineering Dept., Shahid Beheshti University, Tehran, Iran.

Abstract

Nowadays, Wordnet is used in natural language processing as one of the major linguistic resources. Having such a resource for Persian language helps researchers in computational linguistics and natural language processing fields to develop more accurate systems with higher performances. In this research, we propose a model for semi-automatic construction of Persian wordnet of verbs.
Compound verbs are a very productive structure in Persian and number of compound verbs is much greater than simple verbs in this language This research is aimed at finding the structure of Persian compound verbs and the relations between verb components. The main idea behind developing this system is using the wordnet of other POS categories (here means noun and adjective) to extract Persian compound verbs, their synsets and their relations. This paper focuses on three main tasks: 1.extracting compound verbs 2.extracting verbal synsets and 3.extracting the relations among verbal synsets such as hypernymy, antonymy and cause.

Keywords: WordNet; Automatic WordNet development; Natural language processing (NLP); Compound verb; Persian

1. Introduction

The Persian language, also known as Farsi, is the official language of Iran, Afghanistan and Tajikistan with more than one hundred million speakers and also spoken in more than six other countries. There is no doubt in the necessity of constructing basic language processing resources and tools (such as semantic lexicons like WordNet) for it, like many other less-studied languages. WordNet (Miller 1995, Fellbaum 1998) is a semantic lexicon which covers words from four POS (part-of-speech) categories: nouns, verbs, adjectives, and adverbs. The database is organized around the notion of synset (synonym set) between which semantic relations are expressed.

In this paper, we discuss the semi automatic development of the Persian WordNet of compound verbs within the FarsNet project by exploiting the WordNet of other categories (noun and adjective), Persian corpus and dictionaries.

FarsNet_1.0 (Shamsfard et al. 2010) which is the first Persian WordNet contains about 18,000 words and phrases organized in about 10,000 synsets of nouns, adjectives and verbs. In this paper we discuss the semi-automatic development of compound verbs section in FarsNet based on noun and adjective parts.

To conduct the present study, a massive corpus of various written data comprising about 400 million words is collected; besides, Bijankhan corpus as well as Aryanpour Electronic Dictionary are used as electronic resources. The other required resources are nominal and adjectival network of FarsNet. In order to evaluate the results, Sokhan Dictionary of 8 volumes (Anvari, 2003), and contemporary Persian dictionary (Sadriafshar, 2003) are used.

2. Previous studies on Persian compound verbs

Compound verb structure has always been one of the most controversial linguistic issues and several researches have been done on it from different aspects of syntactic, semantic, cognitive and computation, under different topics such as “Light Verb Construction”, “Compound or Complex Verb”, and “Complex Predicate”. There are different reports on the number of Persian simple verbs; Khanlary (2004), for instance, suggests 279 verbs, Sadeghi (1994) discusses 115 verbs and Family (2006) talks about 160 verbs and this shows that Persian native speakers intend to use the productive combination of compound verbs to express verbal concepts and the usage of simple verbs is decreasing intensively. Khanlary (2004) states that compound verbs will gradually replace simple verbs in modern Persian and this process has started from the 13th century. Many linguists try to achieve the child’s mental modeling of compound verb structure from different approaches to this structure in Persian. The specific feature of Persian compound verbs which is rare in other languages has motivated the Iranian and Non-Iranian linguists to provide different analysis of this structure from different approaches.

The common point of all definitions provided in recent studies for the compound verbs can be stated as follows: Persian compound verbs are the composition of a nonverbal element and a verbal element; the nonverbal elements include noun, adjective, adverb, prepositional phrase and particle. Verbal elements comprise of some Persian simple verbs which are called light verbs according to the Jespersen’s definition. Light verb is a simple verb which has undergone semantic bleaching (Vahedi langroodi, 1996; Karimi-Doostan, 1997; 2005). Any Light verb has a heavy counterpart and this seems to be the characteristics of all languages. In the present research, in order to provide a comprehensive definition and classification, Persian verbs are divided into two classes of simple verbs (including a single verbal component) and compound verbs (including nonverbal and verbal components). We refer to compound verb as a verb which is composed of one or more nonverbal components and a simple or compound verbal component.

