Extracting Generalized Semantic Roles from Corpus

Fateme Jafarinejad[1], Mehrnoush Shamsfard[2]

[1] **Faculty of Electrical and Computer Engineering, Shahid Beheshti University, Iran
Tehran, Iran**

[2] **Faculty of Electrical and Computer Engineering, Shahid Beheshti University, Iran
Tehran, Iran**

Abstract

One of the oldest constructs of linguistic theory is semantic role. Automatic extraction of semantic roles in a sentence is a movement towards semantic processing of texts which has been the focus of attention in recent years. Extraction of semantic roles from a text contains some essential parts. Recognition of verb(s) of the sentence, recognition of noun phrases and their heads, and labeling the role of each phrase in the sentence as a semantic argument of verb are general parts of a system that does this task. There is a wide variety of definitions for semantic roles from verb specific roles to some general roles known as thematic roles, This paper focuses on a generalization of thematic roles called proto-roles or generalized semantic roles which includes two roles; actor and undergoer. In this paper we extract proto-roles in a Persian sentence exploiting POS tags. We use Peykareh as our input corpus and apply a rule based approach to extract actor and undergoer of verb(s).

Keywords: *natural language processing, semantic role labeling, , predicate-argument extraction, proto-roles extraction.*

1. Introduction

Extracting the relationships between various constituents in a sentence can be a useful task in many natural language processing applications. The relations may be syntactic and grammatical (as in dependency trees) or semantic (as in case frames and semantic roles). One of the major semantic relations in a sentence is the relation between the verb and its semantic arguments. This task which is also called semantic role labeling (SRL) plays an important role in semantic processing of texts and provides useful information for applications which are related to semantics such as question answering [1,2], information extraction [3], text understanding [4], Machine translation [5,6], Automatic text summarization [7] and coreference resolution[8].
Semantic role labeling is important because in expressing an event, the verb is usually the central point of attention and focus. It is the predicate and we look for its arguments to represent the semantic of the event.

Semantic roles are discussed in three different levels:

(1) Verb specific roles such as runner, recognizer, etc. These roles are different for different verbs.
(2) Thematic roles which are the generalization of the verb specific roles. They are verb-independent roles such as agent, patient, force, goal, etc.
(3) The generalized semantic roles (GRLs) or proto-roles or macroroles which are a generalization of thematic roles. GRLs include just two roles, actor or proto-agent, and undergoer or proto-patient [9, 10].“Actor” is the person or instrument or force or anything that does the event, and “undergoer” is the thing that takes impact of event.
Undergoer can be mapped to thematic roles such as theme, patient, recipient, etc according to the meaning of verb. And actor can be mapped to thematic roles such as agent, cognizer, experiencer, force, instrument, etc.

According to linking theory [11], which indicates that syntactic arguments of a sentence can be predicted from its semantic arguments, there have been some mechanisms developed to extract syntax from semantic. Knowing this fact, and relationship between syntax and semantic some researchers suggest that it may be possible to recognize semantic relationships from syntactic cues too. For example Gildea[12] suggests a system that learns these relations according to syntax.
Nowadays learning approach is widely used in semantic role labeling. Different systems use variety of features and classifiers to do labeling. Example of such systems are introduced in [13, 14, 15] and many other papers.
On the other hand, examples of rule-based approach apply some hand written rules to extract semantic roles. For example [16] uses a head driven phrase structure grammar for this task and [17] uses deverbal nouns information to do so.

Although the interest in exploiting learning approaches in SRL is growing, it is not an appropriate approach for our case, as there is no semantic role-labeled corpus available for Persian language at the moment. Thus we exploit a rule base model which can also be used for developing such a corpus for further researches.

In this paper according to the linking theory and syntax-semantic relationship, we use a rule based approach to extract generalized semantic roles from a POS-tagged corpus for Persian language.

