Int. J. Nonlinear Anal. Appl. (10) No. 2, 189-196
ISSN: 2008-6822 (electronic)
https://dx.doi.org/10.22075/IJNAA.2019.4185

A new representation of open information extraction in Persian language

Mohammad Mahdi Nematollahi[a], Omid Reza Marouzi[b,*]

[a] *Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran*
[b] *Faculty of Electrical and Robotic Engineering, Shahrood University of Technology, Shahrood, Iran*

(Communicated by J. Vahidi)

Abstract

Open information extraction is a new technology in the text mining process which is still at the beginning and requires many attempts and considerations for improvement. These attempts includes both the representation and extraction of information. The complication and instability of language intensify the problems of the open information extraction. In this article an advanced representation of information is presented for the Persian language; a representation which can be a favorable cover for the open information extraction by identifying dependency analysis relationships. In the present article, it is tried to reach the feasibility of this representation, the representation correspondence using syntactic labeling, and plausible representation of information extraction. By making this attempts, the threshold for the information extraction goes far beyond its simple representation state, which is a tripple. Although this article tries to overally outline the approach of using an advanced representation of information extraction, it specifically investigates the use of this representation in QA systems.

Keywords: Open Information Extraction, Representation of Information, Syntactic Labeling, Question and Answer Systems.
2010 MSC: 68T50.

1. Introduction

Information extraction is a method of obtaining simple structured knowledge out of the text [2]. The purpose of information extraction systems is to find and understand related parts of the text in a

*Corresponding author
Email addresses: nematollahi@shahroodut.ac.ir (Mohammad Mahdi Nematollahi), marouzi@shahroodut.ac.ir (Omid Reza Marouzi)

Received: June 2018 *Revised:* *July* 2019

limited way. The information obtained in these systems can be used to create a knowledge base. In fact, by the use of information extraction, information is placed in a precise form with a logical structure that allows other computer algorithms and methods for further inference and investigation.

Since the middle of the last decade, using natural language processing tools, many efforts have been made to obtain a large set of semantic relationships from the enormous amount of textual data available on the Web, without any human intervention [3].

One of the main approaches to achieve this is the *open information extraction.* In fact, the open information extraction is the extraction of knowledge out of a large amount of web information. The term “open” refers to the extraction of any information without a predetermined purpose and specific scope of texts [3].

The “information” refers to the triple formats as [entity, entity, relationship] extracted from the sentence unit. The triple formats are derived from the classical approach of conceptual models such as the class diagram or the entity-relationship model [3]. In this approach, expressions are defined as Subject-Predicate-Object. The predicate denotes the entity characteristic and expresses the relationship between the entity and the object. For instance, in the representation of the sentence ”the sky is blue”, ”the sky” is the sentence subject and ”blue” is the object, and finally ”being” relates this two components of the sentence.

The triple format is being used in extraction of a significant amount of information because it is very simple. But, in practice, this simplicity leads to many fails in the complete extraction of the information in the sentences. Therefore, other representations have been developed, which are discussed in Section 2.

The most notable and complete representation is a nested multiplicity of information that enables the most complex information to be extracted from the sentences.

In Section 3, the necessity of such representation is investigated, specifically in the Persian language. In Section 4, this advanced representation is discussed regarding the considerations required in the Persian language and the feasibility of this representation using the dependency analysis. Finally, the conclusion is drawn from the process of using dependency analysis in the representation.

2. Various representations of information extraction

Triple representation is the most known method for the representation of information extraction. One use of this representation is the extraction of textual information of info boxes in Wikipedia. These boxes are often structured based on the extraction of information from simple sentences as information triplets [13]. For example, the sentence “The capital of Iran is Tehran” gives the information [Tehran, is, the capital of Iran].

Another application of this representation is in the Resource Description Framework (RDF) which used in the Semantic Web [6], but this representation method is not capable of holding information for extraction in many cases. By this representation, the information is subconsciously hidden in one of the triplets where the extraction of information is no longer possible. For instance, the sentence ”Ali saw Hassan” can be represented as [Ali, saw, Hassan]. But if the sentence, ”Ali saw Hassan in the school next to his brother” is used, the spatial information of Hassan’s presence and the quality of his presence alongside his brother are subconsciously hidden in one of the triple entities; for example, the information may be represented as [Ali, saw, Hassan in the school next to his brother].

