Scientia Iranica (2019) 26(1), 26–39

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Special Issue on: Socio-Cognitive Engineering
http://scientiairanica.sharif.edu

A statistical approach to knowledge discovery: Bootstrap analysis of language models for knowledge base population from unstructured text

S. Momtazi[a,*] and O. Moradiannasab[b]

a. *Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran.*
b. *Department of Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany.*

Received 26 April 2017; received in revised form 5 September 2017; accepted 17 February 2018

KEYWORDS
Computational linguistics;
Information extraction;
Statistical language modeling;
n-gram Model;
Relation extraction;
Textual pattern acquisition.

Abstract. This paper proposes a novel approach to knowledge discovery from textual data. The generated knowledge base can be used as one of the main components in the cognitive process of question answering systems. The proposed model automatically extracts relations between named entities in Persian. Our proposed model is a bootstrapping approach based on n-gram (a contiguous sequence of n items from a given sequence of text or speech) model to find the representative textual patterns of relations as n-grams in order to extract new knowledge about given named entities. The main motivation of this work is the characteristic of the sentence structure in Persian which, in comparison to English sentences, is in *subject-object-verb* format. The proposed approach is a purely statistical one, and no background knowledge of the target language is required. This makes our method applicable to any open domain relation extraction task. However, as for our test-bed, the domain of biographical data of international poets and scientists is considered herein to build a knowledge base about them. Qualitative evaluations based on human assessment represent the evidence of the efficacy of our method.

1. Introduction

Question answering is one of the main fields in computational linguistics, which has been addressed using cognitive computing [1]. Cognitive computing can be used as an advanced approach to model how human beings react to any question through a computerized system [2,3].

As described by Kaur and Singh [4], "cognitive systems are complex information processing ones, capable of acquiring information, putting it into action, and transmitting knowledge". This is the exact process that is performed inside a question answering system through the following steps [4]:

1. Understanding natural language and human interactions;
2. Generating and evaluating evidence-based hypothesis;
3. Adapting and learning from user selections and responses.

This process requires detailed knowledge to establish a high coverage and efficient question answering system [5,6]. In such systems, multiple knowledge formats from a large variety of natural language sources, such as textbooks, encyclopedias, newspapers, and literary works, have to be processed to provide a knowledge base [7]. The process of discovering knowl-

*. *Corresponding author. Tel.: +98 21 64542729*
E-mail address: momtazi@aut.ac.ir (S. Momtazi)

doi: 10.24200/sci.2018.20198

edge from textual resources is normally done using information extraction techniques, which are one of the important computational linguistics tasks.

Information extraction focuses on finding relations between named entities. The extracted relations are then used to build knowledge bases that are the main resource in cognitive process of question answering systems.

As mentioned, cognitive computing refers to reading, reasoning, learning, and making inferences from vast sets of content [8]. The generated knowledge bases are the best example of such contents.

Bootstrapping approaches, such as the one presented in this article, hold great potential for knowledge base development. Success of knowledge base development depends on the existence of fast ways to mine the large amount of unstructured texts on the web to extract information that can be added to knowledge bases. By taking advantage of such knowledge bases, question answering systems are enabled to answer complex questions, which require information from different sources. Previous research has shown that automatic relation extraction is feasible for question answering tasks with decent results [9].

Availability of a huge amount of unstructured texts on the web provides us with an opportunity to extract relations between named entities. Among various approaches that have been proposed for this task, semi-supervised machine learning techniques, which use a bootstrapping algorithm, received researchers' attention. In these approaches, in order to extract the relation between named entities, e.g., a person and a location, first, the corresponding named entities in the text are tagged, i.e., all person names and location names in this example. Then, by focusing on the context and mainly on *the terms that occur between the two entities*, representative patterns are extracted which are later used to distinguish the corresponding relation between the two entities. As an example, some of the possible relations between a person and a location include `place-of-birth`, `living-place`, `place-of-death`, and `tomb`. The procedure explained above has achieved promising results for languages such as English with *subject-verb-object* word order, because, in most cases, the two entities appear as subject and object (either direct or indirect) in a sentence, and the verb is served as the critical element to describe the relation between the entities and to match the corresponding predicates. For example, "Albert Einstein *was born in* Ulm", "James Hall *was buried in* New York" or "Daniel Gabriel Fahrenheit *died in* The Netherlands". Such a model that mainly focuses on the text between two entities, however, cannot generate reasonable results for all languages, especially those that do not follow the *subject-verb-object* word order. Applying this method to Persian (as one of the Indo-European languages) suffers from this problem. The word order in sentences of this language is in *subjectobject-verb* format. In this language, the verb always appears toward the end of the sentence, and both subject and object appear next to each other toward the beginning of the sentence. As a result, many straightforward relations, whose representative patterns normally appear between subject and object in English, are not easily distinguishable in Persian.

Such a difference in the structure of the Persian sentences motivated us to study the relation extraction task for this language in more depth. To this aim, this study proposes a language model-based method to find the representative textual patterns of every relation as n-grams to be used for extracting new information. Statistical language models have been widely used for various natural language processing and text retrieval tasks, including opinion mining [10], ad-hoc retrieval [11], sentence retrieval [12], and word prediction [13]. In this paper, we benefit from this approach for information extraction from unstructured texts. n-gram model is an entirely statistical model that measures statistical properties of text strings in a corpus with disregard to the vocabulary, lexical, or semantic properties of the document language. An approach based on n-gram model without employment of any other natural language process tools with promising outcome is worth being studied further. In the current study, we focus on the domain of biographical information of international poets and scientists. For this purpose, we constructed an ontology containing biographical scheme to be expanded with information newly mined by our proposed relation extraction method. Although the use of biographical information of well-known public figures provides clarity and structure to the task, the proposed approach itself is not domain-specific and has the potential to be applied to other open-domain relation extraction tasks. Qualitative evaluations are provided as the evidence for the efficacy of our approach. It is shown that the proposed method yields reasonable results and major improvements over the baseline.

The remainder of this article is structured as follows: Section 2 briefly reviews the state-of-the-art approaches used for relation extraction. Section 3 describes the baseline approach against which we compared our results. The proposed method is presented in Section 4. In Section 5, the steps taken to create a biographical corpus in Persian are described. The results obtained in Section 6 are reported and discussed. Finally, the article is summarized in Section 7.