Organization of the rest of the paper is as fallows. Section 2 introduces Peykareh, the Persian corpus we exploited in our system.. In section 3 we discuss the structure of Persian sentences and the rules dominating verbs, noun groups and sentences in Persian. Section 4 discusses our relation extraction (SRL) system which extracts the verbs and their generalized semantic roles in a sentence. Results will be explained in section 5. And section 6 contains the conclusion and future work.

2. Peykareh: the Persian corpus

Peykareh [18] also known as Bijankhan corpus (according to the name of its developer) contains more than 7 million tagged tokens. In this work, we used a subset of this corpus with about two million words which was available to us freely. Each word is annotated with its part-of-speech tags.. ◌́As the initial tag set is very large, we reduced it to some few useful tags which are essential for our task. The reduced tag set is shown in Appendix 1.

In Bijankhan corpus, each single word is annotated independently, so the tagging system has problem with compound words and verbs. For example the components (verbal and non-verbal) of a compound verb or even the parts of an inflection of a simple verb (which is a multi-token word) may have tags independently. Thus recognizing these words is not an easy and straight forward task.

In the new edition of corpus, versus the old one, we have a different tagging for adjectives as noun modifiers and adjectives as non-verbal parts of verbs.

Also, in this edition number determiners have a new tag which distinguishes them from nouns. and information about the time and person of verbs are added.

As every complete sentence in the Bijankhan corpus is finished with a punctuation "." or "?", segmenting the text into sentences is simple. Compound sentences have more than one verb. Each verb in the corpus is finished with a "V" tag. But detection of whole verb group and detection of the stem of the verb have some processing steps.

In appendix we list tags of Bijankhan corpus in each category which we use in this work.

3. Persian sentence structures

Like other languages, in Persian verb is the core of the sentence. Each complete sentence must have at least one verb. A verb has a subject and may have zero, one or two objects as its related arguments. We assume that subjects and objects are noun phrases. Noun phrases should have a head (which is anoun) and may have some modifiers(nouns, adjectives, determiners, numericals, …). The head of a phrase is the essential part of it which shows the main element of the phrase. The head of a phrase is the noun which should be related to other constituents (such as verb) semantically and should satisfy the selectional restrictions and unification constraints. For example the head of the object phrase of the verb 'to eat', must be edible.

Besides noun and verb groups, ,there are another kinds of groups such as prepositional groups and adverbs. These groups are not useful in the extraction of actor and undergoer of verb, but may have general and extra information about the type of the event indicated by the verb of sentence.

Each verb in Persian is in one of the tenses of past, present and future. Each tense has some types. Also, a verb has a person and number. According to these features and negativeness or positiveness, the verb of sentence is constructed from its past/present stem. In Bijankhan corpus information are encoded via tags. To reach to the stem of the verb, these information and rules must be considered. In the rest of this section we describe some of the features of Persian sentences which we use in our work.

3.1 Passive/Active sentences

Each verb has a subject that do it, but sentence may be written in a manner that subject is not declared in it. These sentences are called passive. Another kind of sentences which has an explicit subject that is cited in the sentence or is eliminated with an indication is called active sentence. In some cases in passive sentences the subject is indicated after some prepositions such as "توسط، بوسیلهٔ"[1] (by). In an active sentence the subject is a candidate to have the actor role. But in a passive sentence possibility of being undergoer is higher.

3.2 Transitive/Intransitive verbs

Verbs, according to their requirement to object, are divided into two categories. Transitive verbs must have an object, and their object is eliminated scarcely.

The object marker in active sentences is 'ra' (را). But it may be eliminated from the sentence in some cases.

Some kind of transitive verbs are content based which means that their undergoers are sentences. The undergoer of such verbs- if not explicitly explained before them, followed by the object marker- is a sentence which is appeared after the verb. Example of such verbs are "گفتن" (say), "فهمیدن" (understand), and "احساس کردن" (feel).

[1] Throughout the paper we show Persian examples within double quotes followed by their translation within parentheses.