To address this problem, one approach is to create multiple triplets from one sentence to cover the information. To illustrate, the sentence “Ali went to school in a hurry” can be converted to two triples [Ali, went, to school] and [Ali, went, in a hurry]. In this case, the approach eliminates the

Table 1: Comparison of Representations and Linguistic Assumptions Used in Popular Information Extraction Algorithms [8]

Extractor	**output**	**Linguistics**
REVERB	binary	verb-based rels, NP args
OLLIE	binary, nested	verb/noun rels, phrasal args
SRLIE	binary, n-array, nested	verb rels, phrasal args
RELNOUN	binary	noun rels, NP args

connection between the two triplets, and in some cases it creates nonsense triplets. In addition, there are numerous examples in which this approach is not possible.

Another approach suggested for covering information in the representation of information extraction is to use constraints for each of the triplets [5]. As proof, the sentence ”Ali went to school in a hurry” says that ”in a hurry” is a condition and constraint of the relationship. So, the sentence can be kept in triple format, but the verb would be constrained to the condition of being ”in a hurry”. The representation of this example can be [Ali, went, to school |in a hurry]. In this case, constraints are separated from entities and predicates with a vertical line and they can contain more than one constraint. This approach also cannot cover information in many cases. Somtimes there are more than two entities in the sentence, and sometimes one can doubt to assign the constraint to which triplet. In sentences where the second object is used or supplementary sentences are added employing the prepositions, such representation has weakness. The sentence “Ali talked to Hassan about Mohsen yesterday at school” is an example of this case.

Another approach is to use multiple instead of triple [1]. To avoid to limit the information to triples is an important step that can cover a lot of information. According to this representation, above example can be converted into [Ali, talked to, Hassan, about Mohsen, yesterday, at school]. Although this representation is much better than previous ones, it has a disadvantage that all elements in the extraction of information are in the same level. This sometimes leads to a wrong answer in the extraction of information. For example, the sentence “ past scientists believed that the nature abhors a vacuum ” is converted into a multiple as [past scientists, believed that, the nature, abhors, a vacuum]. Since all elements are in the same level, this wrong information is obtained: “ the nature abhors a vacuum ”. To address this issue, a representation was proposed that could include nested information [4]. As an example, according to this representation, above example is converted into [past scientists, believed that, [the nature, abhors, a vacuum]], provided that the nested information is dependent on its father’s information and does not have any meaning of its own.

The more precise and complete the representation in the extraction of information, the more complete the extraction of information; on the other hand, the identification of multiple boundaries and nested information makes the extraction more difficult. If there is a way to specify multiple boundaries, then it is certainly easier to migrate from the triple to more comprehensive ways of representing information. There is not much tendency to such approaches in the open information extraction and this may be because of the complexity of identifying these boundaries in the conventional intelligent non-supervised or semi-supervised or supervised methods.
Table 1 shows how to use the representations described in the popular extraction algorithms.

There is a thing that can help the information extraction well. It is the labeling of syntactic dependency relationships based on the dependency grammar and valency theory. If there is a labeled dataset in the syntactic dependency relationships, a model can be designed to identify multiple components by the machine learning methods with acceptable accuracy.

3. Necessity of using advanced representation of information extraction in the Persian language

Languages have generally complicated behaviors and have grown in a way that they are out of control of humans and laws. The Persian language is not an exception either. In many cases a complete sentence contains information, and the incomplete ones that are the sentence complement cannot lonely be a criterion for information extraction. These sentences are most commonly used in Persian. The following are two examples of these sentences in Persian, taken from the syntactic database configuration [10]. In this configuration, it is tried to collect common Persian language sentences:

“ Often the person you are angry with doesn’t know that you are still mad of him and he has his normal life, moving toward more progress and a better life, but ‘you’ are suffering, and ‘you’ are providing a hard life for yourself and only for yourself.”

