Relation Learning from Persian Web: A Hybrid Approach

Hakimeh Fadaei[1], Mehrnoush Shamsfard[1]

[1]NLP Research Laboratory, Faculty of Electrical & Computer Engineering,
Shahid Beheshti University, Tehran, Iran
ha.fadaee@mail.sbu.ac.ir, m-shams@sbu.ac.ir

Abstract. In this paper a hybrid approach is presented for relation extraction from Persian web. This approach is a combination of statistical, pattern based, structure based and similarity based methods using linguistic heuristics to detect a part of faults. In addition to web, the developed system employs tagged corpora and WordNet as input resources in the relation learning procedure. The proposed methods extract both taxonomic and non-taxonomic, specific or unlabeled relations from semi-structured and unstructured documents.

In this system, a set of Persian patterns were manually extracted to be used in pattern base section. Similarity based approach which uses WordNet relations as a guide to extract Persian relations uses a WSD method to map Persian words to English synsets. This system which is one of the few ontology learning systems for Persian showed good results in performed tests. In spite of resource and tool shortage in Persian the results were comparable with methods proposed for English language.

Keywords: Relation learning, knowledge extraction, ontology learning, web, Wikipedia, similarity, Persian.

1 Introduction

Automatic extraction of semantic relations is a challenging task in the field of knowledge acquisition from text and is addressed by many researchers during recent years. As ontologies are widely used in many branches of science, building or enriching them is of great importance. Automatic methods for performing these tasks are so welcome since the process of building ontologies manually is very time consuming. There are no available ontologies for Persian and not much work is done on automatic extraction of ontological knowledge for this language.

In this paper we present a hybrid approach for extracting taxonomic and non-taxonomic relations from Persian resources. In the proposed system our focus is on using web as the learning resource although we use other resources to increase the effectiveness of our system too. The paper is organized as follows: In section 2 a brief review over related works is presented. The third section is dedicated to describing

© A. Gelbukh (Ed.)
Special issue: Natural Language Processing and its Applications.
Research in Computing Science 46, 2010, pp. 219-230

Received 29/11/09
Accepted 16/01/10
Final version 10/03/10

our system and different methods used in it. Finally performed tests and their results are described in section 4.

2 Related Work

Automatic extraction of conceptual relations has attracted many attentions and some researchers work on proposing more efficient strategies in this field. During recent years different approaches were proposed to extract taxonomic and non taxonomic relations from different resources.

Pattern based, statistical, structure based and linguistic approaches are well-known approaches which extract relations from texts. Many systems (2, 3, 8) use combinations of these approaches to accumulate their advantages .

Pattern matching methods are widely used in extracting taxonomic and non-taxonomic relations. In this category, Hearst patterns [1] are among the most famous patterns defined for extracting taxonomic relations and has been used or adapted in many ontology learning systems 2, 3. Patterns may be defined manually [3] or extracted automatically 4. Some other systems (5, 14) use document structures to extract relations, these structures include tables, hyperlinks, html and xml tags and so on.

Some systems use statistical methods and rely on the distribution of words and their modifiers in text to extract relations 3, 6, 7 and 8. Linguistic structure of sentences is another source of information used in some systems 2, 8 and 9. Linguistic methods use morphological, syntactic and semantic analysis to extract relations. These methods need many linguistic tools like chunkers, taggers and parsers and are not easily used in languages such as Persian in which these tools are unavailable.

On the other hand, ontology learning systems use different resources to extract ontological knowledge, these resources include structured, semi structured on unstructured data. Raw or tagged texts are used by many systems such as 3, 5 and 6. Tagged corpora are proper resources for knowledge extraction as they are targeted for this task. These tags (POS, semantic or …) helps systems to better detect relations but they are not available in all languages. During recent years many researchers are attracted to web as knowledge extraction resource.

The main reason that attracted the attentions of researchers to web documents for ontology learning is the huge amount of text in many languages that is available for everybody. Apart from availability and size there are some other features in web documents which make them suitable for the task of ontology learning and specially relation extraction.

Web documents are usually filled with structured and semi-structured data, tables and hyperlinks which can be used in the process of ontology learning, some systems like 5 and 14 use these structures to learn ontological knowledge. One of the other facilities provided by web is the existence of high performance and efficient search engines like Google which are used to search this large body of texts. Many systems like 2, 4, 8, 11 and 12 use search engines in the procedure of relation extraction. Wikipedia is another advantage of using web in ontology learning, since it has a collection of very informative short articles well suited for knowledge extraction. In systems like [5], 8 and 13 Wikipedia is used in relation extraction.

