Extraction of Verbal Synsets and Relations for FarsNet

Fatemeh Khalghani
Faculty of Computer Science and Engineering
Shahid Beheshti University
Tehran, Iran.
f.khalghani@gmail.com

Mehrnoush Shamsfard
Faculty of Computer Science and Engineering
Shahid Beheshti University
Tehran, Iran.
m-shams@sbu.ac.ir

Abstract

WordNet or ontology development for resource-poor languages like Persian, requires composition of several strategies and employment of appropriate heuristics. Lexical and linguistic structured resources are limited for Persian and there is a lot of diversity and structural and syntagmatic complexities. This paper proposes a system for extraction of verbal synsets and relations to extend FarsNet (Persian WordNet). The proposed method extracts verbal words and concepts using noun and adjective words and synsets. It exploits the data from digital lexicon glossaries, which leads to the identification of 6890 proper verbal words and 2790 verbal synsets, with 91% and 67% precision respectively. The proposed system also extracts relations such as semantic roles of verbal arguments (instrument, location, agent, and patient) and also "related-to" (unlabeled) relations and co-occurrence among verbs and other concepts. For this purpose, a combination of linguistic approaches such as morphological analysis of words, semantic analysis, and use of key phrases and syntactic and semantic patterns, corpus-based approach, statistical techniques and co-occurrence analysis have been utilized. The presented strategy extracts 5600 proper relations between the existing concepts in FarsNet 2.0 with 76% precision.

1 Introduction

Semantic or conceptual relation extraction between concepts and appropriate relation labeling forms an important part of ontology learning and ontology construction process that is widely used in information retrieval, question-answering systems, summarization, and word sense disambiguation (WSD) (Girju, 2008).

Learning and labeling of conceptual relations has been introduced as the most complex and challenging element in most of systems, especially in the construction of ontologies or WordNets (Sánchez & Moreno, 2008; Kavalec & Svátek, 2005). This problem can be divided into two separate parts of relationship extraction and labeling. The latter which tries to label an existing unlabeled relation between two concepts has been less addressed in previous studies.

Semantic analysis requires composition of various approaches like pattern-based and corpus-based techniques for languages such as Persian that lack accurate analytical instruments and structured sources and suitable tagged corpora. Thus, several lexical resources including syntactic verbal valency lexicon (Rasooli et al., 2011), comprehensive lexicon of synonyms and antonyms (Khodaparasti, 1997), online and digital lexicons such as Vajehyab browser[1], Wikipedia, Dadegan dependency Treebank (Rasooli et al., 2013) and Wortschatz statistical corpus (Goldhahn et al., 2012) have been used in the presented strategy. In addition, Persian preprocessing tools such as Negar text editing tool, STeP-1 morphological analyzer (Shamsfard et al., 2010) and ParsiPardaz dependency parser (Sarabi et al., 2013), and a composition of linguistic, syntactic, and statistical approaches have been used for semantic relation extraction.

As verbs are the main core of sentences in many languages, extending the verbal part of wordnets may improve their efficiency and application in semantic analysis of texts.

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

This paper focuses on extraction of verbs, verbal synsets and non-taxonomic relations in which at least one of the related terms is verb.

The given strategy in this paper for extraction of verbal concepts emphasizes on wide range of compound verbs and prefixes in Persian with highly metaphorical concepts, and in addition to implementation of verbal construction rules and paying attention to Arabic rhythms of words, starting from concepts of noun and adjective, it derives correspondent verbal concepts from several online lexicons with analyzing of entries and text of explanations and examples in thesauruses.

The rest of the paper is organized as following: section 2 present a brief introduction to FarsNet and its current situation, section 3 discusses related work, section 4 describes the proposed method including verb extraction, verbal synset composition and verbal relation extraction. Section 5 concludes the paper and suggests some further work.