“Thousands of English demonstrators, outraged by unemployment, liquidity shortages, injustice, and economic policies of the world’s so-called top economic countries, gathered in central London on Saturday to voice their protest against the leaders of G20”.

Complete sentences in Persian sometimes consist of several incomplete sentences that are either inflected with each other, or are “main clause” and “subordinate clause”, or are addition of the previous sentence as description or deduction.

Assuming that it is not possible to extract information without the ”complete sentence” unit, the triple representation, of course, cannot handle the information extraction. So a representation is required that can represent all the information and its interdependency in the nested form covering n-components of the language. The nested form of this representation is required because of covering the sentences and the constituents of a complete sentence; and for the n-representations, n is because of the covering of several components of the word roles used in a sentence, such as subject, verb, adverbs, substitutes, clauses, and objects.

4. Advanced representation of information extraction

The first step in a representation is to parse a complete sentence using a shallow parsing, which converts the sentence into parts with syntactic labels. For instance, by using the shallow parsing, above sentences are converted into the following sentences:

(ADVP Often) (NP the person) (NP you) (VP are) (ADJP angry) (PP with) (VP doesn’t know) (PP that) (NP you) (VP are) (ADVP still) (ADJP mad) (PP of) (NP him) (PP and) (NP he) (VP has) (NP his normal life), (VP moving toward) (NP more progress and a better life), but (NP ‘you’) (VP are suffering), (PP and) (NP ‘you’) (VP are providing) (NP a hard life) (PP for) (NP yourself) (PP and) (ADVP only) (PP for) (NP yourself).

(NP Thousands of English demonstrators), (VP outraged) (PP by) (NP unemployment), (NP liquidity shortages), (NP injustice), (PP and) (NP economic policies of the world’s so-called top economic countries), (VP gathered) (PP in) (NP central London) (PP on) (ADVP Saturday) (PP to) (VP voice) (NP their protest) (PP against) (NP the leaders of G20).

The symbols used in the above parsing are defined in Table 2. This process identifies the n elements in the representation, but the relevance and dependency between the sentences and their incomplete components must be extracted and identified in the next stage.

Some of the most important communications in the Persian language whose identification is necessary for the open information extraction are as follows:

- Conjunctive sentences:
  These sentences are linked together using the letters like ”and” and ”or”. So the sentence ”Ali

Table 2: symbols used in shallow parsing

Abbreviation	**Description**
ADVP	adverb group
VP	verb group
NP	noun group
POSTP	post preposition group
PP	preposition group

went to school and then returned home” consists of two components: ”Ali went to school” and ”then returned home” that are linked by the letter ”and”. Therefore, the representation of these components extraction, is nested as follows: [H (NP Ali) (VP went) (PP to) (NP school)] [CH and] [H (ADVP then) (VP returned) (NP home)]

- Clause of Nouns:
  These clauses are used to explain a word or a phrase in a sentence. For example, in the phrase ”Ali, who is exemplar of observing moral principles, must also be exemplary in effort.” the term ”is exemplar of observing moral principles” is a clause that explains ”Ali” using the word “who”. [RB (NP Ali)] [CB who] [B (VP is) (NP exemplar) (PP of) (NP observing moral principles)]

- Verbal clauses:
  Sometimes a clause describes a verb, rather than a word or a phrase. For instance, in the sentence ”This theory is provided so that it is the basis for solving future problems”, the phrase ”It is the basis for solving future problems” is a clause that explains the previous sentence, i.e. ”This theory is provided because”. In this sentence, the word ”that” links the clauses.