Beside the positive points mentioned above for web as a resource in the task of learning, web has some shortages as well. Web documents are mainly written by ordinary people with no NLP background and as they are not basically targeted for NLP applications they may need special processes in comparison with the corpora prepared by language experts.

3 The Proposed Hybrid Approach

Our proposed approach is a combination of statistical, pattern based, linguistic, structure based and similarity based methods. These methods may be used in a serial or parallel order. In serial (sequential) order the output of one is the input of the other while in the parallel approach each method extracts some relations and then we choose the best one by voting. They can also be used separately to extract different types of relations. In this approach we use web to a great extent to be able to use its advantages, but to cover the dark sides of using web as a resource, we also use other resources such as corpora, dictionaries and WordNet. In this section we will describe each method in more details.

3.1 Structure Based Approach

The structure based part of our system uses the Wikipedia pages' structures such as tables, bullets and hyperlinks to extract relations. In many Wikipedia documents we can find some information given via bullets. This information usually shows some taxonomic relations. Given a Persian word the system follows these steps to extract relations from bulleted text:

1. Extract the Wikipedia article of the given word.
2. Find the bulleted parts and extracts their items.
3. Refine the items extracted in the second step by omitting stop words and prepositional phrases, finding the heads of nominal groups and .....
4. Make new relations between the title of bulleted part and the results of the third step.

The translations of some of these relations are presented in table 1.

Table 1. Translation of some extracted relations from bullets

Isa(Versailles , historical place)
Isa(car accident, event)
Isa(Islam, religion)
Isa (hypertension, disease)
Isa(suicide, death)

Disambiguation pages in Wikipedia are also so helpful in extracting taxonomic relations. 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.

While searching the word "tree" in Wikipedia, in disambiguation page we come across the following meanings of the word "tree":

- Tree is a woody plant
- Tree structure, a way of representing the hierarchical nature of a structure in a graphical form
- Tree (data structure), a widely used computer data structure that emulates a tree structure with a set of linked nodes
- Tree (graph theory), a connected graph without cycles

From the above explanations we can extract the relations:

- isa(tree, woody plant)
- isa(tree, way of representing)
- isa(tree, computer data structure)
- isa(tree, connected graph)

Table 2 contains some of the extracted relations from disambiguation pages. The extracted relations are between Persian terms but to be understandable for all readers we mention the English translation of extracted relations throughout this paper.

Table 2. Translation of some extracted relations from disambiguation pages

Isa (milk, dairy product)
Isa (lion, Felidae)
Isa (valve, device)
Isa (electric charge, concept)
Isa (Municipality, administrative division)
Isa (watch , device)

3.2 Similarity Based Approach

The similarity based part of our system uses Persian or English synonyms of a given word to find its related words and it is based on the similarities among the synonyms' contexts. The input of this part is a Persian word and as output, it returns a set of candidate related words to the given Persian word. The relation extraction resource could be WordNet, Persian corpus, Wikipedia or other resources, according to the application. The system finds the parts of the resource, which are related to each synonym, the intersection of which leads us to new relations. The process of finding the related parts of resource and finding the intersection is defined regarding the type of the resource and the task in hand. In the rest of this section we present three similarity based methods, using three different resources to extract taxonomic or non-taxonomic relations. In these methods (apart from the one using WordNet) the type of extracted relations are not defined and the system just extracts related words. The type (label) of these relations can be found by using pattern based method (see section 3.3) or by doing linguistic analysis. In this section we will show the application of this approach on different learning resources.

3.2.1 WordNet

In this part system uses WordNet relations as a guide for extracting Persian relations. The whole idea is to find WordNet synsets with a meaning close to the Persian word's, and then to translate the relations of these synsets to Persian. The problem is how to find these WordNet synsets which is solved by similarity based method. Although in this section we talk about extracting taxonomic relations, this method can be well adapted for any other WordNet relations. To extract relations having WordNet as a resource, the system follows the following steps:

1. Find the English equivalents (translations) of the Persian word using a bilingual dictionary.
2. Find the related WordNet synsets for each English equivalent.
3. Select the synset(s) with greatest number of words in common with all English equivalents.
4. Find the hypernym synsets of the selected synset(s) in step 3.
5. Translate the words in hypernym synset(s) to Persian.
6. Make new relations between the given Persian word and each translation from step 5.