2 FarsNet

FarsNet, the first Persian WordNet (Shamsfard et al., 2010) is a lexical database for Persian words. The first and second versions of FarsNet have been established in Natural Language Processing lab of Shahid Beheshti University at 2008 and 2010 respectively. FarsNet 3.0 which is currently under development is expected to have 100,000 lexical entries (currently about 87,000 are available). FarsNet like other wordnets is formed by a large set of lexical entries (words or phrases) organized in a network of synsets (a set of synonym terms). The edges of this network are semantic relations among synsets, including inner-POS and inter-POS ones. The relations defined between synsets in FarsNet include hypernym/ hyponym, holonym/ meronym, antonym, domain, related-to, co-occurrence, cause, entails, salient-defining feature, potential-defining feature, unit and attribute; besides, some semantic roles as instrument, location, agent and patient. The report of variation trend of FarsNet versions and the existing semantic relations has been presented in (Shamsfard & Ghazanfari, 2016). Also, Table 1 displays statistics of words, synsets, and the relations between senses and synsets in various versions.

FarsNet version	Words	Word senses	Synsets	Synset relations	Sense relations
1.0	17842	24480	10012	6980	360
2.0	30222	36115	19398	36848	7043
2.5	33290	39735	20559	47761	19021
current	86747	98370	37959	91744	28739

Table 1: Statistics of words, synsets and relations of FarsNet

The strategies given in this paper have been adapted to extend and improve FarsNet 3.0 verbal synsets and relations.

3 Related Work

In the related field of automatic wordnet development, several efforts have been made. According to classification of Vossen (1998) for wordnet development approaches, two major approaches can be considered as merge and expansion. The merge approach is constructing a wordnet with independent use of target language resources and language specific properties and usually creating synsets from scratch, whereas the expansion method relies on existing wordnets (especially the English WordNet) and uses multilingual resources to translate words of existing synsets to target language and therefore preserves the source wordnet structure. However developing a wordnet by use of merge method is not always cost effective due to budget constraints and is more time-consuming than expansion method, it leads to a higher quality and extensive wordnet to be effectively used in certain and real NLP applications. Also a wordnet developed with merge approach will reserve the target language culture and region specific concepts and semantic relations and there is no need to deal with translation ambiguity, compared to expansion approach (Prabhu et al. , 2012).

Prabhu et al. (2012) use a hybrid approach of merge and expansion for developing IndoWordNet to benefit from the advantages of both.

Recently, word embedding models, especially Word2Vec (Mikolov et al., 2013), have been the focus of much research in NLP tasks. These models are widely used to calculate semantic relatedness of words and thus they can be applied in synset construction and semantic relation extraction subtasks of a wordnet development process.

The proposed work by Al Tarouti (2016) on Arabic wordnet and Mousavi and Faili (2017) on Persian wordnet use vectors created by Word2Vec to move towards a wordnet by an expansion method.

There are some other efforts to build a wordnet for the Persian language by either semiautomatic or automatic methods. Among semiautomatic ones we can mention (Bagherbeygi & Shamsfard, 2012), (Fadaei & Shamsfard, 2010)

and (Shamsfard, et al., 2010) which mainly use a merge method to build a Persian wordnet.

Among automatic methods we can mention (Dehkharghani & Shamsfard, 2011), and (Taghizadeh & Faili, 2016) which mainly use expansion methods and extract some mappings between Persian words and Princeton synsets. These systems do not build a wordnet but can be used to initiate building a wordnet. They are good in coverage and development time but not as well in precision of result.

Most of the research conducted on extraction and labeling of conceptual relation for Persian language (such as (Shamsfard & Barforoush, 2004) and (Fadaei & Shamsfard, 2010)) work on a limited predefined relations such as synonymy, hyper/hyponymy and holo/meronymy relations; and they lack favorable and needed efficiency for non-taxonomic relations and those relations corresponding to semantic roles (role relations).

Boudabous (2013) proposes a linguistic method based on morpho-lexical patterns to extract semantic relations in order to improve the Arabic WordNet (AWN) performance using Arabic Wikipedia articles as the input corpus.

The methods proposed by Shamsfard & Mousavi (2008) and Jafarinejad & Shamsfard (2012) carry out labeling thematic roles in sentence through rule-based approaches using shallow parsing of text; the mentioned conducted works lack favorable efficiency for extraction of conceptual relations among FarsNet high-level concepts and they don't have appropriate recall either. The work done by Zadeh Khosravi Forooshani & Rezaei Sharifabadi (2016) carries out semantic role labeling in Persian sentences by dependency parsing; that in comparison to works implemented with shallow-parsing, has higher accuracy and better efficiency; but it does not yet propose any solution for extraction of corresponding semantic relations among high level concepts of a wordnet.

