Journal of AI and Data Mining
Vol 8, No 2, 2020, 227-236.
DOI: 10.22044/JADM.2019.8430.1980

A Novel Approach to Conditional Random Field-based Named Entity Recognition using Persian Specific Features

L. Jafar Tafreshi[1*] and F. Soltanzadeh[2]

1. Computer Research Center of Islamic Sciences (CRCIS), Tehran, Iran.
2. General Linguistics Department, Allameh Tabatabaei University, Tehran, Iran.

Received 13 May 2019; Revised 09 October 2019; Accepted 12 December 2019
**Corresponding author: f_soltanzadeh@atu.ac.ir (F. Soltanzadeh).*

Abstract
Named entity recognition (NER) is an information extraction technique that identifies the name entities in a text. Three popular methods, namely rule-based, machine-learning-based, and their hybrid have been conventionally used to extract named entities from a text. The machine-learning-based methods have a good performance in the Persian language if they are trained with good features. In order to get a good performance in conditional random field-based Persian named entity recognition, several linguistic features have been designed to extract suitable features for the learning phase based on dependency grammar along with some morphological and language-independent features. In this implementation, the designed features have been applied to conditional random field to build our model. To evaluate our system, the Persian syntactic dependency treebank with about 30,000 sentences, prepared in Computer Research Center of Islamic Sciences, has been implemented. This Treebank has named-entity tags such as person, organization, and location. The result of this work show that our approach is able to achieved 86.86% precision, 80.29% recall, and 83.44% F-measure, which are relatively higher than those values reported for other Persian NER methods.

Keywords: *Natural Language Processing, Named Entity Recognition, Conditional Random Field, Dependency Grammar.*

1. Introduction

Natural language processing (NLP), a branch of artificial intelligence, is the ability of a computer program to process the human language as it is spoken.

Processing of a natural language requires some basic and specific tools depending on the system's application.

Basic tools as normalizer, tokenizer, lemmatizer, and specific tools as co-reference resolution recognizer are named entity recognizers and relation extractors.

Named Entity Recognition (NER) or entity identification is a sub-task of natural language processing.

This task finds the categories such as the names of persons, organizations, and locations in a text.

NER has been developed in various languages but limited works have been carried out on Persian texts due to the scarcity of the resources and tools in recognizing Persian named entities.

Most of the works done on recognizing Persian named entities have used rule-based methods. These systems are not necessarily perfect in their performance. The rule-based methods do not have a good coating on the dispersion attribute of the components and phrases in the Persian language. Moreover, they do not cover various structures in Persian.

Some of these rule-based systems work based on dictionaries and lists of named entities, and their good performance depends on these resources, which may not cover all the available named entities. Besides, the boundary of a Named Entity (NE) may differ from one to another in those lists or dictionaries.

The obvious disadvantages of the rule-based systems are their need for skilled experts to encode rules from the language structure to NLP, enhance them, and avoid their contracting continuously.

On the other hand, machine learning systems learn a language through the use of statistical methods without being explicitly programmed.

The main problem with using machine learning in NLP is the lack of annotated training data.

By rectifying the mentioned problem, this approach speeds up the development of NLP systems significantly. In this research work, we used entity-rich corpus labeled and checked by the experts.

One of the famous machine learning methods that has been used in many NER systems such as Stanford NER system is Conditional Random Field (CRF), which acts as statistical modeling [1]. CRF is a supervised learning method that specifies the probabilities of possible labeled sequences for an observed sequence.

2. Related work

More than a hundred million people speak the Persian language in the world. However, to the best of our knowledge, very limited research works have been carried out on NER for Persian texts. This is due to several factors such as the lack of the Persian NE resources. However, there are some other problems in processing the Persian language, which will be explained in the following part.

Finkel et al. (2005) [2] have presented an approach for English NER based on some statistical algorithms as HMMs, CMMs, and CRFs.

They used Gibbs sampling, a sample Monte Carlo method used to perform an approximate inference in factored probabilistic models.

They used simulated annealing in the sequence models such as HMMs, CMMs, and CRFs. They achieved 90.2% for F-measure in S&M CRF.

The drawback of their work is their computational cost.

Shamsfard and Mortazavi (2009) [3] have worked on a rule-based system for Persian texts. They used the contextual patterns and lexical evidence to recognize Persian NEs and obtained a 72% precision and a 76% recall.