It is worth mentioning that zero or more translations may be found for each English word or phrase in step 5. In this method the system uses a Persian to English dictionary which gives English equivalents for Persian synsets. The structure of the dictionary is as follows:

Persian Word	Sense Number	Persian Synonyms	English Synonyms

As we didn't have any English to Persian dictionary with the same structure, we decided to use this dictionary reversely to translate English words to Persian ones. So since the English words are not annotated with sense numbers, the reverse dictionary's structure is as follows:

English Word	Persian Word	Persian Sense Number	Persian synonyms

The fact that we can't distinguish different senses of an English word causes some problems in translating hypernym synset(s) to Persian, the system doesn't know which translations are related to the word's sense indicated in WordNet synset. So we should have a mechanism to find the correct translation of the English words. To solve this problem and to increase the precision we used a voting strategy. In this strategy all the words in hypernym synset are searched in three resources to find their translations: bilingual dictionary, Wikipedia and Wiktionary 16 which is a wiki based dictionary. Then within the all retrieved Persian equivalents for all the synset words, we choose the equivalent(s) with greatest frequency. The translation of English words can be directly found by bilingual dictionary and by Wiktionary, but to find the translation of words via Wikipedia the word is searched in English Wikipedia and we see if the retrieved article has link to the related Persian article. If such a link exists, the title of the Persian Wikipedia 15 article is the translation of the original English word.

To more clarify this method we follow what system dose for the first sense of the Persian word "آموزش" "Amuzesh". According to the bilingual dictionary the English equivalents of this word are: pedagogy, pedagogics, learning, instruction, tuition, study, teaching, schooling, educating and education. The related synsets of these words are shown in table 3.

Table 3. English synsets for english synonyms of the word "آموزش"

English word	Related synsets	Common number
pedagogy	(teaching method, pedagogics, pedagogy) ‖ (teaching, instruction, pedagogy) ‖ (education, instruction, teaching, pedagogy, didactics, educational activity)	2, 3, 4
pedagogics	(teaching method, pedagogics, pedagogy)	2
learnign	(learning, acquisition) ‖ (eruditeness, erudition, learnedness, learning, scholarship, encyclopedism, encyclopaedism)	1, 2
Instruction	(direction, instruction) ‖ (education, instruction, teaching, pedagogy, didactics, educational activity) ‖ (teaching, instruction, pedagogy) ‖ (instruction, command, statement, program line)	1, 4, 3, 1
Tuition	(tuition, tuition fee) ‖ (tutelage, tuition, tutorship)	1, 1
Study	(survey, study) ‖ (study, work) ‖ (report, study, written report) ‖ (study) ‖ (study) ‖ (discipline, subject, subject area, subject field, field, field of study, study, bailiwick, branch of knowledge) ‖ (sketch, study) ‖ (cogitation, study) ‖ (study)	1, 1, 1, 1, 1, 1, 1, 1, 1
Teaching	(teaching, instruction, pedagogy) ‖ (teaching, precept, commandment) ‖ (education, instruction, teaching, pedagogy, didactics, educational activity)	3, 1, 4
Schooling	(schooling) ‖ (school, schooling) ‖ (schooling)	1, 1, 1
Educating	No relevant synsets found	
education	(education, instruction, teaching, pedagogy, didactics, educational activity) ‖ (education) ‖ (education) ‖ (education) ‖ (ducation, training, breeding) ‖ (Department of Education, Education Department, Education)	4, 1, 1, 1, 1, 1

The system should select the synset(s) which has more words in common with all the English equivalents i.e. we find the intersection of each synset and the set of English equivalents and we choose the synset with largest intersection. As result we reach the most similar English synset to our Persian word. The number of words in the intersection of each sysnet (which is a sign for similarity and is called as "common number") in our example are indicated respectively in table 3 and we can see that the synset (education, instruction, teaching, pedagogy, didactics, educational activity) hast the larget intersection with 4 common words, so this synset is selected as our target synset in wordNet.

When the target English synset is found we start mapping its relations to Persian. As we mentioned before for now we choose Hypernymy relation to find taxonomic relations for the given Persian word. So we find the hypernym synset(s) of the selected synset which is: (activity). Now we translate all the words in Hypernym synset to Persian, and as result we have some hypernymy relations between the Persian word "آموزش" "Amuzesh" and the translated words. In our example the relation isa(آموزش, فعاليت) (which means isa(instruction, activity)) is extracted.

