A Semantic Approach to Person Profile Extraction from Farsi Documents

Hojjat Emami*
Department of Information and Communication Technology (ICT), Malek-Ashtar University of Technology, Tehran, Iran
h_emami@mut.ac.ir
Hossein Shirazi
Department of Information and Communication Technology (ICT), Malek-Ashtar University of Technology, Tehran, Iran
shirazi@mut.ac.ir
Ahmad Abdollahzadeh Barforoush
Computer Engineering Department, Amir Kabir University of Technology, Tehran, Iran,
ahmadaku@aut.ac.ir

Received: 14/Jun/2016 Revised: 29/Aug/2016 Accepted: 29/Aug/2016

Abstract

Entity profiling (EP) as an important task of Web mining and information extraction (IE) is the process of extracting entities in question and their related information from given text resources. From computational viewpoint, the Farsi language is one of the less-studied and less-resourced languages, and suffers from the lack of high quality language processing tools. This problem emphasizes the necessity of developing Farsi text processing systems. As an element of EP research, we present a semantic approach to extract profile of person entities from Farsi Web documents. Our approach includes three major components: (*i*) pre-processing, (*ii*) semantic analysis and (*iii*) attribute extraction. First, our system takes as input the raw text, and annotates the text using existing pre-processing tools. In semantic analysis stage, we analyze the pre-processed text syntactically and semantically and enrich the local processed information with semantic information obtained from a distant knowledge base. We then use a semantic rule-based approach to extract the related information of the persons in question. We show the effectiveness of our approach by testing it on a small Farsi corpus. The experimental results are encouraging and show that the proposed method outperforms baseline methods.

Keywords: Web Mining; Information Extraction; Person Profiling; Farsi Language.

1. Introduction

Entity profiling (EP) is an active research topic in Web data mining and information extraction (IE). EP aims to gather, infer, refine and group unobservable information of a given entity from observable data about it. There are many ongoing researches on the problem of EP in different languages. However, one of the less studied languages in EP is Farsi. Farsi speaking people constitute 1.5% of the world’s population. They spend hours daily in the Internet and easily publish data on their homepages, news articles, blog entries, item reviews, comments, micro-posts, and social networks. This results a huge volume of valuable Farsi contents on the Internet, which a significant part of them are unstructured, free-text documents. Farsi contents constituent 1% of all the digital contents on the Internet. The volume of Farsi content has increased at a steady rate over the past years (blue line in Figure 1), and this growth is expected to be continuing for the future (orange line in Figure 1). Due to the special and different nature of Farsi such as linguistic phenomena, lack of appropriate natural language processing (NLP) tools and underlying linguistic resources, processing Farsi content is more serious [1],[2]. These challenges highlight the necessity of developing high quality IE approaches in Farsi. In this article, we present an approach to extract profile of persons in Farsi, and report an evaluation of that. Person profiling is a specific variant of the general EP problem. In person profiling, we are given a collection of Web pages about different persons. The process is to extract profile of each given person from his/her relevant Web pages. A central task in person profiling is *attribute extraction*. In recent years, a few efforts have been made to automatically process the Farsi text and to extract attributes of entities. However, these approaches suffer from several fundamental issues.

The first is that many existing work (e.g., [3]-[5]) relying on syntactic information and used pre-specified lexico-syntactic rules or specific machine learning approaches. These methods cannot entirely solve the problems of *synonymy* and *polysemy* that need deep understanding of text. The second problem is that the resulting attributes are surface textual facts and are not linked to an ontology.

* Corresponding Author

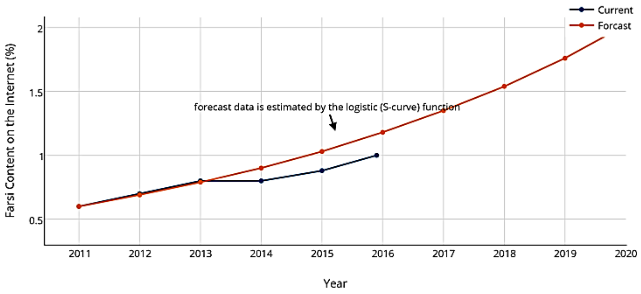

Fig. 1. Farsi content's growth on the Internet; the blue curve shows the growth rate of Farsi content from 2011 to 2016[1], and orange line shows the forecast growth rate up to 2020.

The third is the problem of syntactic variation, presenting the same meaning with different surface linguistic expression forms. This makes it hard for any Web AE to cover all variations of writing patterns. These observations promoted us to developing a new semantic AE approach to overcome the challenges of syntactic variation, and extracting semantic, meaningful attributes.

We use two types of AE methods to extracting profile information of an entity: *verb-based AE* and *noun-based* AE. We use verb-based AE to extract attribute values from semantic role (SRL) frames which their verb predicates serves as an indicator of a given attribute class. We show how verb-based AE can improve the quality of IE, partially solve the syntactic variation. To extract attribute values from noun-based constructions, we use noun-based AE. For noun-based attributes, we map each sentence into a semantic boosted dependency graph, and then use dependency-based patterns to extract the target attributes. The observation underlying our approach is that understanding the text semantically can improve the results. Our approach addresses the problems of *synonymy* and *polysemy*, and makes full use of the merits of both syntactic and semantic analysis of the text. Our approach links the resulting textual surface facts contained in profiles to their possible meaning in a distant ontology. This is very helpful for multi-lingual text processing. The resulting profiles are structured and machine-readable and can simply translate to other more-studied and more-resourced languages, and facilitate understanding and processing of Farsi text documents.

To summarize, our contributions in this article are as follows:

- For verb-based AE, we present a semantic approach, which is effective for rich profile IE from SRL frames. This is helpful to solve the problem of syntactic variation in expressing the same meaning.
- For noun-based AE, we improve the robustness of IMPLIE system [6] by incorporating co-reference information. We apply our extended algorithm on semantic enriched dependency graph to extract both single-value and multiple-value attributes.
- To evaluate the performance of our method, we create a small Farsi dataset drawn from Wikipedia articles. We compare our approach with baseline methods. Our experiments demonstrate that our method is an encouraging approach, and it can extract high quality information about entities.

The remainder of this paper is organized as follows. After a brief survey of related work in Section 2, we describe our personal EP approach in Section 3. Section 4 describes the experiments we performed to evaluate our approach. Finally, we draw some conclusions, and identify future work in Section 5.

2. Related Work

The task of extracting structured and relevant information about entities from text documents is a long-standing task in the IE and NLP community. Early IE systems were based on lexico-syntactic rule-based methods and are domain dependent. For example, Li et al. [7] used a multi-level rule-based approach, which relies on various linguistic IE patterns to extract entity-centric relations from English corpora. However, such approaches to IE are limited by the availability of domain knowledge, the difficulty in designing rules for all types of text, and less accurate results under noisy setting. Later systems to achieve robustness under noisy setting and to extract arbitrary entity-centric relations use probabilistic [8],[9] and statistical methods [10],[11]. However, these approaches do not discover the semantic information contained in text entirely. Recently, machine learning methods are widely used for entity-centric relation extraction. Supervised learning methods achieved high performance in relation extraction, but they need more hand labelled training data in order to be effective [12]. Due to the lacking of high quality of labelled training data, and the low performance of supervised methods for extracting arbitrary relations from large-scale corpora such as Web, semi-supervised learning methods [12],[13] bootstrapping methods [14], self-supervision approaches [15], distant supervision methods [16], and unsupervised clustering methods [17] are developed. However, each of these methods suffers from several challenges. For example, bootstrapping methods suffers from semantic drift problem, and distant supervision suffers from noisy training data. The output of unsupervised learning methods often does not resemble ontological relations and the resulting relations are hard to map to a domain ontology.