2. Related works

Information extraction is a challenging task from different perspectives. The input of this task is either semistructured data or raw text. Semi-structured data include mark-ups such as HTML, Wikipedia info-boxes, and web tables or lists, all of which can effortlessly be processed to extract required information [14-17]. Information extraction from raw text, however, needs more complex analysis. The available approaches for extracting information from raw texts are divided into three main categories: (1) manually built patterns, (2) supervised approaches, and (3) semi-supervised approaches.

Hindle [18] employed manually built patterns for deriving semantic relatedness. He provided handcrafted textual patterns in order to extract specific relations from unstructured text. Later on, Hearst [19] proposed employment of lexico-syntactic patterns, instead of textual ones, in order to increase the precision. Several other computational linguistics methods are also used to improve the quality of the patterns [20]. The most common computational linguistics method for this task is part-of-speech tagging. To have full awareness of grammatical structure of a sentence, a dependency parser is also potentially helpful; in particular, a parser with the ability to detect far dependencies is advantageous for the *subject-object-verb* word order problem. Although the precision of such methods is high, they show low recall. To improve the recall for these systems, pattern repositories should be enriched, which is a labor-intensive task.

Another approach to relation extraction is supervised learning. Jurafsky and Martin [21,22] employed machine learning methods for this purpose. Some of the methods proposed for this approach are based on Hidden Markov Models and Conditional Random Fields [23,24]. Another group of models for this task performs likelihood optimization processes [25-28]. The weakness of these methods is their narrow awareness of context. As a result, finding high-quality patterns in long sentences with scattered keywords becomes complicated. It gets even worse when the order of sentence elements is *subject-object-verb*. Bayesian models, logistic regression, and support vector machines are other methods examined to distinguish whether an input sentence is similar to a set of training sentences or not. By providing manually labeled sentences, a probabilistic model could be trained. However, the need for substantial amount of annotated data for training is the drawback of these approaches.

Brin first proposed semi-supervised approaches that iteratively search for better patterns [29]. This process starts with a small set of records, called seed data, as a set of correct facts. The process then iteratively improves these records by adding new patterns to this initial set. This is done by searching through the corpus to construct patterns of texts or linguistic blocks. Extracted patterns are then utilized to find new relations. The newly extracted relations, which are potentially also correct facts, will be used again to enrich the patterns.

This process continues until a stopping criterion is satisfied. The outcome is used to enrich the knowledge base. So far, a variety of tools working in this scheme have been proposed, including Snowball [30], Semagix/SWETO [31], KnowItAll [32], Text2Onto [33], LEILA [34], TextRunner [35], and SEAL [36]. These systems also take advantage of natural language processing tools to improve the results by employing parts-of-speech tagging, lexical dependency parsing or using heuristics for entity disambiguation, etc.

In addition to the above works, there are a limited number of researches on relation extraction in Persian. These works, however, have only focused on generic relations (e.g., is-a and part-of). Moradi et al. [37] defined 17 relations in a "concept-net" of generic relations, making their work similar to (Pantel and Pennacchiotti) [38] in the sense of relation types. Our baseline is designed to work in a manner similar to [38] as it also implements an iterative extraction of relation triples. However, the distinguishing factor between our work and theirs is that all our experiments focus on non-generic relations in contrast to [38] and [37]. In [37], the authors designed a solution that makes use of a combination of Hearst method, machine translation, and Wikipedia infoboxes. Basically, minimally supervised algorithms, like that of Hearst, do not show high performance for generic patterns since system precision greatly decreases from the introduced noise and bootstrapping deviates from the correct patterns after very first iterations. In order to control this deviation, [37] employed Google machine translation and Wikipedia Infoboxes to support further knowledge with respect to the knowledge extraction.

There are not many works on knowledge base population in Persian. Shamsfard and Barforoush's [39] *Hasti* project is one of the few works that uses small seed sets to extract relations from sentences. The extraction process, as explained in [39], is a combination of logical, template-driven, and semantic analysis methods. Besides the patterns of a text for sentences containing certain relations, they also extracted the implicit knowledge from a given sentence by logical reasoning at an inference engine, which covers basic reasoning, such as inheritance laws, transitive rules, etc. [40]; it is also a simple method for enriching the Persian WordNet by combining "direct semantic contexts" of some initial concepts. This work also mainly covers generic relations such as *hypernymy*, *part-of* and *definition*. Most of the other works on knowledge acquisition in Persian, such as [41] and [42], are focused on extracting terminologies and lexicons, rather than ontological relations.

3. Baseline

To compare our model with the state-of-the-art methods, the work proposed by Ravichandran and Hovy [43] is considered here, which is based on extracting patterns used for question-answer pairs. Their approach involves a fixed set of seeds containing correct question-term-answer-term pairs, e.g., "Mozart" and "1756" as a seed for "birth year" questions. In order to obtain patterns for a question type, terms are extracted for each seed document containing both the question term and the answer term. Occurrences of the question terms in these documents are replaced with <question> and occurrences of the answer terms with <answer>. Suffix trees are employed to extract the longest matching substrings. These substrings, along with their frequencies, are the raw patterns. In the next step, the authors use a precision scoring algorithm to rank these raw patterns. Their scoring mechanism is based on calculating the ratio of correct answers to all answers retrieved for a given pattern. An adaptation of their pattern extraction algorithm is presented to extract patterns for subject-object pairs in a given relation, e.g., born_in_year (Mozart, 1756).

The second important source for implementing our baseline is the work of Pantel and Pennacchiotti [38], which presents the Espresso algorithm for iterative extraction of relation triples from the web by leveraging generic patterns. Their approach is also based on an initial seed set of known relations. These seeds are fed to a pattern extraction algorithm similar to [43]. Unlike that work, however, the authors of [38] extract relations over several iterations, adding the top k most reliable patterns to extract new seeds for subsequent iterations. The contributions of their work further include metrics for pattern and instance reliability based on association strength and mutual information. The ranking algorithm applied to patterns is fairly complex, yet superior to the one described in [43]. Our baseline is designed to also work in a similar manner. Pattern and instance reliability metrics, such as the ones used by [38], are added, and the processing flow over several iterations continues, always adding the highest ranked relations extracted in each iteration as seeds in the subsequent iteration. Figure 1 shows the process flow within such a bootstrapping approach.