The rule-based approaches have some disadvantages. Some rules that work correctly in some domains may make errors in the other ones. We should always determine the domain of an input text to apply the related rule.

Khormuji and Bazrafkan (2014) [4] have presented an approach based on local filters to recognize NEs. They used a look-up dictionary to detect the NE candidates and filter based on false positives. A designed recognizer uses multiple dictionaries created from the entities of the National Library and Archives Organization of Iran (NLAI). Their dictionary-based recognizer performed the Persian language with an 84.86% precision, a 71.40% recall, and a 72.7% F1 score using exact string search (ESEM). The recognizer obtained an 88.95% precision, a 79.65% recall, and an 82.73% F1 score using approximate string search (ASEM).

In the rule-based systems that work based on dictionaries and lists of NEs, a good performance depends on these resources, which may not cover all the available NEs.

Mehdizadeh Seraj et al. (2014) [5] have introduced semi-supervised models to recognize Persian NEs using Parallel Persian-English corpora. They released a Farsi NE identifier (without using specific features of Farsi) for the first time with a 74% F1 score.

Zafarian et al. (2015) [6] have proposed an un-supervised NER using Parallel Persian-English corpora. They obtained a 72.79% precision, a 62.94% recall, and a 67.51% F1 score.

Limited researchers such as Poostchi et al. (2016) [7] have used machine learning methods by focusing on the pipeline word embedding by Hellinger PCA and classification by a structural SVM-HMM using a subset of Bijankhan corpus. Their research scored 72.59% of f-measure for MUC7 and 65.13% for CoNNL.

Abdous et al. (2017) [8] have proposed another approach using morphological rules, adjacency, and text patterns. They evaluated their method using Bijankhan corpus [9] and got 78.79% for f-measure, and could improve this parameter to 81.92% by adding the Izafe feature.

BiLSTM-CRF is a recurrent neural network and conditional random field algorithm, which has been adopted in [10]. In this research work, an approach for Persian NER based on deep learning is presented. In the system, sentences are pre-processed by LSTM, and an intermediate representation is produced. Then the output is used as input for CRF. They also released several word embeddings trained on a sizable collation of Persian texts. The combination of BiLSTM-CRF and the pre-trained word embeddings allowed them to achieve the 77.45 CONLL F1 score.

As we can see, several research works have been done in Persian named entity recognition and most of them have used rule-based, learning algorithms or deep learning to recognize NEs and have compared the results of their system with others but there are a very few works that have focused on the Persian rich linguistic features.

In this research work, we focused on the Persian rich linguistics features.

3. Persian processing challenges

The following shows some of the challenges that have made the processing of Persian language difficult as far as Persian NLP is concerned.

- Limited training annotated data in Persian.
- No preference for capital and small letters in the Persian language, unlike English.
- Separate prefix and suffix makes it difficult to properly detect the boundary of a noun.
- Great freedom in order of words in Persian.

The following states an example of freedom in word order:

“I gave the book to Ali in the school”:

This sentence can be written in various ways with the same meaning, as bellow:

1. “من در مدرسه کتاب را به علی دادم.”
2. “**در مدرسه** من کتاب را به علی دادم.”
3. “**کتاب را** من در مدرسه به علی دادم.”
4. “**به علی** من در مدرسه کتاب را دادم.”

The first sentence has an unmarked word order because it starts with the subject. In the next sentences, other elements are topicalized and located at the beginning of the sentence. In the second sentence, the prepositional phrase that is locational adjunct is topicalized.

In the third sentence, the direct object that is “کتاب” is focused, and in the fourth sentence, the indirect object that is “علی” has appeared at the beginning of the sentences.

4. Dataset

Among the different existing grammatical theories, the dependency grammar theory was found to be the closest and most suitable one to be applied in processing the Persian language.

In this grammar, the dependency relations are shown by the dependency between the words.

Persian syntactic dependency Treebank [11], prepared in the Noor Islamic Science Computer Research Center, is the first syntactic dependency Treebank including approximately 30,000 sentences randomly collected from the web and annotated with dependency, part of speech, and NER tags. Then in the project called Persian Proposition Bank (PerPB), the Noor researchers added a layer of predicate-argument information to the syntactic structures of Persian Dependency Treebank [12].