3. The Proposed Method

This paper describes a semi-automatic method for extraction of Persian compound verbs and their relations from corpus and other resources. The motivation of our proposed approach lies in the construction of WordNet of one category based on its relationship to the WordNet of other categories of the same language. Since the relational pattern over the Persian compound verbs and the nominal and adjectival categories is observable from the structure of most Persian compound verbs, and since this is realized in the form of generative patterns, the design of this pattern and its semi-automatic application on the nominal and adjectival WordNet has been considered as a method for construction of the WordNet of Persian verbs in this research. Among the advantages of this method we can refer to the absence of bias to the WordNet of other languages and considering and retaining the specific features of the language in question.

The proposed method contains 3 phases:

1. Extracting compound verbs,
2. Extracting synonymy relations among compound verbs (constructing synsets), and
3. Extracting hyperonymy, hyponymy, antonymy and cause relations among synsets.

3.1 Extracting Compound Verbs

Persian verbs can be classified into two groups: Simple verbs (containing a single verbal component) and Compound verbs (containing nonverbal and verbal components). In Persian compound verbs, the verbal component can be either a simple (light) verb or a compound verb itself and the non-verbal component may be a noun, an adjective, an adverb or a prepositional phrase. In this study we consider the noun and adjective non-verbal elements and thus we use the noun and adjective categories of FarsNet to create the verb category of it. At first, all nouns and adjectives in FarsNet 1.0 are stored along with their POS tag in a table entitled "*nonverbal components"*. Then, a complete list of simple (light) verbs commonly used in the present Persian written and spoken language along with a total number of 130 compound verbs such as "/negah dAštan/ (to hold) which contribute as a verbal component in the construction of other compounds, are stored in a separate table entitled "*verbal components"*.

Then a Cartesian product of these two tables is produced automatically in a table called *"candidate compounds"* in which each nonverbal component[3] is attached to each verbal component and the result should be tested for being a valid compound verb. There are two tests to be performed to recognize compound verbs as a lexical unit in the corpus, Compound verbs as the verb of the sentence and Infinitivisation test. To test each candidate compound (to see if it is a valid compound verb) we search for its infinitive and also all its inflections in the corpus and count the frequency of occurrences. It is worth mentioning that at this stage the form of each occurrence is considered regardless of its meaning, so all occurrences of a verb which even may have different meanings due to the polysemy of the nonverbal component, are considered identical. The verbs which are used in the corpus are selected as compound verbs and then are evaluated manually by referring to Sokhan Dictionary and in case the verb is not found in this Dictionary, with regard to its frequency in the corpus, or by referring to contemporary Persian dictionary, or ultimately based on the author's intuition. Experimental results show that 81.79% of selected compound verbs have an entry in Sokhan dictionary, 83.38% of them are confirmed by Aryanpour dictionary and 93.01% are confirmed by human intuition based on their high frequency in the corpus. Table 1 shows the number of extracted compound verbs according to the POS category of their nonverbal component.

Noun	Adjectie	Adverb	Verb Particle	Prepositional phrase
4568	2215	77	100	165

Table1- Number of compound verbs according to the category of their nonverbal components

Since some verb prefixes, such as "/mi-/ (verbal prefix) have homograph pairs (ex. /mey /) which is a noun in Persian, the compound verbs extracted from them have obtained a high score in the system while they are wrong candidates.

As it can be seen in table1, although we exploited just the noun and adjective non-verbal elements to make the compounds, there are some compounds made from adverbs, verb particles and prepositional phrases too. These verbs are created due to polysemy of some words. For example although the words "/pas/ (behind) and /piš/ (front) are listed as nouns in FarsNet they have adverbial role (according to Dabirmoghaddam (1995)), in compound verbs such as "/piš raftan/ (front+ to go=to advance)" and "/aghab mAndan/ (behind +to stay= to stay behind).