The main overlap between the work of [38] and our baseline is the process of instance extraction. However, in order to assess the precision and productivity of a given pattern, the patterns based on the frequency of their occurrence in the corpus are ranked considering the corresponding correct subject-object pairs in the knowledge base. The extracted instance relations are ranked based on the frequency value of the patterns used to extract them. An accumulative approach is used to rank instance relations, i.e., if an extracted subject-object relation is verified by more than one pattern, its score will be the summation of the frequency values of all those patterns.

Replacing the seeds in the corpus is the first step towards extracting patterns. For each relation of a person, all its possible subjects and objects are replaced in the corresponding biography file with <subject> and <object>, respectively. Thus, a new version of the corpus is generated for each relation with annotated subjects and objects, called the corpus of *candidate sentences*.

In this study, Racichandran and Hovy's algorithm [43] is followed and a suffix tree-like structure is used for finding frequent patterns. In our implementation, we make use of suffix array structure instead of suffix trees. The suffix array is a stripped-down data structure based on the suffix tree, yet pattern

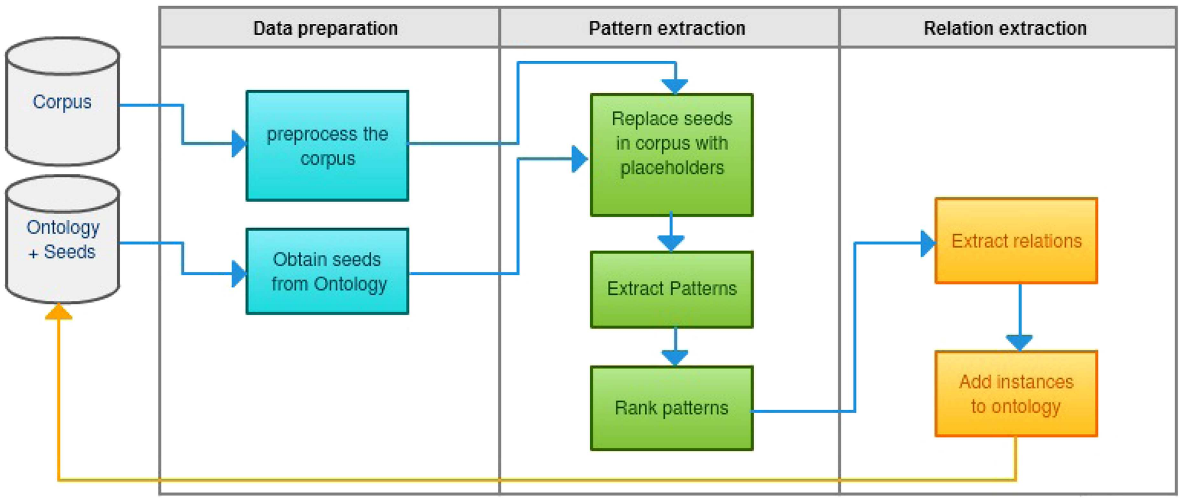

Figure 1. Bootstrapping system architecture.

matching is very fast. The suffix array is slower in some pattern searches than the suffix tree, yet uses less space and is used widely, as compared to the suffix tree. This data structure is used to compute the longest frequent substrings containing both `<subject>` and `<object>`. Each of these substrings is considered a *pattern*. Furthermore, the process we implemented counts the frequency of each pattern and returns this number along with the pattern itself. The quality of a pattern is assessed via its frequency of occurrence in *candidate sentences*, assuming that patterns that are more frequent are more productive and reliable.

Before proceeding to extract relations using the discovered patterns, additional processes are applied to the corpus. One of these pre-processes is Named Entity Recognition (NER) with the aim of improving relation extraction process. Ravichandran and Hovy [43] proposed it to improve precision. The extraction process is applied to the NER tagged corpus. While extracting relations, the strings of the `<subject>` and the `<object>` positions are verified to agree with the expected Named Entity types; for example, the subject of `date-of-birth` relation is a `Person` and its object is a `Date`.

In analogy to the frequency-based ranking of the patterns [43], the extracted instance relations are ranked based on the frequency of the patterns, which are used to extract them. An accumulative approach is used for ranking instance relations, meaning that if a relation is verified by more than one pattern, its score is calculated by summing all the pattern frequencies. In other words, the ranking criterion is a (cumulative) score of the matching patterns.

4. The proposed method

Continuous accretion of knowledge represented in unstructured texts over the World Wide Web affirms the requisite for effective methods to automate information retrieval tasks. Many approaches have been currently proposed to introduce retrieval methods with high precision and/or recall. Our method is based mainly on algorithms integrating n-gram language models to identify segments of the text with the aim of gathering some predefined semantic relations.

Since a bootstrapping approach is used in our work, a set of sample relations is required as the initial seed data. This means that given each semantic relation formulated in our predefined ontology, a number of correct subject-object pairs are concerned with the relations that are inserted in the seed data. We took advantage of Wikipedia info-boxes to aggregate such data. The seed data, in addition to the textual corpus, form the input of our system. The system is designed to use the seeds to extract a ranked set of textual patterns for each given relation. No special care is taken while gathering the seeds. In other words, seeds are selected irrespective of whether the subject-object pairs are mentioned anywhere in the corpus or not. Actually, this makes some of the seeds fruitless in the whole process. As a result, even after providing several records to the system, only a fraction of them might exist in the corpus and the others are flagged as out-of-corpus. The first fraction, altogether, forms an *initializer seed* of effective records. Table 1 shows the number of records used as the seed for each relation in our first experiment.

In order to assess the impact of the number of seeds on the quality of relation extraction, in the second experiment, a seed set of a larger size is also used. Table 2 provides a comparison between the small and large sets.

Having the seed data, each record is represented by a triple with the following structure: <subject–predicate–object>. Subject is always the person about whom we are trying to extract relevant information.

Table 1. Number of effective initializing seeds for each relation in the small set.

Relation name	Total number	Out-of-corpus	Effective
`literary-works`	152	126	26
`tomb`	12	5	7
`date-of-birth`	40	13	27
`date-of-death`	38	16	22
`contemporaneity`	33	18	15
`literary-style`	44	34	10
`place-of-birth`	30	14	16
`living-place`	12	3	9
`place-of-death`	18	10	8
`religion`	12	7	5
`nationality`	21	11	10

Table 2. Number of effective initializing seeds in small and large sets.