Moreover, the Noor researchers added sentence-level relations defined between clauses in complex sentences, and also co-reference information.

They prepared the first Persian Discourse Treebank and (PerDTB) and Coreference Corpus (PerCoref) [13]. For named entity recognition project, Dadegan treebank was tagged with NER labels by experts manually.

As described, the Dadegan treebank consists of several layers of linguistics information that is suitable for many natural language processing.

Table ۱ . An example of our dataset. Columns from left to right show word ID, word, Part of speech, NER, lemma of the word, Head and dependency relation tag of the word, respectively. NER tags get ‘B’ for the first token of NE and ‘I’ for the inner token

word-ID	Word	POS	NER	Lemma	Head	Dependency Relation
1	تسهیلات	NE	O	تسهیلات	11	Subject
2	بنیاد	NE	B-ORG	بنیاد	1	Ezafe Dependent
3	مسکن	NE	I-ORG	مسکن	2	Ezafe Dependent
4	استان	NE	I-ORG	استان	2	Ezafe Dependent
5	یزد	N	I-ORG	یزد	4	Ezafe Dependent
6	به	P	O	به	11	Adverb
7	طور	NE	O	طور	6	Post-Dependent
8	۱۰۰	NUM	O	۱۰۰	9	Pre-Dependent
9	درصد	RESE	O	درصد	7	Post-Modifier of Noun
10	جذب	NE	O	جذب	11	Non-Verbal Element
11	شده	V	O	شده	0	Root
12	.	PUNC	O	.	11	Punctuation Mark

Table۲ . Number of entities in Persian syntactic dependency Treebank.

Number of tokens	Number of entities	Person	Location	Organization
475225	19826	8526	6255	5045

Following shows the different steps in collecting and annotating the Treebank:

- Sentences are randomly collected from the web and stored with their original length.
- Sentences containing colloquial words removed.
- Spellings of the sentences are checked.
- Sentences are tokenized.

- Tokenized sentences are fed into the Persian verb analyzing tool.
- Sentences are annotated with part of speech tags.
- All of the word processing steps are carried out using Virastyar library [14]
- The preprocessed sentences are given to the dependency parser (MST parser) [15].
- NER tags as person, location, and organization are added to the Treebank in IOB standard format, in which NER tags get 'B' for the first token of NE and 'I' for the inner and the end tokens.

In this Treebank, each word has one head, and the head of each sentence depends on an artificial root word. A sample dependency tree is shown in table 1 for a Persian sentence.

The main reasons for using this Treebank are its similarity to the human language understanding and the consistency of these Treebank with great freedom of word order in some languages such as Persian. Table 2 shows the number of entities in Persian Syntactic Dependency Treebank.

5. Methodology

We proposed a Conditional Random Field-based NER that recognizes named entities using many syntactic features based on dependency grammar along with some Persian morphological and language independent features.

The framework of our approach is shown in figure 1. This figure shows the different steps of our system including pre-processing, feature extracting, and machine learning. The NER process starts by normalizing the text using the Hazm normalization tool [21].

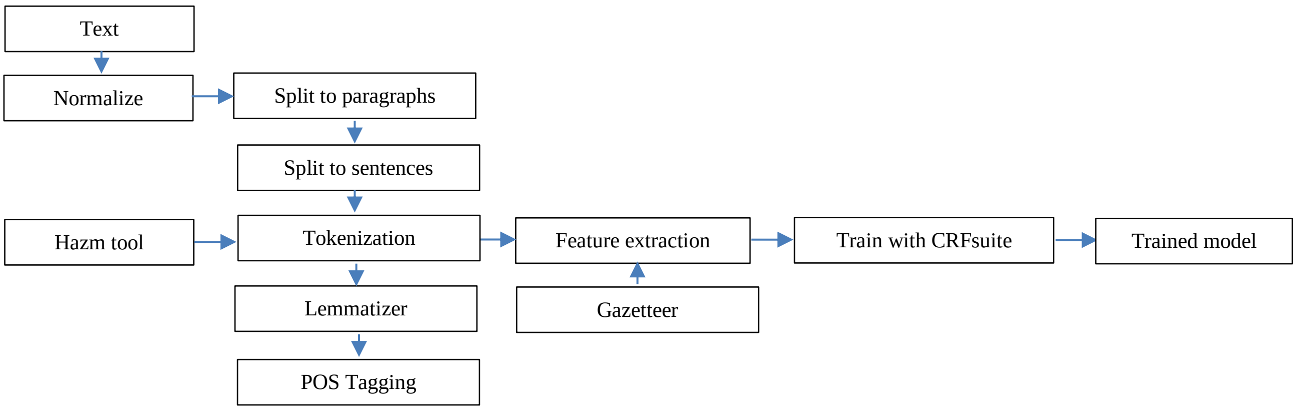

Figure۱ . Model Architecture.