3.3 Compound/Simple sentences

In Persian, like many other languages we have two kinds of sentences; compound and simple. Compound sentences are composed of some sentences that don't have an independent meaning, so in compound sentences we have more than one verb. In these sentences subject of a verb can be eliminated because of similarity and redundancy of it with the subject of the previous verb.

Compound sentences are divided into two categories, Simple and complex. Simples are connected to each other by some conjunctions such as "و"(and) and "یا"(or) serially. For example the sentence "علی به مدرسه رفت و حسن را دید." (Ali went to school and saw Hassan.) is a simple compound sentence. On the other hand, complex compounds are those in which a sentence includes another sentence. In this case usually the inner sentence is attached to a constituent (say noun) of the outer sentence by conjunctions such as "که" (which, that). For example the sentence "علی کتابش را که روی میز بود برداشت." (Ali took his book which was on the table.) is of this kind.

In compound sentences the actor or undergoer of one verb can be omitted and can be extracted from the actor and undergoer of the other verbs in the sentence according to some features such as similarity of verbs in transitivity and voice.

4. Extracting the actor and undergoer of a verb

The task of extracting the actor and undergoer of a verb includes the following steps:

- Chunking the sentence into its phrases and constituents,
- Detection of the verb(s) and their attributes (tense, person, number, active/passive, transitive/ intransitive, …)
- Recognize the candidate noun phrases and their heads in the sentence. These are the noun groups which have potential to be the actor or undergoer of the verb,
- Assigning the proto-roles: This step assigns the role of actor or undergoer to the head of the appropriate candidate noun phrase according to the structure of sentence and the feature of its verb.

In compound sentences a more complicated analysis should be done to find the actor or undergoer of a verb that is omitted in one clause but exists in another clause. Figure 1 shows the overall schema of the system.

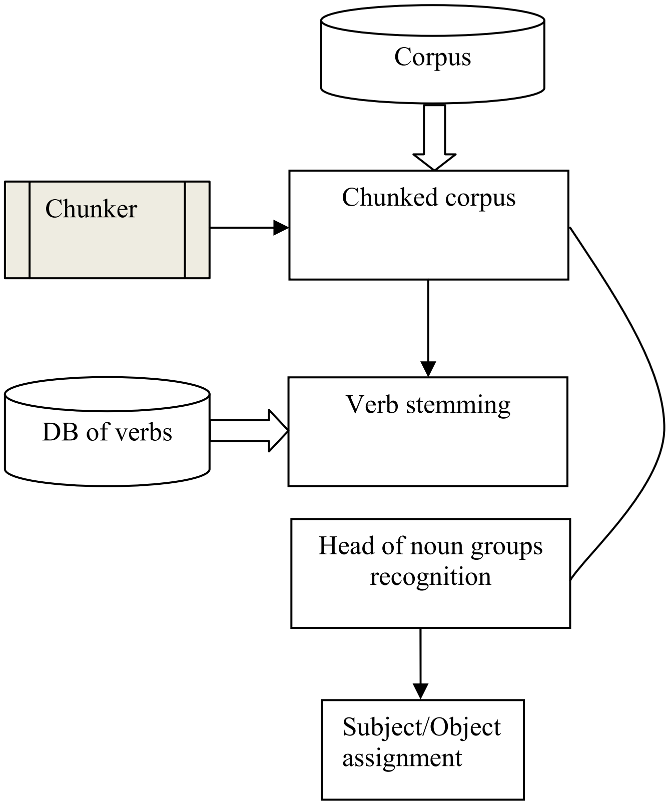

Fig. 1 Stages of generalized semantic role extraction system from text.

4.1 Detection of independent groups

One essential part of a semantic role extraction system is recognition of noun phrases in a sentence. These noun phrases or groups are the candidates for filling the semantic requirements of verb(s) by playing a semantic role of that verb. For detecting the noun phrases we use the Persian chunker [19] developed by NLP lab of Shahid Beheshti University [20]. The chunker performs shallow parsing and chunks the constituents by IOB tags . The output of the chunker is a text file that has sentences and tags of Bijankhan corpus and two extra columns showing that each word is in the beginning, inside or outside of a group.