The strategy suggested by Bagherbeygi & Shamsfard (2012) is one of the works conducted for automatic extraction of Persian verbal concepts in which FarsNet noun and adjective concepts are used for compound verbs extraction. It considers any combination of each noun/adjective and Persian light verbs as a compound verb candidate and then verifies correct words by checking up in Bijankhan Corpus and Arianpour Dictionary. Then it makes verbal synsets by a rule based mothod from noun and adjective synsets. Despite appropriate efficiency of this technique in derivation of phrasal verbs, with respect to reliance of this method on combination of noun and adjective with light verbs and limitation of the used lexical sources, many prefixed and propositional verbs as well as more complex expressions and verbal phrases with metaphorical concepts are not identified.

4 The Proposed Method

4.1 Verbal Synset Extraction

The proposed method for verbal synset extraction uses the existing noun and adjective synsets and it is focused on the principle that automatic learning of concepts by starting from synsets instead of words, reduces processing and time costs for building synsets and extension of database.

The basic concept of the proposed approach is to consider the internal structure of the phrasal and prefixed verbs and verbal and phrasal terms with metaphorical concepts; this has led to expansion of verbs in Persian. The non-verbal parts of compounds are derived from the noun and adjective concepts and the appropriate verbal parts and prepositions and prefixes should be extracted.

To this end, we first consider each noun synset and for each noun in it, apply some rules to find a corresponding verbal concept based on its semantic category, grammatical structure and Arabic rhythm (for words with Arabic origin). Considering all semantic classes of nouns, we extracted the semantic classes for which verb extraction is possible. These classes are act, attribute, possession, motive, feeling, event, cognition, state, relation, and process.

Afterwards, based on structural rules of Persian and Arabic gerunds, the verbal and non-verbal parts are derived for the words for which these rules are applicable. Some of these rules are as following:

- Words with *Fe'Alat* rhythm such as *TebAbat* (طبابت: medicine), *VekAlat* (وکالت: attorney-ship), and *KetAbat* (کتابت: writing) can be combined with the light verb *Kardan* (کردن: to do) to make a compound;
- Words with *Fa'Al* rhythm e.g. *KaffAsh* (کفاش: shoemaker), *AkkAs* (عکاس: photographer), and *NaqqAl* (نقال: narrator) can be used to make a phrasal verb by the rule *word+ ye+ kardan* (~ ی + کردن).
- Words with suffix *Gari* (~ گری) can be participate in verb construction with/without deletion of suffix before adding to *Kardan*, e.g. *Efsha Kardan* (افشا کردن: to disclose) from

EfshA+Gari (افشاگری: disclosure); and *SodA Kardan* (سودا کردن: to speculate) from *SodA+Gari* (سوداگری: speculation).

- From words with suffix *Gi* (گی ~) proper verbs can be made by a set of rules. For example *Kooftan* or *Koofteh Shodan* (کوفته شدن یا کوفتن: to concuss) from *Kooftegi* (کوفتگی: concussion) and *RAnandegi Kardan* (رانندگی کردن: to drive) from *RAnandegi* (رانندگی: driving).
- From the combined words whose structure ends to (present lemma + ی (i)) the corresponding verbs can be obtained by substituting the (present lemma + i) with its corresponding gerund form. *TAj GozAshtan* (تاج گذاشتن: to crown) from *TAjgozAri* (تاج گذاری: crowning), and *Tasmim Gereftan* (تصمیم گرفتن: to decide) from *Tasmimgiri* (تصمیم گیری: decision making).