In some cases no synsets are selected in step 3 and that occurs when all the synsets only cover one word among the English equivalents. In these cases system follows another strategy. In this alternative strategy step 1 and 2 are the same as the previous one and the strategy continues with the following steps:

3. Find the hypernym(s) of all of the synstes retrieved in step2.
4. Select the hypernym synset(s) with most frequency among the hypernyms found in step 3.
5. Translate the words in selected synset(s) of step 4.
6. Make new relations between the given Persian word and each translation from step 5.

3.2.2 Wikipedia

Another resource used in extracting conceptual relations via similarity based method is Wikipedia. In the whole Wikipedia articles each important word which has an article in Wikipedia itself, is linked to its related article. These linked words especially the ones locating in the first section of the text are usually related to the title of the document. We use this fact to extract some taxonomic and non-taxonomic relations. This method can be also categorized under structure based approach. The following steps are followed by system to extract relations from Wikipedia by this method:

1. Find the Persian synonyms of the given word.
2. Find the related Wikipedia articles to the given word and all its synonyms.
3. Extract hyperlinked words of each Wikipedia article of step 3.
4. Find the common hyperlinked words in all extracted Wikipedia documents.
5. Make new relations between the given word and the words found in step 4.

The translations of some of the extracted relations by this method are shown in table 4.

Table 4. Translations of some relations extracted from wikipedia using similarity based approach

Input Word	Related word
Instruction	School
Child	Human
Life	Death
Life	Birth
Calculus	Math
Child	Son

3.2.3 Corpus

The last resource used in the similarity based part is corpus. In this system a general domain corpus named Peykareh 10 is used which is a collection gathered form Ettela'at and Hamshahri newspapers of the years 1999 and 2000, dissertations, books, magazines and weblogs. This method which can be classified as a statistical method contains the following steps:

1. Finding the Persian synonyms of the given word.
2. Finding the words which co-occur with the given word and its synonyms (separately) and their co-occurrence frequencies.
3. Selecting the words among the results of step2 with a frequency above a given threshold.
4. Finding the words of step 3 which co-occur with the given word and all its synonyms.
5. Making new relations between the words extracted in step 4 and the given word.

The threshold used in third step is to increase the precision of the extracted relations. The test results showed us that 8% of the total frequency of the word (the input word or any of its synonyms) is a proper threshold. Some of the relations extracted by this method are presented in table 5.

Table 5. Translations of some relations extracted from corpus using similarity based approach

Input Word	Related Word
Football	Team
Why	Reason
Scene	Movie
Politics	Government
Man	Woman
Actor	Movie
Success	Failure
Increase	Decrease
Death	Life

3.3 Pattern Based Approach

In this section we describe the pattern based part of our relation learning system. This system exploits pattern based approach to extract both taxonomic and non-taxonomic relations from Persian texts. To extract taxonomic relations we define a set of 36 patterns containing the adaptation of Hearst patterns for Persian and some other new patterns. We have also extracted some patterns for some well known non-taxonomic relations such as "Part of", "Has part", "Member of" and "synonymy". The translations of some of these patterns are shown in table 6 (TW stands for target word).

In this system pattern matching method is used in two modes. In the first mode the system is given a pair of related words and the target is to find the type of relation between them. These related pairs are obtained by using structured based or similarity based methods as it was described in section 3.2. In this case the two words are substituted in each pattern and are searched in corpus or web to find the occurrences of the patterns i.e. TW and X in above patterns are given. By following this method the system would be able to detect the type of relation if it is a taxonomic relation or a non-taxonomic one for which we have a template.

Table 6. Translation of some patterns for extracting relations

Pattern	Relation	Pattern	Relation
TW is X.	Hypernymy	TW is a part of X	Part of
TW is a X	Hypernymy	TW includes X	Has part
TW is considered as X	Hypernymy	TW means X	Definition
TW is known as X	Hypernymy	TW is defined as X	Definition
TW is called X	Hypernymy	TW1 or TW2 or … are	Synonymy
TW is named as X	Hypernymy	TW has X	Has

In second mode system is given just one word for which it should find some relations. So in this case the X part of the patterns is known and we should find TW by matching patterns over text. As we can hardly find the occurrences of these patterns through the corpus and since large corpora are not available for Persian we decided to use Wikipedia in the process of relation extraction. In the first section of Wikipedia articles we can usually find some occurrences of our patterns. To start the pattern matching phase we extracted the 1000 most frequent Persian nouns and extracted the Wikipedia articles related to these words. For each word the related article is searched for the phrases matching any of the patterns. The translations of some of the extracted relations by this method are mentioned in table 7.