- Appositions:
  Appositions provide an explanation for a part of a simple sentence, in a nested way. As an example, in the sentence ”Hafez, the famous poet of Shiraz, has composed beautiful poems regarding mystic”, the term ”the famous poet of Shiraz” is an apposition used to describe the term ”Hafez”. [RA (NA Hafez)] [A (NP the famous poet of Shiraz)]

- Additional clauses:
  Additional clauses have the form of main and subordinate clauses in Persian. As proof, in the sentence ”If you withhold water and sun from it, it will wither away”, the phrase ”will wither away” is an additional clause for ”If you withhold water and sun from it”. [CB If] [B (NP you) (VP withhold) (NP water) (PP and) (NP sun) (PP from) (NP it)] [B (NP it) (VP will wither away)]

The symbols applied in the advanced representation of information are given in Table 3. Tree representation can be used for this representation. For example, the sentence ”Ali, who is a good boy and is diligent enough, is a source of pride for his family,” is indicated in Figures 1 and 2, before and after the nested parsing:

5. Method of extraction of advanced representation using dependency analysis

Dependency grammar theory is a constructivist and formalist theory, in which syntactic constructs in various languages are investigated mainly by examining the interdependent relationships between core and dependent elements of the language [11]. Tesnière theories are the first ones used this

Table 3: symbols used in advanced representation of information

Abbreviation	Description
H	Conjunction entity
B	Clause of Noun
A	Appositions
CH	Predicate entity
MB	Main clause
RB	Reference of clause
CB	Clause predicate entity

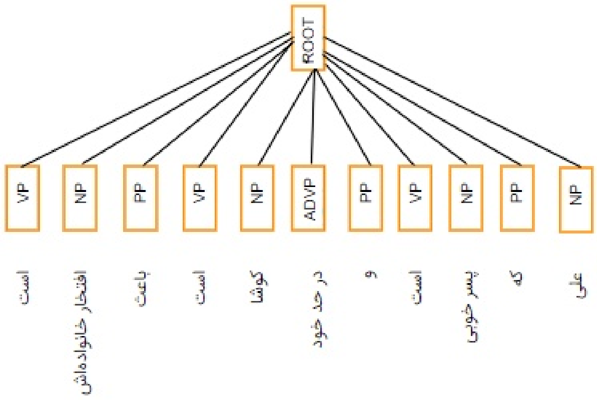

Figure 1: tree representation of information before nested parsing
(*Translation*: Ali, who is a good boy and is diligent enough, is a source of pride for his family)

approach in modern linguistics. He first brought up this view in a compact book named “structural syntax”, which was published in “Fundamentals of structural syntax” [12] after his death.

After Tesnière, various linguists have proposed different methods for the dependency grammar theory. All of these grammars are based on this basic assumption that the structural syntax includes the words which are related to each other with assymetric binary relationships. These are called dependency relationships or dependency [7]. The dependency grammar theory is based on two main assumptions: firstly, each sentence contains a central verb. The second one is that the base structure of sentences containing verbs, can be determined by the type and number of compulsory and arbitrary additions [11]. Figure 3 shows an example of a dependency tree.

The advanced representation can be created by using the labels of the syntactic parsing and by the following rules:

- If the role of a clause in a sentence is “Clause of Noun”, and if the role is “Subordinating Conjunction, Coordinating Conjunction and Pseudo-Sentence” from the syntactic viewpoint, the nested threshold will be after the word until the next verb. Or it is a “verb” syntactically, which will be the nested threshold until the verb.

- If there is an apposition in a clause, the word related to the apposition will be the nested entity of the apposition.

- If a clause role is specified as the “verb added clause” the nested entity threshold will be in the range after the clause until the next verb.