Some of noun words which are very numerous do not follow any certain rule; but they have participation in structure of phrasal verbs as verbal part(s). For example, making verb of *Habs Keshidan* (حبس کشیدن: to imprison) from noun of *Habs* (حبس: prison) and *Shak DAshtan* (شک داشتن: having suspicion) from *Shak* (شک: suspicion) and *Be Haghighat Peyvastan* (به حقیقت پیوستن: to come true) from word of *Haghighat* (حقیقت: truth); extracting a rule for these cases is not easy and they can be validated through analysis on lexicons and the related corpora. For this purpose, the suitable verbal part can be obtained for each noun non-verbal part of a compound verb automatically by searching for the word(s) in entries and body of the group of digital lexicons including Khodaparasti Glossary, Moein Thesaurus, Dehkhoda Dictionary, Amid Thesaurus, and Glossary of Refined Words and by benefitting from Vajehyab Dictionary Browser.

In the next step, the verbs obtained from each noun in a synset, are considered as candidates for making a synset; moreover, for each verb in the synset, its synonyms can be extracted from available lexical sources to participate in the synset. After completion of the verbal synset an unlabeled relation (related-to) will be held between the original noun synset and the derived verbal synset.

Finally after completion of automatic phases, with respect to error possibility, expert supervision for synset verification would be necessary. The possible errors might comprise non prevalency (obsolescence) of the generated verb by means of grammatical rules, and or non-idiomaticness of the verb obtained by surveying in glossaries of the day. It is also possible that the obtained verb might be very specific and rarely used. For example, Persian verb *Derakhshandegi Kardan* (درخشندگی کردن) (to do brighten) that has been derived by means of grammatical rules is not correct, or verb of *KhAb Dookhtan* (خواب دوختن) (To sew sleep) that is found in glossaries is not used today. The other error is forming a synsets with words with similar structure and non-verbal part but different meanings. For example Persian phrasal verbs such as *KhAb Raftan* (to go asleep: خواب رفتن), *KhAb Beh KhAb Raftan* (to die in asleep: خواب به خواب رفتن), and *KhAb Didan* (to see or have night dream: خواب دیدن) that are all derived from Persian term *KhAb* (خواب: sleep) each one has a separate meaning. With respect to these errors, we conclude that building verbal synsets from noun synsets is not 100% automatically feasible and expert supervision and analysis would be inevitable; nevertheless, the approach used might be highly efficient in automatic extraction of new and synonymous terms and the data obtained might be efficient in reducing processing size and time spent for building synsets.

4.2 Non-Taxonomic Relation Extraction

The proposed method for extracting non-taxonomic relations employs lexical sources, various tools and combination of different linguistic, syntactic and statistical methods to improve efficiency and to increase precision and recall. This system primarily extracts a pair of concepts in semantic relation in concept pair extraction subsystem. The type and label of some of the relations can be identified during the concept pair extraction phase. For others, the labeling is postponed to the next phase and just adds the pair as "related-to" into the candidate set. These unlabeled relations will go through the labeling subsystem to determine their labels. These subsystems and their algorithms are discussed in this section.

- **The concept pair extraction subsystem**

To extract concept pairs with a semantic relation we applied three methods:

In the first method, all synsets with words containing any derivational form of a verb, Arabic rhythms, and keywords denoting a semantic role (location, instrument, agent, and patient) have been extracted from FarsNet. Then for each of the above words their corresponding verb (e.g. with the same stem) is extracted. The word and its corresponding verb make a concept pair to be

used as the input of further morphological and semantic analysis.

For instance, between concept pair of *DastgAhe Tasfiyeh HavA* (air refinement system: دستگاه تصفیه هوا) and *Tasfieh Kardan* (to refine :تصفیه کردن) there is an "instrument" relation; and between the concept of *PanAhgAh* (shelter :پناهگاه) and verbs *PanAh DAdan* (to shelter :پناه دادن) and *PanAh Bordan* (to take refuge :پناه بردن) there are "location" relations; and in concept pair of *NAzer* (supervisor :ناظر) and *NezArat Kardan* (to supervise :نظارت کردن) there is an "agent" relation. All of these relations can be extracted by morphological analysis according to derivational affixes or Arabic patterns (rhythms).