4.2 Detection of a verb and its stem

As we explained former, a verb in Persian is constructed from its past/present stem according to some features such as tense, type, and person. These features cause addition of affixes in some parts of the verb. So to detect a verb we should do stemming to extract the verb stem from its inflectional form.

As Bijankhan corpus tags each part of a compound verb and each token of a verb inflection separately, we cannot trust the tokenization done in the corpus and we have to detect the verb boundaries by ourselves. To do this task, we first look for a word with the verb tag in the sentence. Then starting from the verb and going backward in the sentence we look for any word that can be concatenate to the found verb to make a compound verb or an inflection of the verb. Then removing affixes, we search the root in the database of Persian verbs. If the verb exists, we extract its transitivity information. If the verb is in DB so the verb is active, but if the verb by removing its passive affiliate is in the DB the verb is in a passive sentence.

Knowing these is useful in analysis of the structure of sentence to extract the subject and object.

4.3 Detection of candidate noun phrases and their heads

Candidate noun phrases, in our procedure, are the ones that have potential to be the subject or object of a verb in the sentence, so the adverbial phrases and prepositional phrases are eliminated form this set.

As the input of the system is POS tagged, we can simply recognize the nouns in the noun phrase. But for detection of the head of the noun phrase we use some rules. Each phrase includes a head and some pre- and post- nominal modifiers.

Head of a noun group is usually the first noun of the group. But in some cases for example when we have a unit for a thing it is not true, like one meter of fabric in which the head is ‘fabric’, not ‘meter’. Also in groups which are composed of some smaller groups we may have more than one head.

After detection of the head, it should be gone under some sort of lemmatization and morphological analysis to extract its lemma. In this process we remove the plural sign, indefinite markers and so on. Knowing the stem of the head can improve our work in cases that we want this information to be used in other works such as extracting category of verbs’ parameters, and WordNet enrichment.

In this part we use the stemmer included in STeP-1 [21].

4.4 Semantic roles identification

The last part of our approach is assigning the appropriate semantic roles to the heads of the appropriate candidate noun phrases in the sentence. For this task we have defined some rules which use some information such as transitivity or intransitivity of verbs, activeness or passiveness of sentence, selectional restrictions of verb arguments and so on.

In the case of compound sentences if the actor or undergoer of a verb is not found in its clause, these roles can be discovered from the adjacent clauses. They may be the actor and undergoer of the previous verbs in the sentence, or some nouns that appear beyond the inner clause of a complex compound sentence.

In our approach we must first recognize the subject and object of the sentence and then find the actor and undergoer.

Some of the rules for finding the subject and object are following:

- Intransitive verbs don’t have object.
- The object of content-bearing verbs such as "گفتن" (say) can be indicated by an object marker (را) as other transitive verbs. Otherwise if there is no direct object, the object may be a content presented by a sentence which usually appears after the clause containing the verb.
- Formal Persian has SOV ordering so, subjects usually come before objects.
- Probability of omitting a subject, without a sign, is higher than omitting the object.
- In simple compound sentences, sometimes the subject and object of a verb can be distinguished from the subject/object of the previous verb.
- In complex compound sentences, subject and object of the verb of the inner sentence can be recognized from the noun groups which are beyond the inner sentence.

And some of the rules for finding actor and undergoer from subject and object are following:

- Subject can be actor in active voice verb or undergoer in passive voice verbs.
- Subject of Intransitive verbs can be either actor – such as in verb "خندید" (laugh) or undergoer such as in verb "افتاد" (drop). Default selection of our system is actor because of confliction with other mechanisms that find a passive verb.
- if the actor and undergoer of a verb are extracted from the previous clause (verb) and the verbs don’t have similar voice properties then the actor and undergoer may be exchanged.

5. Experimental Results

As there is no semantic role labeled corpus for Persian to be used as a golden standard, we should do the evaluation of our system manually. We use precision and recall to report our systems accuracy.