Table۳ . Information about Gazetteer

Lists	Count	Title	Count
Person	24600	Person	112
Organization	18344	Organization	79
Location	7873	Location	510

In the second step, the text is splitted into paragraphs and sentences, respectively. Then the sentences are tokenized, in which the different words and punctuations such as semicolons and full stops are separated.

In the next step, the POS tag is marked for each word. After that, the designed features are extracted for each word with the help of lemmatizer and gazetteer those designed in this approach.

Finally, in the learning phase, these features are used to train CRFsuite, which is an implementation of the conditional random field method.

In the test phase, the trained model is used to guess the named entities.

5.1. Conditional random field

CRFs, trained by maximum likelihood or MAP estimation, assign a probability distribution over the possible labeling described by the following equations:

$$p(z_{1} = N \mid x_{1} = N) = \frac{1}{z} \exp\left(\sum_{n=1}^{N} \sum_{i=1}^{E} \lambda_i f_i (Z_{n-1}, Z_n, x_{1:N}, n)\right) \quad (1)$$

$$Z = \sum_{x_{1:N}} \exp\left(\sum_{n=1}^{N} \sum_{i=1}^{F} \lambda_i f_i (Z_{n-1}, Z_n, x_{1:N}, n)\right) \quad (2)$$

where Z is the normalization factor, which defines the sum of the exponential number of sequences.

These equations show that Z implicitly depends on $x_{1:N}$ and λ parameters.

A big $exp()$ function has been used historically with connection to the exponential family distribution. Within the $exp()$ function, we sum over $n = 1, \ldots, N$ word positions in the sequence. For each position, we sum over $i = 1, \ldots, F$ weighted features.

The scalar λ_i is the weight for feature $f_i()$. λ_i's are the parameters of the CRF model.
Notably, in contrast to HMMs, CRFs can contain any number of feature functions.

5.1.1. Advantages of CRF

Most of the researches in NER such as Stanford NER have shown that CRF exhibits a better performance when compared with HMM in this field. The following outlines the reasons:

- CRF results in a good labeling when good features are designed (e.g. for NER task).
- Independency of features is not required when CRF is applied. Thus it enhances the flexibility of feature selection.
- CRF can use both linguistic (word, characters) and non-linguistic information (punctuation marks, spaces, etc.).

5.1.2. Disadvantages of CRF

The main disadvantage of CRF comes from its complex computation in the training stage. Thus it is difficult to re-train the model after adding some new data samples. In order to overcome this shortcoming, CRFsuite implementation was used.
In the following section, we briefly describe CRFsuite.

Table 4. NER feature sets.

Type	*Group*	*Features*
Word-based	Morphological	Current word, lemma, Number
Word-based	Syntactic	POS of the current word, Surrounding POS, Placement of the word in the sentence
Entity-based	Gazetteer-based	Membership of the current word, Membership of the Surrounding words and Exclusive Membership in the gazetteers, …
Entity-based	Morphological	Existence of affixes in the current word and surrounding words.
Dependency Parse	Syntactic	Dependency relations between words: Object, Mosnad (MOS), Non-verbal element (NVE), …
Hybrid	Syntactic, Gazetteer-based	- hybrid of the Dependency Parse Tree and Membership in the Gazetteers, - hybrid of POS and Membership in the Gazetteers, - hybrid of POS and Membership in the Gazetteers and Izafe construction, …
Hybrid	Morphological, Syntactic and Gazetteer-based	Hybrid of Morphological patterns, Membership in the Gazetteers and POS, …

CRFsuite [16], as an implementation of CRF among the various implementations, was used for labeling sequential data in our approach. It provides not only fast training but also a simple data training and tagging format as the other machine learning tools. Furthermore, CRFsuite provides outputs such as precision, recall, and F1 scores of the evaluated model.