Open IE [18] as a new emerged IE methodology aims to extract arbitrary domain-independent relations in the text without a pre-determined set of relations and with no domain-specific knowledge engineering effort. Open IE extractions are surface text and do not resemble domain-specific ontological relations [19]. Recently, a few approaches focused on adapting Open IE extractions to domain-specific ontology [6],[19]-[21]. Soderland et al. [19] propose a two-step approach for adapting open domain relational tuples to domain-specific ontological relations. In the first stage, the tuples are annotated by a domain concept recognizer, and then a number of relation-mapping rules are learned by using a cover

[1] The data for draw the current status of Farsi content on the Internet are driven from "Usage of content languages for Websites", [www.W3Techs.com], Retrieved 10 March 2016

learning algorithm to map the tuples to domain relations. Since machine learning approach to learning high precision mapping rules need more training data, and such a high volume data are not available, the authors in [20] chose to create mapping rules manually rather than adopting machine learning approaches. In IMPLIE [6] syntactic dependency rules are used to find relations that are beyond the scope of Open IE extractions. IMPLIE begins with user-collected semantic taggers for a set of target attribute classes, and then uses dependency parse rules to find noun phrase that are modified by terms of a target class. We borrow the idea from their work for extracting noun-based attributes, but we enrich the syntactic information with semantic information obtained from a distant ontology to alleviate the errors produced by syntactic parsing. Exploiting syntactic dependencies for relation extraction is not a new idea and studied in early work. For example, [22] formulates the entity-centric relation extraction as the problem of finding the shortest paths between entities on dependency graph. Some other information extraction works rely on shallow semantic analysis of text [23]-[25]. For example, Surdeanu et al. [23] proposed a rule-based approach, which contains a number of mapping rules to map SRL frames to relations in question. However, these approaches have not been addressed entirely some challenging linguistic phenomena such as *synonymy* and *polysemy*.

Some other works integrate syntactic dependencies and semantic information derived from distant knowledge bases to address the challenges like *synonymy* and *polysemy*. For example, Moro and Navigli [26] combined syntactic dependencies and distributional semantic information to extract ontologized relations. However, the resulting relations are still bound to surface text, lacking actual semantic content. Bovi et al. [27] developed DEFIE system, which extracts semantic relations from Web text through deep syntactic and semantic analysis of the text. They obtained syntactic information from dependency parser, and semantic information from Babelfy [28]. They mainly focused on verb-based relations. We borrow the idea of enriching local document-level information with semantic information derived from distant knowledge base Babelfy from the work of Bovi et al. [27]. We (a) focus on Farsi language; and (b) integrate SRL frames with semantic information obtained from Babelfy to extract verb-based relations, and (c) in addition to verb-based relations, we focused on extracting noun-based relations by adopting and extending the rule-based approach presented in [20].

There are also some relation extraction works in Farsi. Relation extraction is a central task in entity profiling and focuses on learning atomic facts about entities. We notice that in contrast to work in relation extraction, our work addresses the problem of entity profiling, which extracts a richer information structure about a given entity. One of the first approaches to semantic relation extraction in Farsi is based on hand-crafted rules, which uses the syntactic and lexical information [3]. A similar approach is done by Moradi el al. [4]. They adopt the Hearst's approach [29] to relation extraction in Farsi. In their approach, relations are extracted by matching some pre-defined patterns over the text. Their approach has syntactic nature and does not analyze the text semantically. In other work, [5] uses a set of semantic and lexico-syntactic patterns and templates for extracting taxonomic and non-taxonomic relations and axioms from Farsi text. Our own earlier work on personal information extraction in Farsi includes [30]. In the paper [30], we used syntactic-based patterns and attribute-specific gazetteers to extract personal attributes. However, the limitations to our previous work [30] and some existing work in Farsi are that (*i*) the resulting attributes are surface textual facts and are not linked to an ontology, and (*ii*) they did not address the language phenomena of *synonymy* and *polysemy* that need deep understanding of the text, and (*iii*) they suffer from the problem of syntactic variation. In this paper, we use different semantic-based approaches to improve the quality of attribute extraction, alleviate the problem of syntactic variation and the challenges of synonymy and *polysemy*, which was not studied before in previous work. Our approach relies on deep semantic analysis of the text, and enriches local entity-centric information with semantic information obtained from a distant knowledge base. The resulting profile attributes are meaningful and are linked to their possible meaning in a distant ontology. This is greatly helpful for the multi-lingual text processing such that the resulting profiles can simply translate to other language.

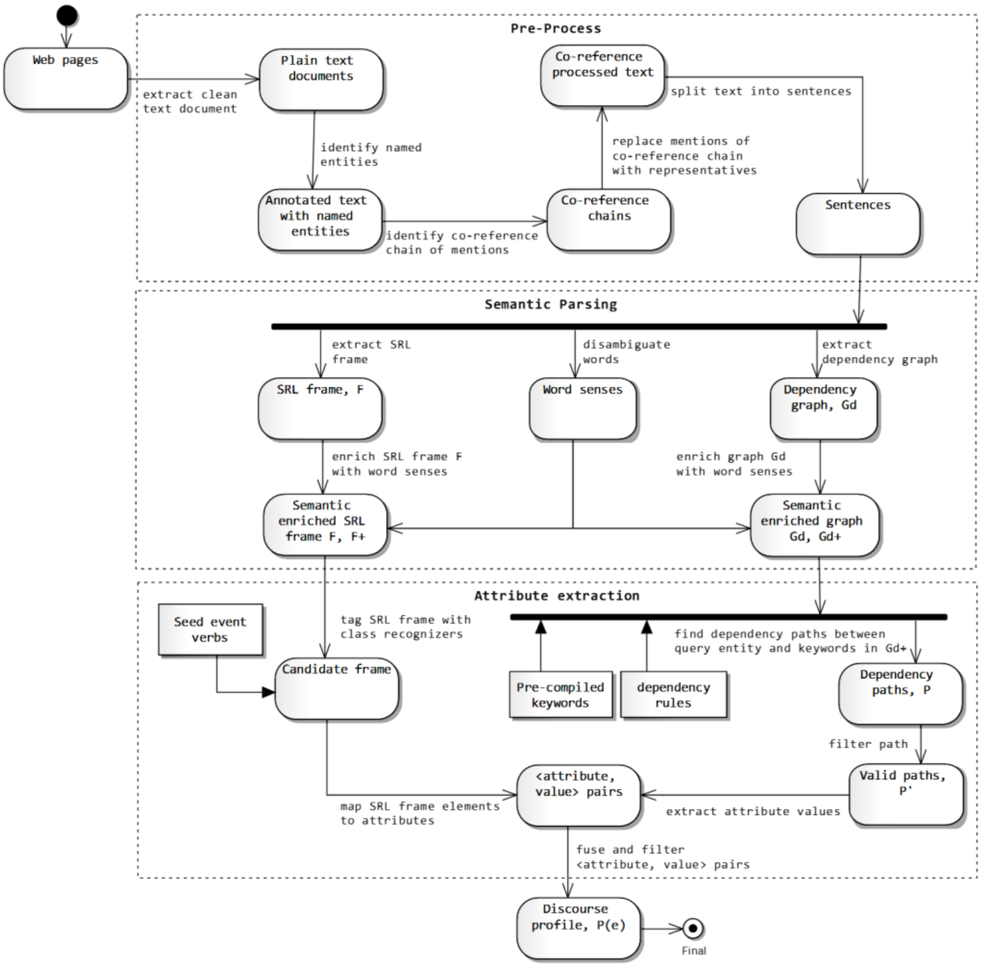
Pre-Process
Web pages
extract clean text document
Plain text documents
identify named entities
Annotated text with named entities
identify co-reference chain of mentions
Co-reference chains
replace mentions of co-reference chain with representatives
Co-reference processed text
split text into sentences
Sentences
Semantic Parsing
extract SRL frame
disambiguate words
extract dependency graph
SRL frame, F
Word senses
Dependency graph, Gd
enrich SRL frame F with word senses
enrich graph Gd with word senses
Semantic enriched SRL frame F, F+
Semantic enriched graph Gd, Gd+
Attribute extraction
Seed event verbs
tag SRL frame with class recognizers
Candidate frame
find dependency paths between query entity and keywords in Gd+
Pre-compiled keywords
dependency rules
Dependency paths, P
filter path
Valid paths, P'
extract attribute values
<attribute, value> pairs
map SRL frame elements to attributes
fuse and filter <attribute, value> pairs
Discourse profile, P(e)
Final