Precision is the ratio of the number of correct roles distinguished by our system to the number of responses that the system gives us.

Recall is the ratio of the number of correct roles distinguished by the system to the number of responses that must be returned by the system.

We calculate the precision and recall in case of subject recognition, object recognition, separately and collectively. We evaluate our system in a small part of the corpus. Table 1 shows the experimental results.

Table 1: Precision and recall of our work.

	actor	undergoer	Overall accuracy
precision	75	60	72
recall	75	95	80

An example of input sentence and output of the system is shown in table 2. In this example, the input sentence is: "در این گونه موارد اغلب اعمال قضاوت حرفه ای ضروری است و هیچ گونه زیانی نباید شناسایی گردد مگر آنکه شواهد موجود آشکارا نشان دهد که زیان واقع شده است." In these cases usually applying (professional judgment is necessary and no kind of loss should be recognized unless available evidences clearly show that loss is occurred.)

The first column of table 2 shows the tokens of this sentence, each in a row .
This column has 3 inner columns which are tagging information of word, word in Persian, and English translation of word, respectively.
The second column shows the result of semantic role extraction. Each verb of the sentence is written in a line, followed by its semantic argument(s) if exist. This column includes semantic role information of the phrase (verb, or its arguments, actor, undergoer), the head of the phrase in Persian, and its English translation .

Table 2: A sample input and output of the system.

input	Output
P در (in) DET,DEMO این(these) CL,SING گونه (type) N,COM,PL موارد (cases) ADV,GENR,SIM اغلب(often) N,COM,SING,EZ اعمال (applying) N,COM,SING,EZ قضاوت (judgment) AJ,SIM حرفه اي(professional) AJ,SIM ضروري (necessary) V,COP,PRES,POS,3 است(is) CONJ و (and) DET,INDF هیچ (no) CL,SING گونه (kind of) N,COM,SING,YA زياني(loss) V,AUX,NIN,NEG نباید (should be) N,COM,SING شناسائي V,SUB,POS,3 گردد (recognized) CONJ مگرآن که (unless) N,COM,PL,EZ شواهد (evidences) AJ,SIM موجود (available) ADV,GENR,SIM آشکارا(clearly) N,COM,SING نشان V,SUB,POS,3 دهد (show) CONJ که (that) N,COM,SING زیان (loss) AJ,SIM واقع V,PASTP شده V,COP,PRES,POS,3 است (is occurred) PUNC .	to) بودن:verb (be اعمال:actor (applying) شناسائی:verb کردن (recognize) زیان:undergoer (loss) نشان دادن:verb (show) شواهد:actor (evidences) -:undergoer جمله-که زیان واقع شده است -) content- that loss is (occurred واقع شدن:verb (occurred) زیان:actor (loss)

6. Conclusions

Subject and object of a verb are two useful parts of sentence in explaining the event that the verb of sentence expresses it. These roles are useful in semantic analysis of a sentence and can guide in knowing semantic roles.
We have presented an approach for automatically extracting subject and object of a verb in Persian sentences using Bijankhan corpus as input. Then we extract the generalized semantic roles according to the extracted grammatical roles using a rule based approach. Experimental results show admissible results.

Appendix

Table 3: Set of tags of Bijankhan corpus that are used in our work.

Category	Tags	Explanation & usage
Verb(V)	1,2,3,4,5,6	Verb stem rec.
	PA, PRES, PASTP, IMPERF, SUB, PREF	Tense & type of verb Verb stem rec.
	NEG, AUX,	Verb stem rec.
Noun(N) Pronoun (PRO)		Recognizing head of noun group.
Adjective(AJ)		Recognizing the head of noun phrases and verb stemming.
Adverb(ADV)	TIME, LOC, SIM	Recognizing the head of noun phrase.
Preposition (PP)		Recognizing the candidate noun phrases.
Conjunction (CONJ) Punctuation (PUNC)		Recognizing the borders of simple and compound sentences.