Lexico-semantic analysis of synset glosses is another technique to extract related concepts. In this method a group of lexico-semantic patterns and key phrases correspondent to each of the semantic roles has been utilized for semantic analysis of glosses. After using a verb detection module to detect simple and compound verbs in the gloss, some patterns are used to extract the relation between the synset and the detected verb. For example, the synset of *Rahbar* (leader:رهبر), *RAhnamA* (guide: راهنما) and *SarjonbAn* (mentor: سرجنبان) is defined as "*someone who leads and commands"*. Applying the agent patterns on this gloss lead to extraction of an "agent" relation between the synset and the verbs "*Hedayat Kardan* (to guide: هدایت کردن) and *FarmAn DAdan* (to command: فرمان دادن)". As another example the synset of "*HammAm* (bathroom: حمام) and *GarmAbeh* (bathhouse: گرمابه)" is defined as "*a location that is built for washing body*". Using a location pattern leads to extracting a "location" relationship between the synset of bathroom and the synset of wash (شستشو کردن).

The other approach for extraction of a concept pair participant in semantic relation is to consider all of the existing verbal synsets (concepts) in FarsNet as the first input and obtaining the second selected concept by means of the following statistical approach. The input of the statistical module is the set of all words (with all of their written forms) Then using Wortschatz statistical corpus for each verb, its co-occurrent nouns are derived and sorted according to their frequency. This way the most frequent co-occurent nouns to each verb are extracted. But we need a synset as a member of concept pair not a word. Thus we extract all the synsets which include the co-occurent noun as a candidate and at the next steps employ a Word Sense Disambiguation (WSD) module to determine the suitable sense (synset).

- **Semantic relation labeling subsystem**

In the previous steps some concept pairs (a pair of two synsets with a relation among them) were extracted and some of their relations were labeled during the extraction process. In this step we are going to extract more labeled relations or label some remained unlabeled ones. In order to enhance precision and recall in the system we employed several aforesaid sources. In this step we first find dependents (synonyms, hyper/hyponyms and instances) for the input concept pair. Then we label the relations between the concept pairs and their dependents. These two steps will be discussed in more details in the following.

- **Finding dependents for input concept pair**

In order to derive dependent for each of input concepts, we have used various sources including FarsNet synsets, Khodaparasti lexicon, and also redirect pages in Wikipedia to find synonymies and FarsNet taxonomic relations, and Wikipedia categories and subcategories to achieve hierarchical relations as father and child concepts for any concept.

We execute a shallow preprocessing on the given dependents to improve system efficiency including text normalization and unifying various word forms, omission of inflectional affixes, refinement of additional descriptors and finding of NP head especially for Wikipedia categories.

After determination of dependents, the labels are acquired for semantic relations among input concept pair and pair of dependent concepts by various techniques. The used approaches include morphological analysis, employing syntactic patterns, and adjustment of these patterns for identifying semantic roles which are discussed in the following.

- **Morphological analysis module**

We have utilized STeP-1 stemmer and morphological analyzer as the main tool in this module. This module tries to find stems and derivational affixes for any input term. We have prepared anaffix lexicon-and a rich set of morpho-patterns that covers various types of derivational affixes for combining with noun, adjective and verb stems.

Likewise, we also utilize a group of pattern or templates (rhythms) in Arabic language from which many words have been made in Persian. These rhythms include construction patterns for

gerund, noun of place, nominative noun, past participle, and noun of exaggeration. For instance for active participles of *NAzer (*supervisor: ناظر) or *TAjer* (merchant: تاجر), these gerunds are derived *NezArat* (supervision: نظارت) and *TejArat* (trade: تجارت) and they refer to "agent" semantic role. The noun of exaggeration also usually refers to a job. For example, the label for relation among *KhayyAt* (tailor: خیاط) and *KhayyAti Kardan* (to sew: خیاطی کردن) is also an agent.

Each word, after morphological analysis is examined for inclusion of an entry of the affix lexicon or obeying of Persian morpho-patterns or Arabic templates (rhythms) and if it is composed of one of meaningful derivational affixes, proportional to the semantic role, a semantic label would be attached to it. Then all the words in a concept pair and are compared with each other. If they have any common infinitive stem we may be able to extract new relations among them. For example consider that the words *ArAyeshgar* (barber-hair dresser: آرایشگر) and *ArAyeshgAh* (barber shop: آرایشگاه) appear in a concept pair. As they have the common infinitive stem *ArAyesh kardan* (hair dressing: آرایش کردن) and the first is the agent and the second is the location of this act, we can include that there is a "location" relation held in this concept pair. We have employed verbs valency lexicon in order to find the dependent stems of an infinitive.