Seed-set size	Minimum	Maximum
`small`	7 (tomb)	27 (date-of-birth)
`large`	133 (literary-works)	166 (place-of-birth)

Predicate is the type of information (attribute) we are interested in, e.g., where the person is buried (`tomb`). Object is the target named entity holding the relation to the subject.

As mentioned, in contrast to languages like English, whose grammatical structures are in *subject-verb-object* order, Persian sentences are in *subject-object-verb* form. As a result, approaches, such as Hearst's [19] which tries to construct patterns of words between subjects and objects, do not seem promising for Persian. To overcome this problem, our approach is based on the idea to find frequent patterns in the form of n-grams, which possibly occur anywhere in the sentence. A ranked list of such patterns with a weight factor assigned to each is then used to nominate candidate sentences, which represent the target semantic relation.

For every relation defined in the ontology, the whole corpus is searched to find all sentences that include both subject and object of the seed records of that relation. These sentences are assumed to most likely include the lexicosyntactic structure representing the corresponding semantic relation in natural language. As an example, for `place-of-birth` relation, all the records with this predicate in the seed are processed and the whole corpus is searched for a set of sentences containing subjects and objects of each of these records. This set is called $U_{\texttt{place-of-birth}}$ or U in a more general view. Sentences in U, altogether, are then processed to create a statistical model as described below:

The following are some other notations that are used in this section to describe our model:

- p is a variable n-gram with a size ranging from that of a unigram to 4-gram;
- w is a variable word token;
- $f(p)$ is the frequency of p in U;
- p_2 (w) is the most frequent bigram in U containing w;
- p_3 (w) is the most frequent trigram in U containing w;
- p_4 (w) is the most frequent 4-gram in U containing w.

Our method uses an integration of most frequent n-gram patterns for each relation in its corresponding U set. At first, all n-grams ranging from unigrams to 4-grams occurring in U are generated and counted. n-grams consisting of nothing but stop-words are removed from the model; however, any combinations of stop-words with non-stop-words are maintained for further use. After that, the process goes through the unigram list and expands each word w in this list to a bigram if and only if:

$$f(p_2(w)) > f(w)/2. \tag{1}$$

Following this rule, for each word w, the most frequent bigrams containing w are maintained if and only if their frequency is higher than half of the frequency of w. This expansion process continues for every n-gram to its higher order n-grams with the following condition, which is a general form of Formula 1:

$$f(p_n(w)) > f(w)/2. \tag{2}$$

As an example, the word w with a frequency of 38 is one of the most frequent terms in U. First, the word itself is stored as a useful unigram; then, the most frequent bigram containing w $(p_2(w))$ is found. If $f(p_2(w)) > 19$, then the bigram is stored and the expansion process continues toward the trigram level by finding the most frequent trigram containing w $(p_3(w))$ and counting its frequency $(f(p_3(w)))$. If $f(p_3(w)) > 19$, we record the trigram and expand the pattern to 4-gram, given that the condition is satisfied. In each step, the expansion process terminates if the condition is not satisfied by the condition.

Afterwards, expansion process goes over all the words w in U, and the score of each pattern is calculated as follows:

$$S(p_n(w)) = f(w) + 2 * f(p_2(w)) + 3 * f(p_3(w)) + 4 * f(p_4(w)). \tag{3}$$

$f(p_n(w))$ is set to 1 if the word is not expanded at that level. This means that n-grams of higher order get higher scores than those of lower order containing the same word w.

In Persian, compound-complex sentences can be of lengths as high as 50 words, i.e., the highest in our corpus. These sentences often represent several facts corresponding to different semantic relations in an ontology. For example, both `place-of-death` and `place-of-birth` of a person can be mentioned in a single sentence:

šāmlu dar tehrān motevaled šod va dar karaǰ dargozasht.

Shamlu in Tehran born was and in Karaj died

Shamlu was born in Tehran and died in Karaj.

In this sentence, '*motevaled šhod*' (born) is a pattern representing `place-of-birth` relation and '*dargozasht*' (died) corresponds to `place-of-death` relation. Therefore, the above-mentioned strategy which relies only on frequency of n-grams to find the best pattern for a target relation fails, because it returns both objects, i.e., Tehran and Karaj, for both relations, i.e., `place-of-birth` and `place-of-death` for the target person, i.e., `Shamlu`, which is of course not precise. Therefore, by applying such a method to a bootstrapping process, after a number of iterations, the results of these two relations will be a mix of both and a deviation from the initializing seeds happens. In this example, after some iterations, there will not be any discernment between '*motevaled šhod*' and '*dargozasht*' in the process.

To overcome this issue, a string metric is defined and is considered as another factor in the computation: the character-wise distance between the object entity and the n-gram patterns. For each occurrence of a n-gram pattern in any sentence of U, the distance of the n-gram pattern from the object entity is calculated and an average of these distances over all sentences in U is computed ($A(p_n(w))$). This metric as well as $S(p_n(w))$ defined in Eq. (3) are used to assign a weight to every pattern. The following formula shows how these two factors are taken into account:

$$R(p_n(w)) = S(p_n(w)) + (50 - A(p_n(w))). \qquad (4)$$

In this formula, the maximum length of a sentence in our corpus is 50. After calculating $R(p_n(w))$ for each n-gram pattern, they are ranked in descending order of their weight and the top five patterns are selected as the representatives of the target semantic relation. The following examples represent the patterns, and their weight values are identified and ranked in the first iteration for `place-of-birth` relation.

1) 892.46
be donia āmad
to world came
'was born'

2) 500.85
dar šahr
in city
'in city'

3) 362.57
češm be jahān gošud
eye to world opened
'was born'

4) 242.92
yeki az
one of
'one of'

5) 214.92
motevaled šod
born became
'was born'

These five patterns and the values assigned to each are used to collect and rank sentences from the whole corpus. We call this set of sentences *candidate segments* or T and anticipate them to represent the target relation between new pairs of entities.

To construct T, we search through all sentences of the corpus and assign a score to each sentence according to the sum of corresponding weights of the patterns that occur in that sentence. Filtering out the sentences with zero score, we have a set of segments, inside each of which at least one of the patterns occurs. Those segments that include both a named entity of a person and a corresponding named entity of the object type of the target relation are selected. Following our assumptions, it is very likely that these sentences represent a subject-object pair holding the target semantic relation. Segments not including the aforementioned named entities are filtered out.