Fig. 2. The state diagram of our EP approach

3. Our Proposed Method

Our profile extraction system takes as input a query entity *e*, and extracts its discourse profile, which contains a number of <attribute, value> pairs, each of which represents a certain characteristic of the entity *e*. Formally, we define the discourse profile of the entity *e*, *P(e)* as follows:

$$P(e) = \{(a, v) | a \in A, v \in V(a)\} \qquad (1)$$

where a is a given attribute, *v* is a value for attribute a, A is the vocabulary of attributes that can be used to describe characteristics of the entity *e*; and $V(a)$ represent a set of filler values for attribute $a \in A$. Similarly, we define the entity profiling problem as follows:

$$Given\ \{D, A, e\}, desing\ a\ system\ \varphi,\ \ P(e) \leftarrow \varphi(D, e) \quad (2)$$

This formulation says that given a text document *D*, a query entity *e*, and a vocabulary of attributes *A*, our goal is to design a profiling system φ to extract structured information $P(e)$ related to entity e from document *D*. Figure 2 shows the state diagram of our proposed method. We decompose person profile extraction problem as three major subtasks: (*i*) pre-processing, (*ii*) semantic analysis, and (*iii*) attribute extraction. Pre-processing provides the input text as system's desired format using existing pre-processing tools. Semantic analysis takes a pre-processed sentence as input to produce its semantic representation. This component extracts the syntactic information (dependency graph) and semantic information (SRL frame) from pre-processed text, and enriches syntactic dependency graph and SRL frames with word senses and disambiguated entity mentions. The attribute extraction component, is given the query entity *e*, and a vocabulary of attributes *A*, and must find a set of filler values *V* for each attribute $a \in A$ from annotated text generated by semantic analysis component. The extracted <attribute, value> pairs are validated and integrated to form discourse profile of entities in question. In the following, we describe these tasks in more detail.

3.1 Pre-Processing

In this article, we focus on the textual part of the Web pages, because the majority of the information about entities on the Web is often expressed by natural language text. Web pages need to be pre-processed and prepared according to system's desired format. Pre-processing includes four main stages: (*i*) html tag removal, (*ii*) named entity recognition, (*iii*) co-reference resolution, and (*iv*) sentence splitting. First, for each Web page, Jsoup (Java HTML Parser, [https://jsoup.org]) is run to cast it into plain text document. We then use a multi-lingual named entity tagger [31] to annotate the text for coarse-grained lexical entity types including person, location, and organization. The annotated text documents are passed to a rule-based co-reference resolution module to identify co-reference chains for all the entities mentioned in each document. The mentions in every co-reference chain of interest are then replaced with their corresponding representative mentions. Next, for the co-reference chain of interest within each document, we split the document to sentences. We note that, in our implementations, we focused on formal-style sentences. A formal-style sentence follows prescribed writing standards, and prepared for a fairly broad audience [32], [33] A formal-style sentence is often complete and contains a subject, verb and an object. The pre-processing may produce errors, which propagate to the latter stages. However, improving the pre-processing tools is beyond the scope of this paper. The remainder of the processing described in the following use this pre-processed text.

3.2 Semantic Analysis

The semantic analysis component takes as input the pre-processed formal-style text, extracts SRL frames and dependency graphs from the sentences of pre-processed text, and augments them with word senses derived from a distant knowledge base. Semantic parsing consists of four subtasks: (*i*) word sense disambiguation, (*ii*) SRL frame extraction, (*iii*) dependency parsing, and (*iv*) semantic enrichment. In the following we describe these subtasks in more detail.

3.2.1 Word Sense Disambigua

Word sense disambiguation (WSD) provides a sense mapping from surface words and entity mentions in a given text to concepts and named entities in an ontology. In WSD stage, we first disambiguate the word senses using Babelfy [28], a state-of-the-art entity linking and word sense disambiguation system. We filter the resulting senses by pruning the senses corresponding to short-tail mentions that covered by other long-tail mentions. We then map surface textual words and entity mentions to word senses and named entities in BabelNet ontology [28]. Figure 3(a) shows the WSD result for a sample sentence. In Figure 3(a), notation bn:in refers to the i-th BabelNet sense for the given word.

3.2.2 SRL Frame Extraction

In the stage of SRL frame extraction, we assign a "who did what to whom, when, where, why, and how" structure to the sentences of text. We use a rule-based semantic role labeller system [34] to annotate constituents of the sentences with semantic roles. We extract SRL frames from the output of semantic role labeller system. An SRL frame consists of a verb predicate and a number of semantic role elements. Let $F = \{p, E_1, \dots, E_m\}$ be an SRL frame in which p denotes the verb predicate, and E_i is the *i*th SRF element in the frame. Each element E_i is a (s, g) pair, where s indicates the type of semantic role, and g denotes the value for the underlying argument. We note that there may be multiple SRL frame in a sentence depending on the number of verbs in the sentence. In our implementations, types of semantic roles in the output of semantic role labelling system follow the annotation guideline in VerbNet [35]. Figure 3(b) shows a sample sentence along with SRL frames extracted by semantic role labelling system. Semantic role labelling system produces errors (e.g., incorrect argument boundary, or incorrect associated semantic role labels to words), which propagate to the later stages. However, improving the semantic role labelling system is orthogonal to our problem and out of the scope of this paper.