Whereas STeP-1 stemmer does not analyze compound nouns and verbs, thus we have improved function of stemmer for morphological analysis of compound words. For example in the initial stemmer, some terms like *DAneshAmooz* (student:دانش آموز), *Ashpaz* (cook :آشپز), and *GolkAr* (gardener :گلکار) are identified as single noun or adjective words; while it will be very useful for labeling their corresponding relations, if they are analyzed into constructional terms with saving all constituent stems. We have utilized glossary of verbs to solve this problem and we check ending of noun or adjective compound words with present stems. If the compound word passes the check, we save the stems of both terms as stem of the given word and create a semantic label of "agent" for infinitive of the present stem. For example, label of relation among concept pair *GolkAr- KAshtan* (gardner-to plant: گلکار- کاشتن) would be "agent".

Similarly, input concepts that are noun phrases are analyzed in this module in terms of presence of keywords correspondent to role relations. For example, many categories are expressed in Wikipedia pages by descriptors e.g. "*VasAyel- Vasileh- abzAr- abzArAlAt- LavAzem- TajhizAt* (devices- means- tools- apparatuses-equipment:وسایل-وسیله-ابزار-ابزارآلات-لوازم-تجهیزات)" and or most of concepts in FarsNet include descriptors e.g. "*MakAn- Mahal- Zamin- Mo'asseseh- OtAgh- EdAreh- Sherkat* (place-location-land-institute-chamber-department-company:مکان-محل-موسسه-اتاق-اداره-شرکت)". Therefore, proposing an approach for morphological analysis on them increases system recall. To this end, we save any word, including one of the given descriptors with correspondent semantic label e.g. instrument and location and stem of the term after descriptor. In order to achieve its semantic relations with the other input concepts we act similar as above-said process. For example, "instrument" will be assumed as label for relation of concept pair of *TajhizAt SAkhtemAni- SAkhtan* (constructional equipment- to build: تجهیزات ساختمانی- ساختن).

- **Syntactic analysis module and dependency analysis**

Dependency treebanks include a group of sentences which have been analyzed according to dependency command, and generally verb of sentence is selected as root and origin and the relation of other words of the sentence with each other and the verb would be characterized. These corpora are considered as rich sources for finding deep syntactic patterns and the resulted processing would be highly accurate; though frequency of occurrence and recall in them is not that much high.

The studied concept pair is analyzed in terms of nature (being noun or verb) after entering into this module; for this purpose we employ stemmer and also utilize lexicon of verbs to identify the compound verbs. Then, we survey corpus to find sentences including both of them. Whereas the concept may occur in corpus in singular or plural form, or other inflection such as a noun preceding an unknown Persian article (Ya-e-Nakareh: ی نکره) and also Dadegan dependency treebank comprises of root of words in sentence, therefore, input concepts are compared with the specified roots in corpus as well.

By finding dependency of noun on verb and application of some rules and conditions and adjustment of semantic patterns to syntactic patterns, we label these relations for semantic role of noun to input verb. For instance, if the dependency of concept-to-verb relation is of subjective and the given verb is of active voice the label of conceptual relation or semantic role will correspond to agent, and if the verb is of passive

voice the label will be of patient type. For example, in this sentence: *"Flags were hoisted as symbol of lament"*, the concept of "flag" will have role of "patient" for the concept of "to hoist". Likewise, the additional composition including infinitive is examined with left and right neighbors; for example, the label of hidden relation in additional composition *FAsh Kardan RAz* (to disclose secret: فاش کردن راز) will denote "patient".

In order to find supplementary relations and to increase precision and efficiency of labeling system, if a noun is related to a preposition with a verb, that preposition is also used for semantic analysis and identifying of label of relation. For instance, Persian prepositions like *BA-Dar-Az-Tavasot-Bevasileh* (with- in- from- to- via- by: با- در- از- به- توسط- به وسیله) can represent various semantic roles, for example label of role relation for concept pair *'Goldoozi Kardan-Charkh-e-KhayyAti' (*needlework- sewing machine: گلدوزی کردن- چرخ خیاطی) with respect to the presence of preposition *Ba* (by: با), through participation with them and using of semantic category and semantic analysis of the gloss for concept of *"sewing machine"* would be determined as "instrument".