After this process, each of the remaining sentences includes a person name, a target named entity, and at least one or more patterns and has an assigned weight value. The person named entity forms the subject, and the target named entity forms the object of the relation triple <subject–predicate–object>. The triple also has a weight as a confidence score. In case of re-occurrence of the same triple, the score of the already-existing triple will be additively incremented. This is according to the assumption that if a triple relation is repeated through the text, it is a more reliable triple.

An example of an extracted sentence is as follows:

doctor eric bern dar 10 mei 1910 dar šahre montreāl be doniā āmad.

doctor Eric Bern on 10 May 1910 in city Montreal to world came

Doctor Eric Bern was born on May 10, 1910 in Montreal.

This sentence is one of the top elements in $T_{\texttt{place-of-birth}}$ because it contains two of the best patterns for `place-of-birth` relation: "be doniš āmad" and "šahr". The confidence score of this sentence is the sum of the weights of these two patterns: $892.46 + 500.85 = 1393.31$. Both the person entity (*Eric Bern*) and the location entity (*Montreal*) are tagged in the corpus. Therefore, the following relation is extracted from the above sentence:

Eric Bern | `place-of-birth` | *Montreal*

In case of compound-complex sentences, it is likely for the sentence to include more than one target entity. In such a situation, an adjunct policy is followed. The following is an example of such a sentence with two location entities, i.e., *Tehran* and *Karaj*:

šāmlu dar tehrān motevaled šod va dar karaǰ dargozasht.

Shamlu in Tehran was born and in Karaj died

Shamlu was born in Tehran and died in Karaj.

The above-mentioned process yields two relations between the person and the two location entities. In this case, besides the confidence score, a supplementary measure is specified to make a preference between these two location entities. $D(x)$ is defined as the average of the string distance of every one of the patterns included in the sentence from the object entity. In case of naturally *one-to-many* relations, e.g., `literary-works`, there is no need to pick up only one of the target entities. That is why simply all object entities in such cases can be added to knowledge base holding the corresponding relation to the subject. However, in case of *one-to-one* relations, e.g., `date-of-birth`, we follow the assumption that the object entity with the lowest average distance to the representative patterns of a semantic relation holds that relation with the subject entity. This means that the object entity is chosen with the lowest $D(x)$.

The extracted triples in each iteration are ordered by the confidence score assigned to each. After the end of each iteration, several candidate triples are extracted; however, only the top 10 triples are added to the knowledge base. According to an assumption similar to Hearst's [19], the newly extracted triples are used as additive records to the initial seed. In the subsequent iteration, all the above-mentioned procedure is repeated by the updated seed in order to improve the patterns and extract more triples. In subsequent iterations, new triples are found and ranked and 10 most confident triples are added to the initial seed data. This procedure enriches the seeds and, as a result, is presumed to improve the patterns and the extraction mechanism in each step.

5. Data

In order to evaluate the proposed relation extraction approach in this article, availability of a named-entity tagged corpus of raw texts, including biographies, is necessary. This section describes the process of preparing such a corpus.

5.1. Corpus

The corpus used in the current study consists of 1,932 Persian text documents including 6412 tokens. This corpus contains biographies of international scientists and poets, including both contemporary and classic ones. It is meant to be a corpus of commonplace biography texts, representing various writing styles. The corpus used in this study is arranged by collecting textual documents from several sources available on the web. The data are crawled using a crawler to collect web pages from specific online sources. The fetched documents go through further processing steps explained in Section 5.2.

The corpus is formed as a set of numbered text documents. These documents do not necessarily include the name of the target person in the first few lines due to removing the title of the web document in the cleaning process. For some of the target persons, more than one biography with different writing styles exists in the corpus. However, the respective features of that person, e.g., `date-of-birth`, are guaranteed to be the same in all of such documents (not necessarily in the form of representation though).

5.2. Pre-processing

5.2.1. Cleaning and normalizing

The corpus documents are cleaned by removing HTML markups, i.e., scripts, styles, and tables. A text normalization process is also applied to the text as a post-process. The current implementation of the normalization process leaves some noise such as conflated words due to stripped white-space, or conflated sentences resulting from processing tables. This noise affects the performance of the implemented system in our experiments. More tuning of the corpus, therefore, is highly recommended for future evaluations.

5.2.2. Sentence splitting

In order to extract textual patterns from *within* sentences, a sentence splitting process is also applied on the corpus. Manual inspection reveals that the sentence splitting is performed reliably, except in cases where some noises from the cleaning process interfere, e.g., when punctuation is accidentally removed or is missing due to processing tables of a web page.

Table 3. Third-person pronouns in Persian.

Persian	English
u	he/she
vei	he/she
xod	self
xodaš	himself/herself
xiš	self
'išān	he/she
'in šāer	this poet

5.2.3. Co-reference resolving

Due to the lack of a co-reference resolver in Persian at the time of implementing our system, a minimal one was developed in the course of this study in order to replace the named entities and pronouns, referring to the target person of each biography text.

For this purpose, all occurrences of the target person's name and also all mentions of third-person pronouns in Persian were counted as a reference to the target person of the biography text. Table 3 lists all Persian third-person pronouns. This process is of course subject to many errors that definitely affect the precision of our method in a negative way.

5.2.4. Named entity tagging

We implemented a named entity tagger able to identify the entities in our ontology that makes use of both gazetteers and regular expressions. Table 4 shows the entity types and the methods employed in each case.

5.3. Ontology schema

The proposed approach to relation extraction is thoroughly statistical. Therefore, it should be capable of extracting relations of any kind with the exception of the generic relations discussed in [38]. However, for evaluation purposes, we designed an ontology schema of biographical information by importing a DBpedia ontology dataset with minor modifications. Protégé ontology editor (http://protege.stanford.edu/) is employed for the design purpose. The ontology is stored in RDF format and JENA semantic web framework (https://jena.apache.org/) is used to insert newly extracted relations into the ontology. Table 5 represents an overview of the schema used in the design of our ontology. The purpose of this ontology is storing the biographical attributes of a person. That is why the subject of nearly all relations is a person entity, and the object is an attribute of that person.

Table 5. Ontology schema.