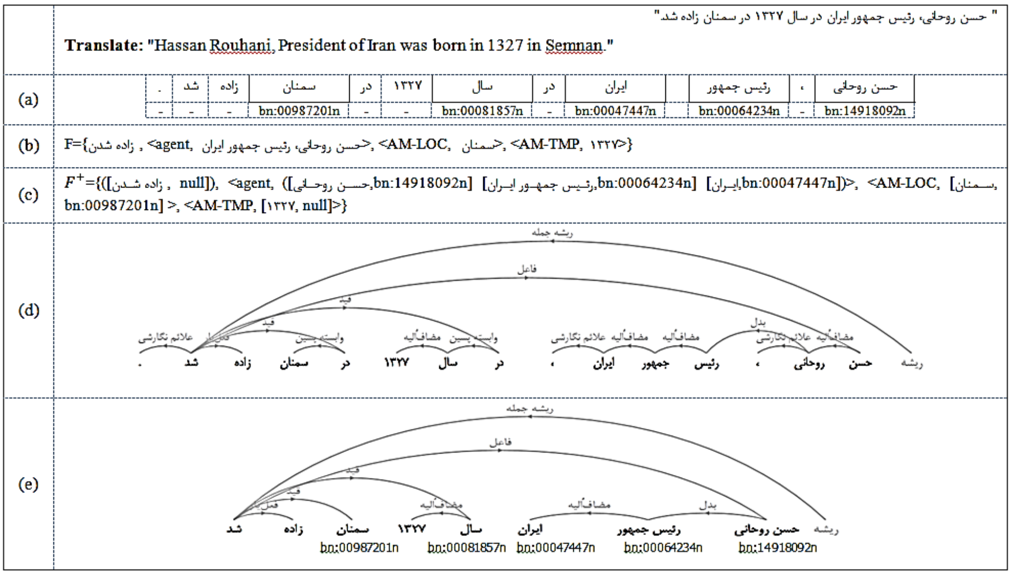
" حسن روحانی، رئیس جمهور ایران در سال ۱۳۲۷ در سمنان زاده شد."
Translate: "Hassan Rouhani, President of Iran was born in 1327 in Semnan."
(a)
. | شد | زاده | سمنان | در | ۱۳۲۷ | سال | در | ایران | | رئیس جمهور | ، | حسن روحانی
- | - | - | bn:00987201n | - | - | bn:00081857n | - | bn:00047447n | | bn:00064234n | - | bn:14918092n
(b)
F={زاده شدن , <agent, حسن روحانی، رئیس جمهور ایران>, <AM-LOC, سمنان>, <AM-TMP, ۱۳۲۷>}
(c)
F+={([زاده شدن , null]), <agent, ([حسن روحانی,bn:14918092n] [رئیس جمهور ایران,bn:00064234n] [ایران,bn:00047447n])>, <AM-LOC, [سمنان, bn:00987201n] >, <AM-TMP, [۱۳۲۷, null]>}
(d)
ریشه جمله
فاعل
قید
قید
علائم نگارشی
وابسته پیشین
مضاف‌الیه
وابسته پیشین
علائم نگارشی
مضاف‌الیه
مضاف‌الیه
بدل
علائم نگارشی
مضاف‌الیه
. شد زاده سمنان در ۱۳۲۷ سال در ، ایران جمهور رئیس ، روحانی حسن ریشه
(e)
ریشه جمله
فاعل
قید
قید
مضاف‌الیه
مضاف‌الیه
بدل
شد زاده سمنان ۱۳۲۷ سال ایران رئیس جمهور حسن روحانی ریشه
bn:00987201n bn:00081857n bn:00047447n bn:00064234n bn:14918092n

Fig. 3. semantic analysis for a sample sentence; (a) disambiguated word senses; (b) SRL frame F; (c) semantic enriched SRL frame; (d) dependency graph G_d; (e) syntactic-semantic graph G_{sd}

3.2.3 Dependency Parsing

In dependency parsing stage, we first parse each sentence of the text to obtain corresponding dependency graph G_d. In G_d, each single word figured as a node and word-word dependencies are represented as directed edges between nodes. In other words, G_d represents binary relationships between words of a sentence, in which words are connected with their parent words with a unique edge labelled with a syntactic function. The definition of syntactic functions is given in [36]. Figure 3(d) shows the dependency graph G_d for a sample sentence. Our syntactic analysis component uses Hazm parser[1], and generates a dependency syntactic graph for each sentence. The erroneous syntactic analyzing of a sentence degrades the performance of later components of EP system. However, we alleviate this problem by enriching the syntactic dependencies with semantic information generated at WSD stage.

3.2.4 Semantic Enrichment

The aim of this stage is to augment the SRL frame and the dependency graph G_d with semantic information obtained in WSD phase. To enrich SRL frame elements with semantic information, we replace each SRL element with its corresponding disambiguated sense. Figure 3(c) shows the semantic enrichment result for the SRL frame given in Figure 3(b). To enrich dependency graph G_d with word senses, and create a syntactic-semantic graph G_{sd}, we start from the dependency graph G_d of sentence s, and a set of disambiguated senses for that sentence. If a disambiguated sense is a single token and covers a single node in G_d , it simply attach to the corresponding dependency node. If a disambiguated sense is a multi-word expression and covers more than one node in G_d, we merge the sub-graph referring to the same concept or entity to a single semantic node. Figure 3(e) shows the enrichment result for the graph G_d given in Figure 3(d).

3.3 Attribute Extraction

The attribute extraction (AE) component takes as input the query entity e, and a vocabulary of attributes A, and extracts filler values for the attributes in A. We focus on six kinds of attributes include: ‘تاریخ تولد/ tarikhe tavallod/ date of birth’, ‘محل تولد/ mahale tavallod/ birth place’, ‘بستگان/ bastegan/ relatives’, ‘ملیت/ meliat/ nationality’, ‘شغل/ shoghl/ occupation’, and ‘مدرک/ madrak/ degree’. These attributes are those extensively studied in IE tasks including slot/template filling, and knowledge base population tasks [37], [38]. We observe that the filler values for these attributes are from noun-based constructions, or sentences having a verb, which serves as an indicator for different attribute classes. Hence, we use two types of AE rules to extract attribute filler values: verb-based rules and noun-based rules. We applied verb-based rules on semantic augmented SRL frames and noun-based rules on semantic boosted dependency graphs. In the following, we give more detail about these methods.

3.3.1 Verb-based Attribute Extraction

The procedure of extraction filler values for the given attributes from semantic boosted SRL frames includes two stages: (*i*) frame marking, and (*ii*) frame element mapping. The frame marking is responsible for labelling the verbs or phrases in SRL frames as indicators of possible values for target attribute classes, but it does not specify which of the elements should be considered as a filler value for any given attribute. The frame mapping looks at the elements of marked SRL frames, and decides which element corresponds to which attribute of the person in question.

3.3.1.1 Frame Marking

The frame marking identifies a set of potential SRL frames containing possible values for a given attribute class. We mark SRL frames by selecting an event verb for each attribute class of interest, and tagging frames for that target class. The main idea in using event verbs as attribute class indicator is that event verbs typically convey the main idea of a sentence. Let $X = \{a_1, \dots, a_k\}$, $X \in A$ be the list of attributes on question that their values can be extracted using verb-based AE rules. To mark SRL frames for a given attribute class $a_i \in X$, we define a seed event verb v specific to a_i. We supplement the event verb v with its synonymous verbs using Farsi version of WordNet [39], FarsNet [40] and form synonym vector, $S = \{s_1, \dots, s_m\}$, where $s_i \in S$ is the ith synonymous verb of v, and m indicates the number of synonymous verbs in S. Using synonymous verbs for SRL frame marking solves the problem of syntactic variation, and prevents inducing several patterns for extracting values for the same attribute class. The decision to using FarsNet comes from the fact that it has a flexible and well-defined lexicon schema, which is publicly known and accepted. We notice that an SRL frame argument may have multiple instances of a given attribute class, and could be considered as candidate value for multiple attributes.