The same thing is true for verb particles such as "/bar/ (on), and /dar/ (in) which are listed as nouns too.

3.2 Extracting Verbal Synets

The second phase constructs verbal synsets. In this phase, we consider the polysemy of compound verbs. Table 2 shows the number of distinct compound verbs extracted from cited resources considering polysemy.

Intuition based on high frequency in the corpus	Aryanpour Dictionary	Sokhan Dictionary
7139	6386	6302

Table2 Number of extracted distinct compound verbs, considering polysemy

The second phase will be accomplished in two stages:

1- Semantic tagging the extracted compound verbs according to their meanings on the semantic continuum,
2- Semi-automatic construction of synsets.

For the first stage, we classify compound verbs into two categories from semantic point of view: composites and idiomatic (Karimi, 1997). We focus on composite verbs where the meaning of the verb is somehow predictable from the meaning of its components; transparent, semi-transparent or semi-ambiguous. There is also another category 'old' which includes the old compound verbs which can be found in dictionaries but are not frequent

[3] The number of FarsNet nouns and adjectives that are used as nonverbal components is 10973 and 3810 respectively

now. Table 3 shows the statistics of semantic tagging of extracted compound verbs.

Old	Idiomatic with non-metaphorical meaning	Composite with metaphorical meaning	idiomatic	composite
97	8	75	693	6432

Table3- Results from the first stage of the second phase based on semantic criteria

At the second stage, the semi-automatic construction of verbal synsets, or in other words, determining the synonym compound verbs and inserting them into a single set, is performed by application of the following three rules:

Rule 1: Insert compound verbs with synonym nonverbal components and identical verbal components into a single set.

Rule 2: Insert compound verbs with identical nonverbal components and synonym verbal components into a single set.

Rule 3: Insert compound verbs with synonym nonverbal components and synonym verbal components into a single set.

For some verbs the application of the first two rules causes the application of the third one implicitly. Following the first principle, we will automatically insert compound verbs with synonym nonverbal components and identical verbal components into a single synset, for example adding the nominal synset {/zarar/, /xesArat/, /ziAn/ (loss)} to the identical verbal component 'zadan' makes the verbal synset {/zarar zadan/, /xesArat zadan/ (to damage)}

According to our approach, verbal components of the compound verbs are polysemic with some degrees of overlap. Any sense of a verbal component may be equivalent to the sense of one or some other verbal components. For example "/zadan/ (to strike)"as a light verb may be the synonym of light verbs like "/resAndan/ (to carry), /vAred kardan/ (to enter), or sometimes "/kardan/ (to do)", depending on its various meanings. Accordingly 75 verbal synsets containing the verbal components of compound verbs, were distinguished. This may be used in the second and third rules. For example the verbal component synset /kardan,GozArdan,goftan/ (do/say) can be added to the nominal synset {/Sokr/ sepAs/ (thank)} and create the verbal synset {/Sokr kardan/, /sepAs gozArdan/,/sepAs goftan/ (thank giving)} following the third rule.

After applying the above rules, we manually test their accuracy with regard to every synset, by using Sokhan Dictionary. Table 4 shows the results.

Accuracy	Total number of verbal synsets
94.3	4218

Table4 Results from the second phase

As the accuracy shows the rule has some exceptions. For example the verbal component synset {/kardan/, /?Avardan/} can be added to the nominal synset {/bAr/} and create wrong candidates {/bAr ?Avardan/(raise),/bAr kardan/(load)} while the same verbal synset can be added to the word bahAne andcreate true verbal synset {/bahAne ?Avardan, bahAne kardan (excuse)}.The construction process of verbal synsets in the second phase shows that the synonymy relation among Persian compound verbs is achieved from both verbal and nonverbal components of the compound verb.