The other technique which has been designed in this module to determine semantic relation among input concept pair comprises of using ParsiPardaz tool for dependency analysis of example sentences of any synset in Synset table of FarsNet database. After dependency analysis of these sentences, we act as what was mentioned above and determine label of relation by adjustment of syntactic and semantic patterns.

- **Word sense disambiguation module**

Finally, after identifying and labeling conceptual relations among a concept pair, it is necessary to adapt a method for selecting the best and most appropriate synset for ambiguous words. To this end, a method has been designed that preserves recall and efficiency of the system while having reasonable precision. In this technique, we primarily select the appropriate synset among candidates according to their semantic categories and its relation to the label of the identified conceptual relation; for instance, if we embed word Cinema (سینما) in a "location" relation we expect that its corresponding synset has *location* in its semantic category.

In the next step, we apply a Lesk-like algorithm for WSD. To find the most appropriate synset for a polysemous word or for a new synset to be merged with, we compare the word (or words in the new synset) with the words in the gloss and example of candidate synsets after omitting the stopwords; the synset with more common words is more appropriate.

The precision of this method is low when the candidate synsets have just one word or if the candidate synsets are semantically similar and so there is textual similarity between their glosses and examples. In these cases human supervision is needed to resolve the ambiguity. For instance, there are several synsets semantically close together for these words *NaghAsh* (painter :نقاش) and *Rang Kardan* (to paint :رنگ کردن) or words of *BAnk* (bank :بانک) and *Poul* (money :پول) that makes difficult automatic recognition of the most appropriate synset. therefore presence of these commonalities in glosses and examples of all of them makes automatic recognition of the most appropriate synset difficult.

5 Results and Conclusion

This paper discusses the application of various automatic linguistic, syntactic and statistical methods on various resources to extend FarsNet by a merge method. The proposed method not only has reasonable precision and coverage, but also covers culture and language specific concepts and relations which cannot be captured by expansion methods. It can either extend the existing verbal synsets by a new verb or create a new synset for new and specific verbs of Persian lexicons with metaphorical meanings

This strategy significantly increases recall and the number of verbs and extracted semantic relations. Although it is applied to Persian, it can be used for extracting and labeling semantic relations in other languages as well.

The experimental results show that the proposed verb-extraction method, extracts 6890 correct verbs - regardless of polysemy and number of senses for each word and add them to FarsNet 2.0–that already had 7820 verbs. The synset extraction method added 2790 verbal synsets to 3670 verbal synsests existing in FarsNet 2.0. The synsets need manual judgment and semantic disambiguation of senses by lexicographers. Table 2 demonstrates the results of the proposed method for verbal word and synset extraction. The results show that the hybrid method (using structural rules plus digital lexicons) significantly increases both the number of extracted verbs and their precision; however using lexicons decreases the precision of results while increasing the number of correctly extracted synsets. This happens due

to polysemous words with different meanings in a synset.

Verb extraction approach	No. of correctly extracted verbs	Precision for verb Ex.	No. of Correctly extracted synsets	Precision for syn-set Ex.
Applying structural rules	750	79%	396	76%
Applying structural rules and digital lexicons	6890	91%	2790	67%

Table 2: Number of correct words and synsets and precision of the proposed method for verb and synset extraction

The given results for automatic extraction of non-taxonomic relations contain 5600 correct relations among existing synsets in FarsNet 2.0; with accuracy rate of 76%. FarsNet 2.0 had 1040 semantic relations (excluding hyper/hypo-nymy, domain, and holo/mero-nymy) before applying the proposed strategy which formed only about 2.8% of the relations in FarsNet 2.0. This rate reached to 15.7% after implementation of the suggested method. Thus, the proposed automatic method has efficiently contributed to improve the number of non-taxonomic relations corresponding to thematic roles and co-occurrence relations and reduced size of manual processing for relation extraction. The presented strategy still leads to extraction of further and more accurate conceptual relations by increase in number of synsets and words and examples for each of the concepts by extension of FarsNet.