Given attribute class recognizers, a semantic frame F is considered as a potential candidate for attribute class $a_i \in X$, if the predicate p in frame F matches up with one of the seed verbs defined for the attribute class a_i. For example, in the SRL frame given in Figure 3(d), the SRL frame marking have found that the verb predicate ‘ زاده شدن/ zAde shodan/ born’ is an indicator for the attribute class of ‘محل تولد/mahale tavallod/birth place’. Frame marking is important, since the tagged SRL frames form the pool of candidates for attributes of interest in the following stages.

3.3.1.2 Frame Mapping

Our procedure for frame mapping takes as input the SRL frame F that has been processed by frame marking, and maps the elements of frame F to corresponding attributes. The final representation we attempt to create

[1] http://www.sobhe.ir/hazm/

from SRL frames is similar to the frames in FrameNet [41]. In other words, we map the universal and verb-specific roles in SRL frames to template-specific roles. In English, there are resources to direct mapping the elements of SRL frame to FrameNet frame elements such as SemLink [42], but there are not still such resource in Farsi. To map the elements of SRL frame to attributes of interest, one can use different machine learning and data mining approaches proposed for slot/ template filling tasks [37], [38]. Since we have not sufficient training data, and the vocabulary of given attributes is a small closed-class set, here, we use a rule-based method to map the SRL frame elements to corresponding attributes. The overall strategy of our approach is similar to the rule-based methods taken by Angeli et al. [43], and Soderland et al. [20] in English, but our way of defining trigger words and extracting attributes' filler values is different.

For each target attribute class, we create manually a number of rules, based on error analysis over the SRL frames obtained from Wikipedia articles written in Farsi. Each rule is expressed as a number of regular expression patterns containing attribute-specific semantic and lexical constraints. These constraints ensure that the candidate SRL frame element to be a valid filler value for corresponding attribute. Each rule is run over given SRL frame and extracts a value for the attribute of interest when all constraints are met. A sample rule is shown in Figure 4, which determines the filler value for attribute class ‘محل تولد/ mahale tavallod/ birth place’. In Figure 4, *p* refers to verb predicate of SRL frame, and *s* and *g* respectively refers for semantic role label and argument value in SRL frame element. It should be noticed that each predicate *p* in an SRL frame *F* may correspond with several syntactic frames in verb valency lexicon, which depends on its sense given in the sentence. Because there is not appropriate verb lexicon representing information about verb senses in Farsi, in our implementation, we assume that syntactic alternations belong to only one sense. However, this assumption makes some errors in the AE phase.

Extracting filler value for attributes using SRL frames has two important advantages: (*i*) since diverse expression forms of sentences with the same meaning are reduced into a single SRL frame, extracting attribute values from SRL frames is much simpler than those relying on syntactic information contained in sentences; (*ii*) because verb-based AE are easy to understand, one can extend and revise the initial AE rules with high quality rules.

Terms *in* rule	***Value***
Trigger seed verb	<*p*: ?birth place>
Query entity in	<*s*: Agent, *g*: ?person entity>
Entity type	<Person>
Attribute value in	<*s*:AM-LOC, *g*: ?location named-entity >
Attribute value type	<Location name>

Fig. 4. A sample rule designed for the attribute class of ‘محل تولد/ mahale tavallod/birth place’. This simple rule specifies the target attribute class and filler values for its arguments.

3.3.2 Noun-based Attribute Extraction

The filler values for attributes that their values contained in noun-based constructions cannot be extracted by verb-based AE rules. For example, in the sentence given in Figure 3, the verb-based AE cannot include that the phrase ‘رئیس جمهور/raeis jomhour/President’ is a filler value for the attribute of “occupation” for the person ‘حسن روحانی/Hassan Rouhani’. To extract the attribute values contained in noun-based constructions, we use a rule-based approach, which exploits the syntactic information produced by dependency parser, and the lexical information in the form of pre-compiled keywords and named entities. We collect the list of keywords from online information resources such as Wikipedia[1], DBpedia[2] and FreeBase[3] ontology, and tables found on the Web. For example, to find candidate values for the attribute of ‘شغل/occupation’, we collect a list of occupations from DBpedia and Wikipedia, and form keyword set ‘*occupation*’. The overall strategy of our AE approach is similar to the implicit relation extraction method developed by Soderland et al. [6]. However, the limitation to their approach are that (*i*) the attribute filler that is a multi-word expression and covers more than one word in dependency graph cannot be extracted, and (*ii*) it cannot be directly applied in Farsi. We modify and extend their approach to extract both single-word and multi-word attributes' fillers from dependency graph. Borrowing the idea from the work of Bovi et al. [27], we couple syntactic dependencies and fully disambiguated entity mentions and word senses to solve the problem of multi-word filler value extraction. Let $Y = \{a_1, \dots, a_l\}$, $Y \in A$ be the list of attributes on question that their values can be extracted using noun-based AE rules. The procedure for noun-based AE is summarized in Figure 5.

```
Input: query entity $e$, query attribute $a \in A$, keyword list $I$ specific
to attribute $a \in A$, and graph $G_{sd}$.
Output: values for attribute $a$, $V$
if ($G_{sd}$ contains keyword $k \in I$ )
        P=shortestPath($G_{sd}$, $e$, $k$); // shortest path between $e$ and $k$
        in $G_{sd}$
        P'=validatePath(P) // filter by constraints
        if (P' is valid)
                $v := k$;
                $V := V \cup \{v\}$;
        end
end
Return $V$;
```

Fig. 5. The AE algorithm for extracting noun-based attribute values

The input to the algorithm is the query entity e, syntactic-semantic graph G_{sd} generated for sentence s, and a set of pre-compiled keywords I specific to attributes Y. The algorithms then iterates on the graph G_{sd}, and looking for a co-occurrence of a keyword $k \in I$, and the query entity e. An entity e and a keyword $k \in I$ in the

[1] https://www.wikipedia.org/

[2] http://wiki.dbpedia.org/

[3] https://www.freebase.com

graph G_{sd} is considered to be related, (*i*) if there is a dependency path between them in which every dependency edge on the path tagged with one of the following labels: *app, moz, npostmod, npremod, apostmod, apremod, nez, npp, nadv, mos, ncl, acl, posdep,* and *predep; and (ii)* if the keyword k and the focus entity co-refer, i.e., they refer to the same entity. This constraint filters the meaningless and erroneous extractions. Since dependency arcs in G_{sd} is directed, there is no guarantee in finding a path between named entities and concepts. Thus, we use an undirected version of the graph G_{sd}, and follow the the assumption of tokens' locality information [44] to find the path between entity pairs. Among all of the paths found between the entity e and the keyword k, we chose the path with the shortest length. The idea behind this constraint follows the shortest path hypothesis [22], which states that the most valuable information about a relation is contained on the shortest path between two relation's argument nodes in the graph. In the dependency graph G_{sd} given in Figure 3(e), the resulting shortest path between entity 'حسن روحانی/Hassan Rouhani' and the keyword 'رئیس جمهور/raies jomhour/President' contains dependency tag *'بدل/badal/app'*. This path is a valid path and meets the constraint, thus the keyword 'رئیس جمهور/raeis jomhour/President' is a valid filler value for the attribute 'شغل/occupation'.