Subject	Predicate	Object
Person	`literary-works`	literary work
Person	`date-of-birth`	date
Person	`date-of-death`	date
Person	`place-of-birth`	place
Person	`place-of-death`	place
Person	`tomb`	place
Person	`living-place`	place
Person	`nationality`	nationality
Person	`religion`	religion

6. Experimental results

This section presents the results obtained from our approach, compared to the baseline, in order to assess the strength and utility of the proposed approach. We carried out an experiment with the whole corpus and the seed set as the input in order to compare the performance of our approach with that of the baseline in terms of precision at different ranks and mean average precision.

Tables 6 and 7 show an overview of the number of relations extracted using this procedure in the first iterations for the baseline as well as our approach, respectively. According to the statistics, the small seed set produces the overall new relations which are much fewer than the large seed set in case of the baseline. In contrast to the baseline, this difference is minor in our approach, meaning that our approach is less sensitive to the number of initiating seeds than to the baseline, and it can generate more patterns and, as a consequence, introduce more new triples even in case of smaller seed sets. It is worth noting that the number

Table 4. Methods used for tagging named entities.

Entity	Method	Format
`place`	gazetteer + regular expression	<[PLACE:xxxxx]>
`country`	gazetteer	<[COUNTRY:xxxxx]>
`nationality`	gazetteer	<[NATIONALITY:xxxxx]>
`religion`	gazetteer	<[RELIGION:xxxxx]>
`literary-style`	gazetteer	<[STYLE:xxxxx]>
`literary works`	regular expression	<[BOOKS:xxxxx]>
`date`	regular expression	<[DATE:xxxxx]>

Table 6. Overview statistics of extracted relations by baseline.

Relation	Iteration	Small seed set	Large seed set
literary-works	1	18	124
	2	20	128
	3	45	128
	4	45	128
	5	44	130
	6	44	130
	7	–	130
	8	–	130
	9	–	131
	10	–	133
tomb	1	0	882
	2	0	883
	3	-	883
	4	-	884
	5	-	884
	6	-	884
	7	-	884
	8	-	884
	9	-	884
	10	-	884
date-of-birth	1	1649	1654
	2	1649	1654
	3	1649	1654
	4	1750	1654
	5	1750	1654
	6	1750	1656
	7	1750	1688
	8	1752	1688
	9	1752	1688
	10	1752	1688

Table 7. Overview statistics of extracted relations by the proposed approach.

Relation	Iteration	Small seed set	Large seed set
literary-works	1	476	467
	2	486	467
	3	502	467
	4	502	467
	5	511	467
	6	511	467
	7	511	467
	8	511	504
	9	511	504
	10	511	504
tomb	1	1370	1692
	2	1406	1692
	3	1587	1663
	4	1587	1663
	5	1738	1663
	6	1768	2193
	7	1877	2193
	8	1794	2193
	9	1794	2193
	10	1794	2193
date-of-birth	1	1261	1290
	2	1261	1248
	3	1115	1248
	4	1493	1248
	5	1493	1248
	6	1289	1248
	7	1289	1248
	8	1289	1248
	9	1289	1248
	10	1289	1248

of relations also varies greatly from one relation type to another; while `date-of-birth` is very productive, `literary-works` can be really rare.

Despite the fact that, in each iteration, only the top 10 records are used in the subsequent iterations, the top 50 retrieved records are evaluated for correctness. Since no ground through data is available for this task, the results of each iteration are manually assessed and precision at ranks 10, 20, and 50 as well as mean average precision are reported. Figure 2 shows these measures for the small seed set in the first iteration while comparing our approach with the baseline. We can see that the results of our approach significantly outperform those of the baseline. Figure 3 shows the same metric using the large seed set in the first iteration. We have the same observation when comparing our method with baseline using large seed data. By comparing the results of using smaller and larger seed sets (Figure 2 versus Figure 3), we can see that while the baseline method fails to produce enough results for `literary-works` and `tomb`, our approach is able to produce results for all relations even with small seed, showing the robustness of our model. It indicates that our model is less sensitive to the size of seed data, i.e., it is able to produce results even with a few number of seed data.

To show the performance of our approach after passing some iterations, we also represent results of the 10th iteration in Tables 8 and 9. As can be seen in the results, our approach outperforms the baseline in the next iterations, too.

The difference in performance of the proposed method over different relation types can be partially caused by the natural difference of the patterns and grammar of sentences representing each relation. In addition to that, the discrepancy in the efficiency and

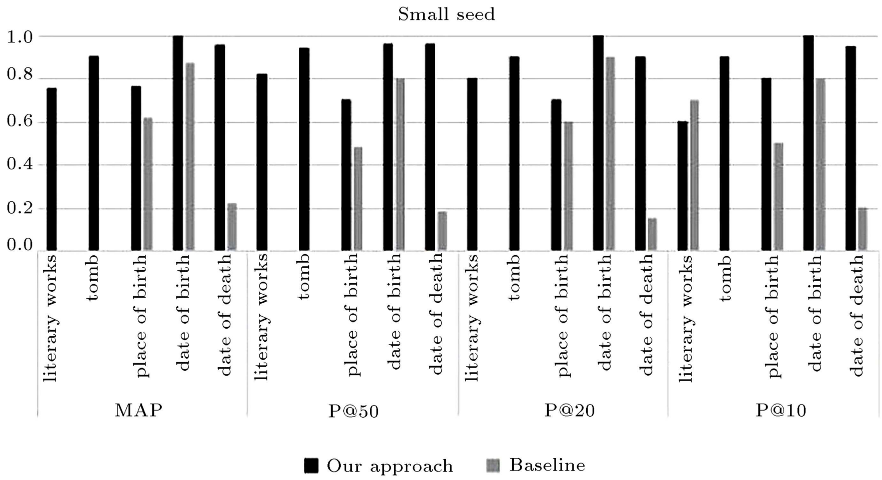

Figure 2. Comparing the results of our approach and the baseline at the first iteration for small seed set.