3.3 Extracting Conceptual Relations among Compound Verb Synsets

In the third phase we look for relations such as troponymy, antonymy and cause. Results show that for some verbal concepts, the opposition relation is merely extracted from that of their nonverbal components, i.e. nominal and adjectival concepts; for others this relation is extracted from opposition relation between their verbal components and for some concepts, this relation can be extracted from that of nonverbal as well as verbal components. The system has extracted 59 antonymy relations from verbal components, showing an accuracy of 55.67%. {/?az dast raftan/ (to be lost)} versus {/be dast ?Amadan/ (to be obtained)} is an example for these relations.

To automatically extract the cause relations among synsets, a list of causative/non-causative alternations of the verbal components is first prepared. Then this relation is automatically established between them, with attention to the verbal components of the compound verbs in the two synsets. For example as the following pairs of verbal components as a light verb have cause relations when combined with a nonverbal component like a noun or an adjective: /zadan/ (to strike) - /xordan/ (to receive), /zadan/ (to strike) - /didan/ (to see), /resAndan/ (to carry) – /didan/ (to see), /resAndan/ (to carry) - /xordan/, their extracted synsets grouped in A and B have cause relations as well:

A) {/sadame zadan/ , /sadame resAndan/ , /sadame vAred kardan/, /latme zadan/, /latme resAndan/, /latme vAred kardan/, /?Asib resAndan/, /?Asib zadan/, /gazand resAndan/ (to hurt)}

B) {/sadame xordan/, /sadame didan/, /latme xordan/, /latme didan/, /?Asib didan/ (to be hurt)}

A number of 911 cause relationships has been extracted which was evaluated manually and showed an accuracy of 89.24%.

Finally, findings show that troponymy relation which forms the hierarchical structure of verbal concepts can be extracted from both nonverbal and verbal components. To construct the hierarchical relation of verbal concepts, we first tried to consider a semantic classification for Persian compound verbs somehow compatible with the semantic classification of verbs in the Princeton WordNet so that the Persian WordNet of verbs can easily be mapped to Princeton WordNet of verbs. From semantic point of view, one of the verbs classification criteria is based on their aspectual features which have been considered both in WordNet verbs classification and Karimi-doostan's (1997) compound verb classification. We expanded this classification and introduced a method based on it for extracting troponymy relation according to nonverbal components semantic fields in WordNet. The model for determining the troponymy in verbal network is as follow:

Suppose we have the two nominal concepts M and N so as the verbal concepts V1 and V2 are related to the concept N and the concepts V3 and V4 are related to M and a

hyponymy relation is held between N and M (whether N is the superordinate or M), in order to determine the hyponymy relation between each of the concepts V1 and V2 with V3 and V4, we should consider the following:

The relation is established when the verbal components of the verb are identical in the first appearance or otherwise the two verbal concepts share the same aspectual features; i.e. their verbal components are in a single set.

Example :
N: {/?ahamiat/ (importance)
V1: {/?ahamiat dAdan (to emphasize)
V2: {/?ahamiat dAštan/ (to be important)

M: {/?arzeš (value)}
V3: {/?arzeš dAdan/ , /?arzeš baxšidan/ (to value)}
V4: {/?arzeš dAštan/ (to be worth)

Respectively:
N hyponym →M ⇨ V1 hyponym→V3
V2 hyponym→V4

This method had an accuracy of 71.59% in extracting troponymy relation.