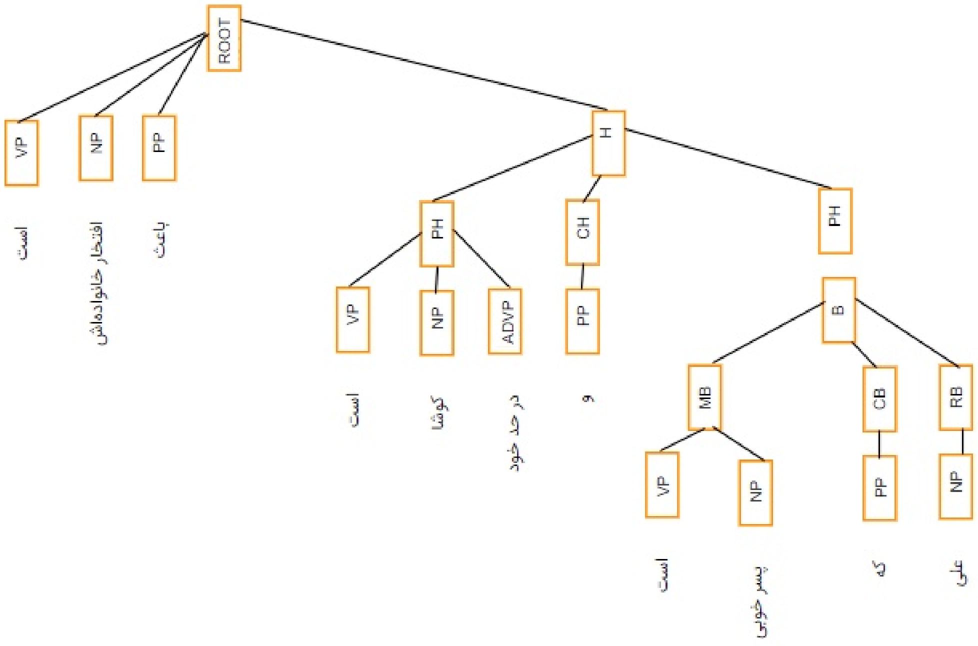

Figure 2: tree representation of information after nested parsing
(*Translation*: Ali, who is a good boy and is diligent enough, is a source of pride for his family)

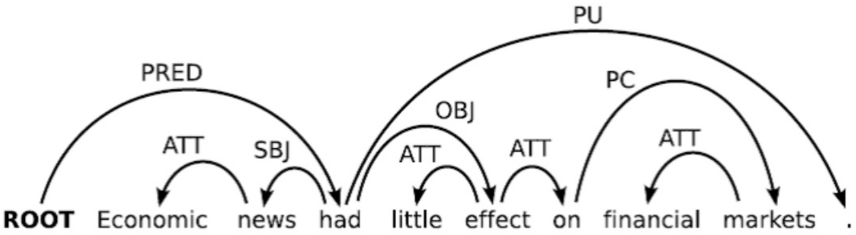

Figure 3: an example of a dependency tree in the English language [9]

- If there is a clause with a role of “Conjunction of Verb”, there will be two nested entity thresholds: 1- After the Conjunction word (like “and” and “or”) until the next verb. 2- Befor the Conjunction word until the beginning of a sentence.

6. Conclusion

One of the advantages of the advanced representation of information extraction based on the syntactic labeling is the hidden power of the labels in managing the information [**?**]. With this approach, most of the questions asked from the sentence will be addressed. So, not only the information of any sentence can be extracted, but also acceptable questions will be addressed. To specify the exact keywords of the questions, syntactic labeling is also required which is not the focus of present article.

Considering above labeling, the information can be extracted so precisely that most of typical methods of information extraction are not capable of it.

References

[1] Alan Akbik and Alexander Löser. Kraken: N-ary facts in open information extraction. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 52–56, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[2] Douglas E Appelt. Introduction to information extraction. *Ai Communications*, 12(3):161–172, 1999.

[3] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *Ijcai*, volume 7, pages 2670–2676, 2007.

[4] Nikita Bhutani, H Jagadish, and Dragomir Radev. Nested propositions in open information extraction. pages 55–64, 01 2016.

[5] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[6] Graham Klyne and Jeremy J. Carroll. Resource description framework (rdf): Concepts and abstract syntax. W3C Recommendation, 2004.

[7] Sandra Kubler, Ryan McDonald, Joakim Nivre, and Graeme Hirst. *Dependency Parsing.* Morgan and Claypool Publishers, 2009.

[8] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 523–534, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[9] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[10] Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 306–314, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[11] Omid Tayyebzade. *The capacity of verbs and sentence basic structures in Persian today.* Center Publication, 1393. (In Persian).

[12] L. Tesnière. *Esquisse d'une syntaxe structurale.* C. Klincksieck, 1953.

[13] Wikipedia. Infobox — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Infobox&oldid=924877551, 2019. [Online; accessed 24-November-2019].