Table 7. Translation of some extracted relations

Isa (Iran, country)
Isa (newspaper, publication)
Isa (water, liquid)
Isa (man, human)
Isa (pen, tool)
Has part (personality, specificity)
Has part (Tehran, Tajrish)
Has (Greece, history)
Has (Iraq, source)
Synonym (thought, idea)

It should be mentioned that these patterns are used for simple phrases and while encountering complex phrases having different syntactic groups, the precision of the method decreases. To avoid this problem some text processing tools (e.g. chunker) are needed to find the constituents of sentences. As there is no efficient chunker for Persian, we did some post-processings 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.

4 Experimental Results

The proposed methods were tested separately and the results are presented in this section. As it was mentioned in section 3.3 the second mode of pattern based approach is tested over 1000 most frequent nouns of Persian. The extracted relations were mainly taxonomic and the number of non-taxonomic relations was much less. Since the given words were not domain specific and no reference ontology was available, human evaluation was used to evaluate our system. The results of pattern based section were evaluated by an ontology expert and a precision of 76% was obtained. Most of the error rate in this part is related to the lack of efficient linguistic tools like chunker for Persian. Although some refinement strategies were applied as was described in section 3.3, but still more process is needed to refine the extracted relation. Despite the unavailability of linguistic tools our system has a precision comparable with English pattern matching systems like 2 and 4.

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 problems mentioned above are also present in structure based methods but a great portions of them in handled by following the same rules described in 3.3. For this method and also for pattern based approach we had no efficient way to count the whole number of relations in the input resource to calculate the recall of our system.

Similarity based approach was tested regarding the used resource. Again we used human evaluation and according to our ontology expert the similarity based approach using WordNet has a precision of 73% and a recall of 54%. The similarity based method which uses Wikipedia has a precision of 76% according to human evaluation. This method has high precision but the number of extracted relations was low like in other proposed methods using Wikipedia and this fact has two major reasons: the first is that the Persian Wikipedia is sparse and there are many words for which no Wikipedia articles are created. Only about 35% of the selected words had correspondent articles in Wikipedia. The second reason is that the most frequent Persian common nouns which were used in test are mostly abstract nouns and Persian Wikipedia articles are usually about proper nouns or concrete nouns.

The final test was performed on the output relations of similarity based approach which uses corpus as input. This method was tested on 300 most frequent common nouns of Persian and the results were verified by three ontology experts. As it was mentioned in section 3.2.3 to increase the precision of extracted relations we considered a threshold and we accepted the co-occurrence relations with a frequency above the defined threshold. To find the proper threshold we performed an initial test that showed us this threshold should be between 5% and 10%. Then we tested our system with different threshold between these boundaries and asked our experts to mark the extracted relations as "Acceptable" or "Unacceptable" to appear in a domain independent ontology. As we had no means to calculate the recall we compared different results regarding their precision and number of correctly extracted relations.

The test results are shown in figure 1 in which horizontal axis shows precision and vertical axis indicates the number of correctly extracted relations.