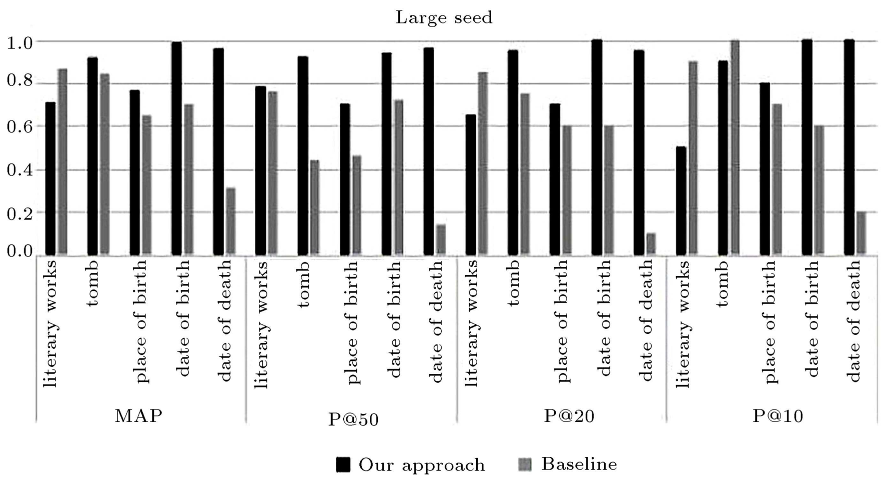

Figure 3. Comparing the results of our approach and the baseline at the first iteration for large seed set.

Table 8. Results at the 10th iteration: small seed set, baseline.

Slot/metric	P@10	P@20	P@50	MAP
`literary-works`	—	—	—	—
`tomb`	—	—	—	—
`place-of-birth`	0.7	0.6	0.54	0.669
`date-of-birth`	1	0.9	0.88	0.920
`date-of-death`	0.1	0.1	0.1	0.129

Table 9. Results at the 10th iteration: small seed set, our approach.

Slot/metric	P@10	P@20	P@50	MAP
`literary-works`	0.6	0.8	0.8	0.761
`tomb`	0.1	0.25	0.24	0.238
`place-of-birth`	0.6	0.65	0.66	0.681
`date-of-birth`	1	1	1	1
`date-of-death`	1	1	0.92	0.975

precision of the named entity tagger employed for detecting different entities can also be a reason for that.

It is also worth mentioning that the above results are achieved while our study suffers from different limitations including:

- The limited size of corpus compared to related studies;
- The lack of a reliable and easy-to-use co-reference resolution mechanism in Persian to facilitate detection of co-references mentions;
- Lack of a high accurate NER tagger in Persian.

These issues do not detract from the project's main aim of demonstrating the efficacy of the proposed approach in comparison to the baseline over a target language of *subject-verb-object* word order. However, possible solutions or improvements to the limitations mentioned above would increase the overall performance of the system developed in the course of the current study.

7. Conclusion and perspectives

In this research work, an approach was proposed to identify textual patterns for predefined semantic relations of a biographical domain from a corpus of Persian textual documents. For this purpose, a corpus of raw texts about people biographies was gathered. The corpus was then tagged with a named entity tagger developed in the course of the current research work to specifically identify entities within the biography domain.

Our ontology schema consisted of a number of triple relations corresponding to biographical attributes of a person. The method worked with some records of correct relations as seed data. The implemented system operated in an iterative manner to add newly extracted records of information to the knowledge base. Our method was thoroughly based on a statistical model of n-grams. The results of the experiment with the proposed method both with small and big seed data are promising, beating the suffix-tree method as the baseline.

Since the proposed method is entirely relying on statistical methods, we believe it should not be that difficult to adapt it to new domains and new languages. The main motivation for this work was the difference in the structure of Persian sentences which, in contrary to English sentence structure, is in *subject-object-verb* format and, as shown, state-of-the-art approaches do not work well on such languages. However, the proposed method is not dependent on any specific sentence structure, and we believe it has the potential to perform well on other languages, too. Moreover, it was shown that the proposed approach does not depend on the size of seed data and can achieve reasonable performance even on small seed sets.

Expanding the current research on other domains and also evaluating the results of a full QA system are part of our future works.

Acknowledgment

The authors would like to thank Manuela Hürlimann and Esther van den Berg who were of great help with implementation of the baseline.

References

1. Chen, Y., Argentinis, J.E., and Weber, G. "IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research", *Clinical Therapeutics*, **38**(4), pp. 688-701 (2016).
2. Gowda, N. and Rekha, K. "Implementation of cognitive approaches in question-answering system", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, **5**(10), pp. 2548-2551 (2016).
3. Bhati, R. and Prasad, S.S. "Open domain question-answering system using cognitive computing", *6th International Conference-Cloud System and Big Data Engineering (Confluence)*, pp. 34-39 (2016).
4. Kaur, S. and Singh, I. "Cognitive computing: Building a smarter planet", *International Journal of Computer Science Trends and Technology (IJCST)*, **4**(2), pp. 325-329 (2016).
5. Aghaebrahimian, A. and Jurcicek, F. "Open-domain factoid question-answering via knowledge graph search", *In Proc. of the NAACL Workshop on Human-Computer Question Answering*, pp. 22-28 (2016).
6. Yahya, M., Berberich, K., Ramanath, M., and Weikum, G. "Exploratory querying of extended knowledge graphs", *Very Large Data Bases (VLDB) Endowment*, **9**(13) pp. 1521-1524 (2016).
7. Furbach, U., Schon, C., and Stolzenburg, F. "Cognitive systems and question-answering", *Industrie Manageme*, **31**, pp. 29-32 (2015).
8. High, R., *The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works*, In IBM Redbooks: Watson (2012).
9. Yih, W. and Ma, H. "Question answering with knowledge base, web and beyond", In *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1219-1221. ACM (2016).
10. Karimi, S. and Shakery, A. "A language model-based approach for subjectivity detection", *Journal of Information Science*, **43**(3), pp. 356-377 (2017).
11. Lafferty, J. and Zhai, C. "Document language models, query models, and risk minimization for information Retrieval", *SIGIR Forum*, **51**(2), pp. 251-259 (2017).
12. Momtazi, S. and Klakow, D. "A word clustering approach for language model-based sentence retrieval in question-answering systems", In *Proc. of the 18th ACM Conference on Information and Knowledge Management*, pp. 1911-1914 (2009).
13. Ghayoomi, M. and Momtazi, S. "An overview on the existing language models for prediction systems as writing assistant tools", *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, pp. 5083-5087 (2009).
14. Kushmerick, N., Weld, D., and Doorenbos, R. "Wrapper induction for information extraction", In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)* (1997).
15. Hsu, C. and Dung, M. "Generating finite-state transducers for semistructured data extraction from the web", *Information Systems (Special Issue on Semistructured Data)*, **23**(9), pp. 521-538 (1998).

16. Hammer, J., Garcia-Molina, H., Cho, J., Aranha, R., and Crespo, A. "Extracting semistructured information from the web", In *Proc. of the Workshop on Management of Semistructured Data* (1997).