4. Experiments and Results

In this section, we first describe benchmark datasets and performance metrics, and then give the results obtained by our approach and its counterparts.

4.1 Dataset

A key challenge to evaluate our EP approach is the lack of Farsi dataset suited for EP problem. Thus, we created a small Farsi corpus for evaluating our approach. All evaluations were carried out based on manual assessment. We first chose 30 typical person names and then queried Wikipedia for these names. The reason to using Wikipedia articles as benchmark comes from the fact that Wikipedia articles are rich source of knowledge on the Web and they frequently accessed by millions of users. We create our dataset by selecting a sample of 100 sentences from collected Wikipedia articles. Each sentence in the sample dataset contains at least a candidate filler value for one of the six attributes: ' تاریخ تولد /date of birth', 'محل تولد /birth place', 'بستگان /relatives', 'ملیت /nationality', 'شغل /occupation', and 'مدرک /degree'. In order to create ground truth for evaluation, two human annotators independently examined the sampled sentences to identify the relevant attributes, with an inter-annotator agreement. This type of evaluation follows previous work in the field of information extraction [25], [27], [45]. The annotators reached an agreement score of $\kappa = 70\%$ measured by Cohen's kappa coefficient, which considered to be within the substantial agreement boundaries [46]. The number of resulting <attribute, value> pairs in ground truth is 160.

However this dataset is small to evaluate the scalability of EP approach, but it have the desired characteristics that enables us to study the effectiveness of our EP approach in extracting entity-centric information. To the best of our knowledge, we are the first to investigate the EP in Farsi, thus our dataset to study EP is unique.

4.2 Performance Measures

In the experiments, we conducted evaluations using three criteria: (*i*) precision, (*ii*) recall, and (*iii*) F1 measure. For more detail about these metrics refer to [47]. The quality of the results is evaluated by comparing the profile <attribute, value> pairs obtained by the system and those attribute values in ground truth annotated by annotators. Formally, precision (P), recall (R), and F1 measure ($F1$) is defined as follows:

$$\mathrm{P} = \frac{|\mathrm{S} \cap \mathrm{G}|}{|\mathrm{S}|} \tag{3}$$

$$\mathrm{R} = \frac{|\mathrm{S} \cap \mathrm{G}|}{|\mathrm{G}|} \tag{4}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{5}$$

where S is the set of <attribute, value> pairs generated by the system, and G is the set of attribute, value> pairs in the annotated gold standard set.

4.3 Numerical Results and Discussion

Table 1 shows the performances obtained by our AE approaches. In Table 1, we give the average performances of the pure verb-based AE (VAE) method, pure noun-based AE (NAE) method, and the combination AE (EP^{+}) method. In Table 1, we observe that the performance of AE method is increasing when incorporating both VAE and NAE, while either the VAE or the NAE cannot achieve good performance. The EP^{+} approach achieves the best scores. However, the performances are far from ideal. This shows that profile extraction in Farsi text resources is a big challenge, and justifies that more effort is needed in this field.

Table 1. Performances of our AE approaches on the benchmark dataset

Method	P(%)	R(%)	F1(%)
VAE	32.86	20.51	25.26
NAE	43.64	21.14	28.48
EP^{+} (VAE + NAE)	43.69	32.38	37.19

Table 2 shows the detailed performance for the six individual attributes obtained by our EP^{+} system on benchmark dataset. As shown in Table 2, the attributes of 'تاریخ تولد/date of birth' and 'محل تولد/birth place' have achieved good performance, because their instances often expressed by easily predictable patterns in formal-style format. On the other hand, for the attributes of 'شغل/occupation', and 'بستگان/relatives', the approach cannot achieve good performance. The low score for these attributes is partially due to the fact that the set of values, which such attributes can take are often expressed with various forms in syntactic structure and vocabulary.

For the attributes 'ملیت/nationality' and 'مدرک/degree', the approach reports moderate result. Our approach achieved around 18-56% precision, 10-50% recall, and 13-53% F1 score for the given profile attributes.

Table 2. Detailed performance of the six individual attributes obtained by our approach (EP^{+}) on benchmark dataset

Attribute class	**P (%)**	**R (%)**	**F1 (%)**
Birth place	56.25	50	52.94
Date of birth	53.85	43.75	48.28
Degree	40	33.33	36.36
Nationality	53.33	34.78	42.11
Occupation	40.54	22.39	28.85
Relatives	18.1	10	12.9
Overall	43.69	32.38	36.91

We implemented five AE methods as our baseline methods. These baseline methods include: (*i*) SRL-based AE, (*ii*) IMPLIE system [6], (*iii*) UvA_2 system [48], (*iv*) PolyUHK [49], and (*v*) our recent work in [30]. SRL-based AE is appropriate for the extraction of verb-based attributes. This baseline uses a set of hand-crafted mapping rules to map SRL frame elements to attributes in question. The overall strategy of SRL-based AE method is similar to those presented in [23] and [24]. IMPLIE is developed by the University of Washington team for TAC-KBP 2015 track. IMPLIE uses a set of syntactic rules to extract implicit noun-based relations from dependency graph. UvA_2 is developed by the university of Amsterdam team at WePS 2009 sharetask [50]. This method uses lists of pre-compiled keywords and Web-specific patterns for the personal AE. The overall strategy of UvA_2 is similar to AE methods presented in [7], [20], [48], [49], [51]. PolyUHK is a rule-based AE approach, which achieved the best performance in the WePS 2009 sharetask. For each attribute, PolyUHK first identifies a set of keywords and named entities. It then looks for a co-occurrence of one of the keywords regarding the focus attribute and the target person in a sentence. If a co-occurrence is found then the candidate keyword would be considered as a filler value for the focus attribute. Our previous work [30] is a simple pattern-matching method relying on pre-compiled keywords and hand-crafted rules. To fair comparison, we compare our EP^{+} approach with UvA_2 [48], PolyUHK [49] and our previous work [30]; our noun-based AE (NAE) with IMPLIE [6]; and our verb-based AE (VAE) with SRL-based AE method.

Table 3 summarizes the results obtained by baseline methods and our EP^{+} on the benchmark dataset. Comparing to baseline methods, our method outperforms the baseline methods. Our method achieves higher overall F1 score, 10.12% better than UvA_2, 8.34% better than [30], and 3.11% better than PolyUHK. This indicates that incorporating both verb-based AE and noun-based AE, and considering semantic enrichment is effective in increasing performance of the attribute extraction approach.

Table 3. Comparison of results obtained by baselines and our method on benchmark dataset

Method	**P (%)**	**R (%)**	**F1 (%)**
UvA_2 [48]	37.64	21.14	27.07
PolyUHK [49]	42.61	28.4	34.08
Emami et al. [30]	38.15	23.2	28.85
Our method	43.69	32.38	37.19

Table 4 shows the results obtained by our VAE method and SRL-based AE. From Table 4, we notice that VAE is outperformed pure SRL-based AE method. VAE method achieves a F1 score of 25.26% providing an improvement of about 1.16 F1 score points. This clearly shows the effect of semantic enrichment in the extraction of verb-based relations.