4. Discussion

The project successfully completed the construction of Persian WordNet of verbs. As results show, by employing morphological and semantic patterns specific to Persian compound verbs and establishing the relational pattern over Persian compound verbs and their components, the semi-automatic construction of Persian WordNet of verbs is applicapable. The method employed in this research is considered to be a modern one as compared to others. As an efficient method, it resulted in considerable theoretical findings on Persian compound verbs. Results from applying three rules in phase 2 show that although the primary meaning of a composite compound verb is obtained from the meaning of its nonverbal components, the meaning is partly understandable from the verbal component. Findings show that the nonverbal components of the composite compound verbs are polysemic with some degrees of overlap. Therefore, theoretical findings of this phase will lead us to a semantic classification of nonverbal components of the composite compound verbs which proves that any sense of a verbal component may be equivalent to the sense of one or some other verbal components. In other words, the meaning of a compound verb is obtained from the collective meaning of its components; so by entering verbal components in separate synsets and applying the three rules, we can extract verbal synsets. As an interesting result from this phase we can refer to the two meanings of the verbal component "/kardan/ (to do)" derived from its nominal and adjectival components. Results show that the nonverbal components of all verbs whose verbal components are in the synset "/kardan/ (to do) - /sAxtan/ (to make)" are adjective. Examples of these verbs are shown in the following synsets.

{/?AšekAr kardan/, /?AšekAr sAxtan/, /namAyAn kardan/, /namAyAn sAxtan, /?alani kardan/, /ma?lum kardan/, /mo?ayan kardan/ (to reveal)}

Semantic examination of these verbs shows that these verbal components transfer the meaning of causality to their compound verbs.

Results from the third phase show that for some verbal concepts, the opposition relation is merely extracted from that of their nonverbal components, i.e. nominal and adjectival concepts; for others this relation is extracted from opposition relation between their verbal components and for some concepts, this relation can be extracted from that of nonverbal as well as verbal components. Finally, findings show that troponymy relation which forms the hierarchical structure of verbal concepts can be extracted from both nonverbal and verbal components.

5. Further works

The following items are considered as the future works:

1. Developing a method for automatic extraction of compound verbs with adverbial and prepositional phrases as nonverbal components is among our future works.
2. Mapping the Persian WordNet of verbs with the Princeton WordNet with attention to the classification of compound verbs provided in this research can be the subject of a future project in this field.
3. Data from the present study can also be used in performing linguistic theoretical studies in future. As an example, the study of verbal and nonverbal components of the compound verbs extracted from the corpus and their frequency can be the basis for future studies on the Persian compound verbs. A more detailed study on composite compound verbs on semantic continuum, by focusing on the meaning of the verbal component can also be a subject for future studies.

References

Anvari, H. (2003). Sokhan Great Dictionary. Tehran. Sokhan Publishing House.

Dabirmoghaddam, M. (1995). Compound Verbs in Persian Language. Linguistic Periodical. 12th year. 1st and 2nd issue. P. 46-2.

Family,N. (2006). Explorations of semantic space: the case of light verb construction in Persian.Ph.D. Dissertation. Ecole des haudes en Sciences Sociales,Paris.France

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press

Karimi, S. (1997). Persian complex verbs: idiomatic or compositional. Lexicology 3(2): 273-318.

Karimi-Doostan, Gh. (1997). Light Verb Constructions in Persian. Ph.D. Dissertation. University of Essex.

Khanlary, P. (2004). History of Persian Language. 3 volumes. Neu Publishing House.

Miller, G. (1995). WordNet: a lexical database for English. Communications of the ACM archive. Volume 38, Issue 11. Pages: 39 – 41.

Sadri Afshar, Q. (2003). Contemporary Persian Dictionary. Tehran. Farhang-e- Moaser Publishing House.

Shamsfard M., Hesabi A., Fadaei H., Mansoory N., Famian A., Bagherbeigi S., Fekri E., Monshizadeh M., Assi M. (2010) Semi Automatic Development of

FarsNet; The Persian WordNet, Global WordNet Conference, Mumbai, India.
Sadeghi, A. (1994). On Pseudo-Verbs in Persian Language. Article Series of Persian Language and Scientific Language Seminar. Tehran. University Publication Center. P. 236-246.
Vahedi-Langrudi, M. (1996). The syntax, Semantics and Argument Structure of Complex Predicates in Modern Farsi. Ph.D. Dissertation.University of Ottawa.