17. Mecca, G., Merialdo, P., and Atzeni, P. "Araneus in the era of xml", In *Proc. of the IEEE Data Engineering Bullettin, Special Issue on XML* (1999).

18. Hindle, D. "Noun classification from predicate-argument structures", In *Proc. of the Annual Meeting on Association for Computational Linguistics (ACL)*, pp. 268-275 (1990).

19. Hearst, M.A. "Automatic acquisition of hyponyms from large tex tcorpora", In *Proc. of the International Conference on Computational Linguistics (CoLing)* (1992).

20. Califf, M.E. and Mooney, R.J. "Relational learning of pattern-match rules for information extraction", In *Proc. of the International Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 328-334 (1999).

21. Jurafsky, D. and Martin, J.H., *Speech and Language Processing* (2nd Edition), Prentice Hall (2008).

22. Manning, C. and Schütze, H., *Foundations of Statistical Natural Language Processing*, MIT Press (1999).

23. Sarawagi, S., *Information Extraction*, Now Publisher (2008).

24. Sutton, C. and McCallum, A., *Introduction to Conditional Random Fields for Relational Learning*, MIT Press (2006).

25. Freitag, D. and McCallum, A. "Information extraction using HMMs and shrinkage", In *Proc. of Workshop on Machine Learning for Information Extraction*, pp. 31-36 (1999).

26. Freitag, D. and McCallum, A. "Information extraction with HMM structures learned by stochastic optimization", In *Proc. of the International Conference of the Association for the Advancement of Artificial Intelligence (AAAI)* (2000).

27. Seymore, K., McCallum, A., and Rosenfeld, R. "Learning hidden Markov model structure for information extraction", In *Proc. of the AAAI Workshop on Machine Learning for Information Extraction* (1999).

28. McCallum, A., Freitag, D., and Pereira, F. "Maximum entropy Markov models for information extraction and segmentation", In *Proc. of the International Conference on Machine Learning (ICML)*, pp. 591-598 (2000).

29. Brin, S. "Extracting patterns and relations from the World Wide Web", In *WebDB '98: Selected Papers from the International Workshop on The World Wide Web and Databases*, pp. 172-183 (1999).

30. Agichtein, E., Gravano, L., Pavel, J., Sokolova, V., and Voskoboynik, A. "Snowball: a prototype system for extracting relations from large text collections", In *Proc. of the International Conference on Management of Data (SIGMOD)* (2001).

31. Aleman-Meza, B., Halaschek, C., Sheth, A., Arpinar, I.B., and Sannapareddy, G. "SWETO: Large-scale semantic web test-bed", In *SEKE: Workshop on Ontology in Action* (2004).

32. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. "Web-scale information extraction in KnowItAll", In *Proc. of the International Conference on World Wide Web (WWW)*, pp. 100-110 (2004).

33. Cimiano, P. and Völker, J. "Text2Onto - a framework for ontology learning and data-driven change discovery", In *Proc. of the International Conference on Natural Language and Information Systems*, pp. 227-238 (2005).

34. Suchanek, F.M., Ifrim, G., and Weikum, G. "Combining linguistic and statistical analysis to extract relations from web documents", In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD)* (2006).

35. Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., and Soderland, S. "TextRunner: Open information extraction on the web", In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT)* (2007).

36. Wang, R.C. and Cohen, W.W. "Language-independent set expansion of named entities using the web", In *Proc. of the IEEE International Conference on Data Mining (ICDM)* (2007).

37. Moradi, M., Vazirnezhad, B., and Bahrani, M. "Commonsense knowledge extraction for Persian language: A combinatory approach", *Iranian Journal of Information Processing and Management*, **31**(1), pp. 109-124 (2015).

38. Pantel, P. and Pennacchiotti, M. "Espresso: Leveraging generic patterns for automatically harvesting semantic relations", In *Proc. of the International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics (CoLing-ACL)*, pp. 113-120 (2006).

39. Shamsfard, M. and Barforoush, A.A. "learning ontologies from natural language texts", *International Journal of Human-Computer Studies*, **60**(1) pp. 17-63 (2004).

40. Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., Monshizadeh, M., and Assi, S.M. "Semi automatic development of farsnet; the Persian wordnet", In *Proc. of the Global WordNet Conference*, **29** (2010).

41. Hashemi, H.B. and Shakery, A. "Mining a Persian-English comparable corpus for cross-language information retrieval", *Information Processing & Management*, **50**(2), pp. 384-398 (2014).

42. Shamsfard, M. "Towards semi automatic construction of a lexical ontology for persian" In *Proc. of the Language Resources and Evaluation Conference (LREC)* (2008).

43. Ravichandran, D. and Hovy, E. "Learning surface text patterns for a question-answering system", In *Proc. of the Annual Meeting on Association for Computational Linguistics (ACL)*, pp. 41-47 (2002).

Biographies

Saeedeh Momtazi is currently an Assistant Professor at Amirkabir University of Technology, Tehran, Iran. She completed her BSc and MSc at Sharif University of Technology, Tehran, Iran. She received the PhD degree in Artificial Intelligence from Saarland University, Germany. As part of her PhD, she was a visiting researcher at the Center of Language and Speech Processing at Johns Hopkins University, U.S. After finishing the PhD, she worked at the HassoPlattner Institute at Potsdam University, Germany and the German Institute for International Educational Research, Germany as a post-doctoral researcher. Natural language processing is her main research focus. She has worked in this area of research for more than 12 years.

Omid Moradiannasab earned his BS degree at Iran University of Science and Technology, Iran. He received two Master degrees: one in Language Science and Technology from Saarland University, Germany and the other in Language and Communication Technologies from Groningen University, the Netherlands. He has been an Erasmus Mundus Scholarship Awardee from 2013 till 2015 and contributed to a number of EU-funded research projects such as ALIZ-E with the goal to develop interactive robots able to verbally and socially interact with humans. His contribution to Persian computational linguistics includes creating tools and resources such as an ontology-based question answering system, a dependency parser, a named-entity tagger, etc. After completing his master studies, he has worked in several companies Europe-wide in NLP and AI sector. He is currently the director of NLP team in a Berlin-based company, which provides large-scale automatic sentiment analysis on online and printed financial news. His particular fields of interest mainly include machine learning and its applications in language technology.