References

Al Tarouti, F. a. (2016). Enhancing Automatic Wordnet Construction Using Word Embeddings. *In Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*.

Bagherbeygi, S., & Shamsfard, M. (2012). Corpus based Semi-Automatic Extraction of Persian Compound Verbs and their Relations. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, (pp. 2863-2867). Istanbul.

Boudabous, M. M. (2013). Arabic wordnet semantic relations enrichment through morpho-lexical patterns. *In Proceeding of 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA).*

Dehkharghani, R., & Shamsfard, M. (2011). *Bilingual Ontology Mapping.* Germany: LAMBERT Publisher.

Fadaei, H., & Shamsfard, M. (2010). Extracting conceptual relations from persian resources. *Seventh International Conference on Information Technology: New Generations (ITNG)*, (pp. 244-248).

Girju, R. (2008). Semantic relation extraction and its applications. *ESSLLI*.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *LREC*, (pp. 759-765).

Hwang, C. L., & Yoon, K. (1981). *Multiple Attributes Decision Making Methods and Applications.* Berlin: Springer.

Jafarinejad, F., & Shamsfard, M. (2012, March). Extracting Generalized Semantic Roles from Corpus. *IJCSI International Journal of Computer Science Issues, 9*(2).

Kavalec, M., & Svátek, V. (2005). A study on automated relation labelling in ontology learning. (P. Buitelaar, P. Cimiano, & B. Magnini, Eds.) *Ontology learning from text: Methods, evaluation and applications*, 44-58.

Khodaparasti, F. (1997). *A Comprehensive Dictionary of Persian Synonyms and Antonyms.* Shiraz: Daneshnameye Fars.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *hlt-Naacl, 13*, 746-751.

Mousavi, Z. a. (2017). Persian Wordnet Construction using Supervised Learning. *arXiv preprint arXiv:1704.03223*.

Prabhu, V., Desai, S., Redkar, H., Prabhugaonkar, N., Nagvenkar, A., & Karmali, R. (2012). An efficient database design for IndoWordNet development using hybrid approach. *In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language*, (pp. 229-236). Mumbai, India.

Rasooli, M. S., Kouhestani, M., & Moloodi, A. (2013). Development of a Persian Syntactic Dependency Treebank. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT).* Atlanta, USA.

Rasooli, M. S., Moloodi, A., Kouhestani, M., & Bidgoli, B. M. (2011). A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank. *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, (pp. 227–231). Poznań, Poland.

Saaty, T. L. (1980). *The Analytic Hierarchy Process.* New York: McGraw-Hill.

Sánchez, D., & Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data and Knowledge Engineering*, *64*(3), 600–623.

Sarabi, Z., Mahyar, H., & Farhoodi, M. (2013). ParsiPardaz: Persian Language Processing Toolkit. *Computer and Knowledge Engineering (ICCKE)* (pp. 73-79). IEEE.

Shamsfard, M., & Barforoush, A. A. (2004). Learning Ontologies from Natural Language Texts. *Int. J. Hum.-Comput. Stud.*, *60*(1), 17-63.

Shamsfard, M., & Ghazanfari, Y. (2016). Augmenting FarsNet with New Relations and Structures for verbs. *8th Global Wordnet Conference.* Bucharest.

Shamsfard, M., & Mousavi, M. (2008). Thematic role extraction using shallow parsing. *International Journal of Computational Intelligence , 4*(2), 126-132.

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.

Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. *LREC.*

Taghizadeh, N., & Faili, H. (2016). Automatic Wordnet Development for Low-resource Languages Using Cross-lingual WSD. *J. Artif. Int. Res.*, *56*(1), 61-87.

Taghizadeh, N., & Faili, H. (2016). Automatic Wordnet Development for Low-Resource Languages using Cross-Lingual WSD. *J. Artif. Int. Res.*, *56*(1), 61-87.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic.* Springer.

Zadeh Khosravi Forooshani, P., & Rezaei Sharifabadi, M. (2016). Automatic semantic role labeling of Persian sentences by the aid of dependency Treebank. pp. 27-38.