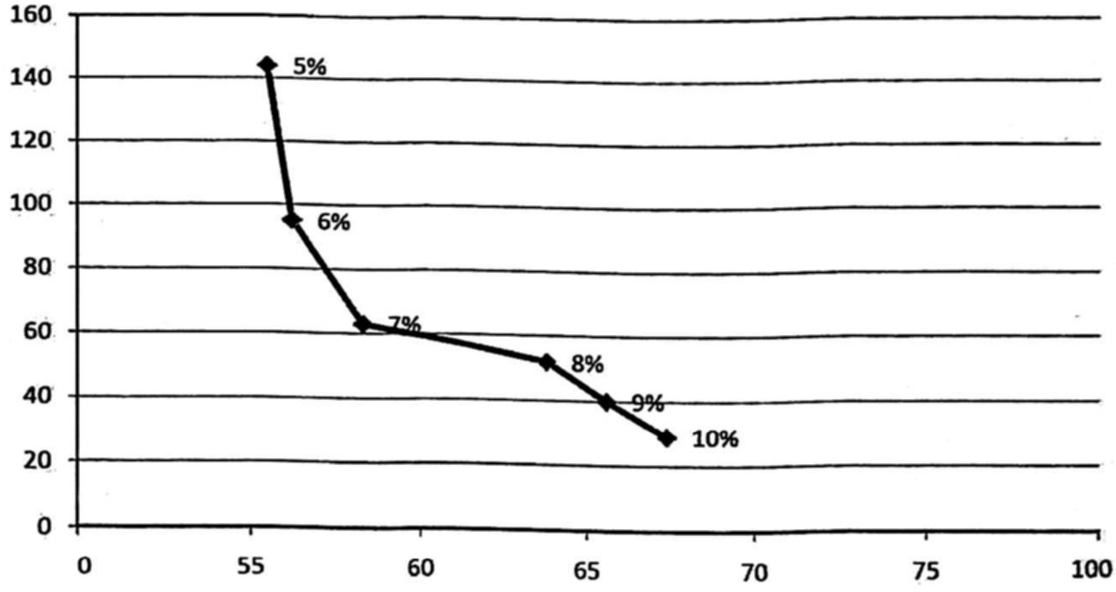

Fig. 1. Test results for different thresholds

As it can be seen in figure 1 the precision increases by raising the threshold while the number of correct relation decreases. To have a reasonable precision and not to miss many correct relations we decided to set 8% as our threshold.

5 Conclusion and Further Work

In this paper a hybrid approach was presented for extracting conceptual relations from Persian resources especially from web. This approach is a combination of pattern based, structure based, similarity based and statistical relations, enriched with linguistic heuristics. This system which is one of the few works done on ontology learning for Persian, uses different methods and resources to use their advantages and to cover the disadvantages of each method with others. The proposed approach is able to extract a noticeable number of relations and is used in the process of building Persian WordNet.

To increase the precision of extracted relations more linguistic heuristics could be applied. Extracting more patterns for taxonomic relations and covering more non-taxonomic relations could be considered as future work. Some more complex methods could be used to find the constituents of Persian sentences. In this way we can extract more relations, especially the non-taxonomic ones, from text and also the precision of pattern based method will increase. More advanced ways to find the types of relations via searching the web and linguistic analysis could be found.

References

1. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: 14th International Conference on Computational Linguistics, pp. 539--545, (1992).
2. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidenc. Ontology learning from text: Methods, Evaluation and Applications. IOS press (2005).
3. Shamsfard, M., Barforoush, A.: Learning Ontologies from Natural Language Texts. International journal of human- computer studies. 60, pp. 17--63 (2004).
4. D. Sanchez, A. Moreno, Discovering non-taxonomic relations from the Web. In: LNCS, vol. 4224, pp. 629--636. Springer, Heidelberg (2006).
5. Ruiz-Casado, M., Alfonseca, E., Okumura M., Castells, P.: Information Extraction and Semantic Annotation of Wikipedia. Ontology learning and Population: Bridging the Gap Between Text and Knowledge. IOS press (2008).
6. Reinberger, M., Spyns, P.: Unsupervised Text Mining for the Learning of DOGMA-Inspired Ontologies. Ontology learning from text: Methods, Evaluation and Applications. IOS press (2005).
7. Ryu, P., Choi, K.: An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning. Ontology learning from text: Methods, Evaluation and Applications. IOS press (2005).
8. Suchanek, F. M., Ifrim, G., Weikum, G.: Combining linguistic and statistical analysis to extract relations from web documents. In: 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, (2006).
9. Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. Ontology learning and Population: Bridging the Gap Between Text and Knowledge. IOS press (2008).
10. M. Bijankhan: Role of language corpora in writing grammar: introducing a computer software, Iranian Journal of Linguistics, No. 38 : 38-67, (2004).
11. Sanchez, D., Moreno, A.: Automatic Generation of Taxonomies from the WWW. In: 5th International Conference on Practical Aspects of Knowledge Management. Vol. 3336 of LNAI., pp. 208--219, (2004).
12. Cimiano, P. and Staab, S.: Learning by googling, ACM SIGKDD Explorations Newsletter, vol.6 n.2, p.24--33, (2004).
13. Herbelot, A., Copestake, A.: Acquiring Ontological Relationships from Wikipedia Using RMRS. In: Web Content Mining with Human Language Technologies workshop, USA, (2006).
14. Hazman, M., El-Beltagy, S.R. and Rafea, A.: Ontology learning from textual web documents. In: 6th International Conference on Informatics and Systems, NLP, pp.113--120, Giza, Egypt, (2008).
15. Persian Wikipedia, the free encyclopedia, http://fa.wikipedia.org
16. Persian Wiktionary, the free dictionary, http://fa.wiktionary.org