Acknowledgments

We would like to thank Ms. Noferesti for providing the chunker, and also Ms. Jafari for sharing the stemmer.

References

[1] D. Shen, M. Lapta, "Using Semantic Roles to Improve Question Answering", ACL Anthology Network, 2007.

[2] M.W. Bilotti, P. Ogilvie, J. Callan, E. Nyberg, "Structured retrieval for question answering", Proceedings of SIGIR 2007, 2007, PP. 351–358.

[3] S. Harabagiu, C. A. Bejan , P. Mor˘arescu, "Shallow Semantics for Relation Extraction", IJCAI'05 Proceedings of the 19th international joint conference on Artificial intelligence, 2005.

[4] B. Coppola, A. Moschitti, G. Riccardi, "Shallow semantic parsing for spoken language understanding", NAACL-Short '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009.

[5] D. Liu, D. Gildea, "semantic role features for machine translation", COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, 2010.

[6] S. Wu, M. Palmer, "Semantic Mapping Using Automatic Word Alignment and Semantic Role Labeling", SSST-5 Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2011.

[7] G. Melli, Y. Wang, Y. Liu, M. M. Kashani, Zh. Shi, B. Gu, A. Sarkar, F. Popowich, "Description of SQUASH, the SFU question answering summary handler for the duc-2005 summarization task", In Proceedings of the HLT/EMNLP Document Understanding Workshop, 2005.

[8] P. S. Ponzetto, M. Strube, "Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution" In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2006, pp. 192–199.

[9] Robert D. Van Valin, "Generalized Semantic Roles and the Syntax-Semantics Interface", 1999.

[10] D. Dowty, "Thematic Proto-roles and Argument Selection", language 67, 1991, pp. 547- 619.

[11] B. Levin, M. R. Hovav, "From lexical semantics to argument realization", 1996.

[12] D. Gildea, D. Jurafsky, "Automatic Labeling of Semantic Roles", In Association for Computational Linguistics, 2002.

[13] J. LI, R. Wang, W. Wang, B. Gu, G. Li, "Automatic Labeling of Semantic Role on Chinese FrameNet Using Conditional Random Fields", IEEE, Computer Society, 2009, PP. 259-262.

[14] H. Shi, G. Zhou, P. Qian, X. Li, Semantic Role Labeling based on dependency Tree with multi-features", IEEE, Computer Society, 2009, PP. 584-587.

[15] M. Surdeanu, L. Marquez, X. Carreras, P. R. Comas, "Combination Strategies for Semantic Role Labeling", Journal of Artificial Intelligence Research 29, 2007, PP. 105-151.

[16] R. D. Levine, W. D. Meurers, "Head-Driven Phrase Structure Grammar Linguistic Approach, Formal Foundations, and Computational Realization", Encyclopedia of Language and Linguistics, Second Edition, Elsevier, 2006.

[17] L. M. SÁNCHEZ, "Deverbal Noun Complementation Rules Applied to Semantic Role Labeling", Vol. 7, No. 2, 2008, PP. 9-42.

[18] http://ece.ut.ac.ir/dbrg/bijankhan.

[19] http://nlp.sbu.ac.ir.

[20] S. Noferesti, M.Shamsfard, Developing a Persian Chunker, Technical Report, NLP Lab, Shahid Beheshti University, 2011.

[21] M.Shamsfard, H.S. Jafari, “STeP-1: A Set of Fundamental Tools for Persian Text Processing”, LREC 2010- 8th Language Resources and Evaluation Conference, 2010.

Fateme Jafarinejad is currently pursuing the MS degree at faculty of Computer Engineering in Shahid Beheshti university. Her research interests are in the fields of natural language processing and semantic role labeling.

Mehrnoush Shamsfard has obtained her PHD in artificial intelligence from AmirKabir University of Technology. She is currently with the faculty of Computer Engineering in Shahid Beheshti university. Her main fields of interest are natural language processing with a special focus on semantics, ontology engineering, knowledge extraction and semantic web.