Table 4. Comparison of results obtained by our VAE method SRL-based AE on benchmark dataset

Method	**P (%)**	**R (%)**	**F1 (%)**
SRL-based AE	31.05	19.7	24.1
VAE	32.86	20.51	25.26

Table 5 shows the results obtained by our NAE method and IMPLIE system. The results clearly show that NAE outperforms IMPLIE, and achieved higher scores. The main reason to the low score of IMPLIE is that it cannot correctly extract the multi-word attributes, while some noun-based attributes are multiple-word mentions.

Table 5. Comparison of results obtained by our NAE and IMPLIE on benchmark dataset

Method	**P (%)**	**R (%)**	**F1 (%)**
IMPLIE [6]	42.55	19.38	26.63
NAE	43.64	21.14	28.48

Our method still suffers from several challenges that need to be addressed. Our manual investigation over incorrect extractions indicates that the performance scores for profile attributes can be raised if the following conditions are hold.

- *Creating more precise AE rules:* overall, our profiling approach reports low F1 score for some attributes on question. This fact indicates that EP in Farsi is still a big challenge. Obviously, the more precise the AE rules are, the higher the performance scores are. Therefore, if we spend more time in the development of more robust AE rules, the system performance will pick up.
- *Improving the performance of pre-processing components:* our manual investigation reveals that almost half of the incorrect extractions were because of the inefficiency of pre-processing and semantic analysis stages, and not because of the inefficiency of our AE method. Errors in pre-processing and semantic analysis stages are propagated to AE step and cause wrong extractions. Thus the low performance of pre-requisite stages is a bottleneck for efficient EP. However, improving pre-processing and semantic analysis is orthogonal to our problem and therefore out of the scope of this paper. Nonetheless, to alleviate the errors in semantic analysis stage, we enrich the analyzed text with semantic information extracted from a distant ontology. The effect of

semantic enrichment is shown in Table 6. In the table, EP^+ shows the scenario in which semantic enrichment is considered, and -EP shows the scenario when the semantic enrichment is completely disregarded. We observe that semantic enrichment improve the result of EP, and the best results are obtained by EP^+. The errors in semantic analysis stage leads to degradation of EP^- performance from 36.91 to 30.57% in terms of F1 score. This proves the importance of semantic enrichment in EP. We manually correct the errors in the output of pre-processing and semantic analysis, and give correct input to the AE stage. We observe that the results are improved, and the overall F1 score is raised to over 62%. This shows that errors in extractions were not completely because of the inefficiency of AE method.

- *Enriching discourse profile*: in this paper, we focused on the extraction of attributes on question only from the content provided in given dataset. One of the promising solutions to improve the result and alleviate the problem of data sparseness is to enrich local discourse profile with semantic information inferred from distant knowledge bases. This task is considered as our future work.

Table 6. The results obtained by EP^+ and EP^- in terms of F1score

Attribute class	EP^-	EP^+
Birth place	44.4	52.94
Date of birth	45.16	48.28
Degree	25	36.36
Nationality	39.02	42.11
Occupation	22.6	28.85
Relatives	7.15	12.9
Overall	30.57	36.91

5. Conclusion

Entity profiling (EP) in poor-resource languages like Farsi is suffering from several challenges regarding the tools of language processing and annotated data. As an element of EP research to address these challenges, in this paper, we have investigated a specific variant of the general EP problem, namely the person profile extraction from Farsi Web documents. Our approach identifies the persons in question from the text, and extracts their profile information. Our approach first parses each sentence of the text syntactically and semantically, and augments the local information with global semantic information derived from a distant knowledge base. It uses a semantic rule-based method to extract the attributes of persons, and form their discourse profile. We evaluated our EP approach with a small corpus collected from Farsi Wikipedia articles. Experimental results indicate that our approach is capable to extract the entity-centric information with a high performance.

On the whole, our EP approach can be considered as a foundation for more robust approaches to EP. There remain several important points to improve our research. First, we plan to automate the induction of attribute extraction rules which might to improve the performance and decrease manual engineering effort. Second work is to design a generic EP system to cover more entities and accurately extract their profiles. As the final results of EP system depend on the performance of three subtasks including pre-processing, semantic analysis, and attribute extraction, therefore, third interesting future work is to improve the pre-requisites' performance, which eventually can improve the overall quality of EP system. Since the problem of EP is far from being solved, our fourth future work is to integrate different information extraction and machine learning methods to complement the shortcomings of AE approach, and further improve the overall performance. We chose not to tackle cross-document EP, and instead spent our energy on document-level EP. Our EP approach identifies the entity-centric information only within a document, which is not enough for Web data as some information occur across documents. Our fifth future work is to work on algorithms for cross-document EP, which aims to gather the information about an entity distributed on multiple documents. In present study, we only focused on EP in formal-style text, while most of the entity-centric information in Web data is expressed in informal-style text. Finally, we would like to investigate EP in informal-style text.

References

[1] P. Saeedi, H. Faili, and A. Shakery, "Semantic role induction in Persian: An unsupervised approach by using probabilistic models," Lit. Linguist. Comput., 2014.

[2] M. Shamsfard, "Challenges and open problems in Persian text processing," in Proceedings of 5th Language & Technology Conference (LTC), Poznań, Poland, 2011, pp. 65–69.

[3] H. Fadaei and M. Shamsfard, "Extracting conceptual relations from Persian resources," in ITNG2010 - 7th International Conference on Information Technology: New Generations, Las Vegas, Nevada, USA, 2010, pp. 244–248.

[4] M. Moradi, B. Vazirnezhad, and M. Bahrani, "Commonsense Knowledge Extraction for Persian Language: A Combinatory Approach," Iran. J. Inf. Process. Manag., vol. 31, no. 1, pp. 109–124, 2015.

[5] M. Shamsfard, "Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts," Int. J. Comput. Sci. Eng., vol. 2, no. 6, pp. 2190–2196, 2010.

[6] S. Soderland, N. Hawkins, G. L. Kim, and D. S.Weld, "University of Washington System for 2015 KBP Cold Start Slot Filling," in Proceedings of TAC-KBP 2015, Maryland, USA, 2015.

[7] W. Li, R. Srihari, C. Niu, and X. Li, "Entity profile extraction from large corpora," in Pacific Association for Computational Linguistics Conference (PACLING-2003), Harifax, Canada, 2003.

[8] X. YU and W. LAM, "An Integrated Probabilistic and Logic Approach to Encyclopedia Relation Extraction with Multiple Features," in Proceedings of the 22nd

International Conference on Computational Linguistics (Coling 2008), Manchester, UK, 2008, pp. 1065–1072.

[9] T. Lee, Z. Wang, H. Wang, and S. Hwang, "Attribute extraction and scoring: A probabilistic approach," in ICDE 2013, Brisbane, Australia, 2013, pp. 194–205.

[10] F. M. Suchanek, G. Ifrim, and G. Weikum, "Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents," in Proceedings of KDD, Philadelphia, Pennsylvania, USA, 2006, pp. 712–717.

[11] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen, "StatSnowball : a Statistical Approach to Extracting Entity," in Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009, pp. 101–110.

[12] N. Bach and S. Badaskar, "A review of relation extraction," Lit. Rev. Lang. Stat. II, 2007.

[13] A. Sun, R. Grishman, and S. Sekine, "Semi-supervised relation extraction with large-scale word clustering," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 521–529.

[14] F. Xu, "Bootstrapping Relation Extraction from Semantic Seeds," Saarland University, Saarbrücken, Germany, 2007.

[15] F. Wu and D. S.Weld, "Autonomously Semantifying Wikipedia," in Proceedings of CIKM' 07, Lisboa, Portugal, 2007, pp. 41–50.

[16] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec , Singapore, 2009, pp. 1003–1011.

[17] K. Eichler, H. Hemsen, and G. Neumann, "Unsupervised relation extraction from Web documents," in Proceeding of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 2008, pp. 1674–1679.

[18] M. Banko, M. Cafarella, and S. Soderland, "Open information extraction from the web," in International Joint Conferences on Artificial Intelligence, Hyderabad, India, 2007, pp. 2670–2676.

[19] S. Soderland, B. Roof, B. Qin, and S. Xu, "Adapting Open Information Extraction to Domain-Specific Relations," AI Mag., vol. 31, no. 3, pp. 93–102, 2010.

[20] S. Soderland, J. Gilmer, R. Bart, O. Etzioni, and D. Weld, "Open Information Extraction to KBP Relations in 3 Hours," in Proceedings of TAC-KBP 2013, Maryland, USA, 2013.

[21] M. Yahya, S. E. Whang, R. Gupta, and A. Halevy, "ReNoun: Fact Extraction for Nominal Attributes," in Proceedings of EMNLP 2014, Doha, Qatar, 2014, pp. 325–335.

[22] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, 2005, pp. 724–731.

[23] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, "Using predicate-argument structures for information extraction," in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03, Morristown, NJ, USA, 2003, pp. 8–15.

[24] M. Gregory, L. Mcgrath, E. Bell, K. O. Hara, and K. Domico, "Domain Independent Knowledge Base Population From Structured and Unstructured Data Sources," in Twenty-Fourth International FLAIRS Conference, Palm Beach, Florida, USA, 2011, pp. 251–256.

[25] P. Exner and P. Nugues, "Using semantic role labeling to extract events from Wikipedia," in CEUR Workshop Proceedings, Bonn, Germany, 2011, pp. 38–47.

[26] A. Moro and R. Navigli, "Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 2013, pp. 2148–2154.

[27] C. Delli Bovi, L. Telesca, and R. Navigli, "Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis," Trans. Assoc. Comput. Linguist., vol. 3, pp. 529–543, 2015.

[28] A. Moro, A. Raganato, and R. Navigli, "Entity Linking meets Word Sense Disambiguation : a Unified Approach," Trans. Assoc. Comput. Linguist., vol. 2, pp. 231–244, 2014.

[29] M. A. Heart, "Automatic Acquisition of Hyponyms from Large Text Corpora Lexico-Syntactic for Hyponymy Patterns," in Proceedings of the 14th conference on Computational linguistics, Stroudsburg, PA, USA, 1992, pp. 539–545.

[30] H. Emami, H. Shirazi, A. A. Barforoush, and M. Hourali, "A Pattern-Matching Method for Extracting Personal Information in Farsi Content," U.P.B. Sci. Bull., Ser. C, vol. 78, no. 1, pp. 125–138, 2016.

[31] R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, "Polyglot-NER: Massive Multilingual Named Entity Recognition," in Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, 2015, pp. 586–594.

[32] E. Minkov, R. C. Wang, and W. W. Cohen, "Extracting Personal Names from Email : Applying Named Entity Recognition to Informal Text," Comput. Linguist., pp. 443–450, 2005.

[33] Y. Chen, S. Y. Mei Lee, and C. R. Huang, "A robust web personal name information extraction system," Expert Syst. Appl., vol. 39, no. 3, pp. 2690–2699, 2012.

[34] Z. M. Arani and A. Abdollahzadeh Barforoush, "Semantic Role Labeling using Syntactic Dependency Analysis and Noun Semantic Catergory," in 20th Annual Conference of Computer Society of Iran, Mashhad, Iran (In Farsi), 2015, pp. 619–624.

[35] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, "A large-scale classification of English verbs," Lang. Resour. Eval., vol. 42, no. 1, pp. 21–40, 2008.

[36] H. Mohagheghiyan, "Comparison of Persian Syntactic Dependency Parsers," Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran, 2015.

[37] M. Surdeanu, "Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling," in Proceedings of the Sixth Text Analysis Conference (TAC 2013), Maryland, USA, 2013.

[38] M. Surdeanu and H. Ji., "Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation," in Proceedings of Text Analysis Conference (TAC2014), Maryland, USA, 2014.

[39] C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, 1998.

[40] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi, "Semi Automatic Development Of FarsNet: The Persian Wordnet," in Proceedings of 5th Global WordNet Conference, Mumbai, India, 2010.

[41] C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck, "Background to FrameNet," Int. J. Lexicogr., vol. 16, no. 3, pp. 1–28, 2002.

[42] C. Bonial, K. Stowe, and M. Palmer, “Renewing and revising SemLink,” in The GenLex Workshop on Linked Data in Linguistics, Pisa, Italy, 2013, pp. 9–17.

[43] G. Angeli, A. Chaganty, A. Chang, K. Reschke, J. Tibshirani, J. Y. Wu, O. Bastani, K. Siilats, and C. D. Manning, “Stanford’s 2013 KBP System,” in Proceedings of the Sixth Text Analysis Conference (TAC2013), Maryland, USA, 2013.

[44] J. Christensen, S. Soderland, and O. Etzioni, “Semantic Role Labeling for Open Information Extraction,” in Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, Los Angeles, California, 2010, pp. 52–60.

[45] A. Fader, S. Soderland, and O. Etzioni, “Identifying relations for open information extraction,” in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, UK, 2011, pp. 1535–1545.

[46] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol. 33, no. 1, pp. 159–174, 1977.

[47] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation,” J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37–63, 2011.

[48] K. Balog, J. He, C. Monz, M. Tsagkias, K. Hofmann, V. Jijkoun, W. Weerkamp, and M. De Rijke, “The University of Amsterdam at WePS2,” in 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, Madrid, Spain, 2009.

[49] Y. Chen, S. Lee, and C. Huang, “Polyuhk: A robust information extraction system for web personal names,” 2nd Web People Search Eval. Work. (WePS 2009), 18th WWW Conf. Madrid, Spain, 2009.

[50] J. Artiles, J. Gonzalo, and S. Sekine, “Weps 2 evaluation campaign: overview of the web people search clustering task,” in 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, Madrid, Spain, 2009.

[51] I. Nagy, “Person Attribute Extraction from the Textual Parts of Web Pages,” Acta Cybern., vol. 20, no. 3, pp. 419–440, 2012.

Hojjat Emami received his BSc degree in Software Engineering from University of Tabriz, Iran. He received his MSc degree in Artificial Intelligence from University of Tabriz, Iran. He is currently a Ph.D student at the Department of Information and Communication Technology (ICT), Malek-Ashtar University of Technology, Tehran, Iran.

Hossein Shirazi received his BSc from Mashhad University, Iran. He received his MSc and Ph.D in Artificial Intelligence from the University of New South Wales, Australia. He is currently an associate professor at the Malek-Ashtar University of Technology, Iran.

Ahmad Abdollahzadeh Barforoush is a professor in Computer Engineering and IT Department of Amir Kabir University of Technology, Iran. He is the author of books entitled “Introduction to Distributed Artificial Intelligence” and “Software Quality Assurance Methodology”. His research areas are: data quality, artificial intelligence, agent-based systems, automated negotiation, expert systems, natural language processing, decision support systems, business intelligence, data mining, data warehouse and software engineering.