5.2. Feature extraction

In our new approach, in addition to the language independent features, the specific Persian language features such as syntactic features extracted from dependency grammar were used in order to recognize named entities in the text. In summary, we used the morphological-based features as prefixes and suffixes, gazetteer-based features, and syntactic features.
In the process of designing this system, valuable gazetteers of persons, locations, and organizations, described in table 3, are prepared and used. It should be noted that, contrary to the dictionary-based systems, a word belonging to a gazetteer is used only as a feature but not as a direct rule for recognizing NEs.
All the gazetteers in table 3 were gathered from various resources, especially the web. Then they were checked and corrected by Persian linguists.
In table 4, we summarized all features (from all types) used in the suggested approach**.** Here, we explain the features in more details.

1.Word-based features:

- The word,
- The lemma of the word,
- Singularity or plurality of the word,
- POS of the word,
- The previous and next words with the windows of size two and their POS,
- The placement of the word in the sentence.

2.Entity-based features:

♦Location:

- Does the word exist in the location gazetteer?
- Does the word exclusively exist in the Location gazetteer?
- Is the word a location title?

"Mehrabad airport"
"فرودگاه مهرآباد"

- Is there a locational suffix in word?

"آباد" in "علی آباد"

- Is there a locational suffix in the previous and the next words with the window of size three?
- Is the word's suffix a location title?

"Bookstore"
"کتابفروشی"

♦ Person:

- Does the word exist in the person gazetteer?
- Do the previous and two words before exist in the person gazetteer?

"Ms. Parvin Vaezi Kashani"
"خانم پروین واعظی کاشانی"

If the current word is "کاشانی", as we see two previous words are in person gazetteer.

- Is the word a person title?

"Mr. Ahmadi"
"آقای احمدی"

- Are the previous and next words with the window of size three a person title?
- Does the word have the "prefix + person name" pattern?

[پور] + [مهدی] -> [پور مهدی]

- Does the word have "person name + suffix" pattern?

[جمشید] + [لو] -> [جمشیدلو]

- Does the word have "prefix + person name + suffix" pattern?

[ابو] + [تراب] + [ی] -> [ابوترابی]

- Does the word have person suffix?

[رشت] + [چی] -> [رشتچی]

- Does the word have the "location + suffix" pattern?

[کاشان] + [ی] -> [کاشانی]

- Does the word have a person prefix?

[پور] + [مرتضی] -> [پورمرتضی]

- Does the word have "person-title + suffix" pattern?

[آقا] + [یی] -> [آقایی]

♦ Organization

- Does the word exist in the organization gazetteer?
- Do the previous and next words with a window of size three exist in the organization gazetteer?
- Is the word an organization title?

"Office"
"اداره"

- Is the word before or two words before an organization title?

"Whole country ports organization"
"سازمان برنامه کل کشور"

If "کل" is the current word, the two words before is an organization title.

- Does the word exist in the organization gazetteer exclusively?

3. Hybrid features

- Is the word a location title and its POS is a noun?
- Is the word or its next or previous word with the window of size three a person title and its POS is a noun and has Izafe construction?
- Does the word, its previous, and next word with the windows of size three belong to organization title with POS of noun and Izafe construction?
- Does the word belong to location gazetteer and the previous word is an organization title?

"مازندران" in "استانداری مازندران"

(Note that in this example, "مازندران" is a location but "استانداری" is an organization title, so the whole "استانداری مازندران" is an organization)

- Does the word belong to person gazetteer and the two words before is a location title?

"امام" in "حرم امام"

(Note that in the above example, "امام" is a person and "حرم" is a location)

One of our system problems was finding the exact boundary of an entity. In fact, the system could not recognize the full boundary of an NE correctly. Thus we overcame this problem by designing special kinds of features such as the following:

- If the word is an organization title and has Izafe construction, it means that the noun phase is continuing.

"Country assessment training organization"
"سازمان سنجش آموزش کشور"

A number of these features were designed, and finally, some of them were selected by the help of Information Gain (IG), which will be described in the evaluation section.

In the appendix, we listed all these features in a table.

5.3. Dependency features

Dependency grammar has largely developed as a form for syntactic representation used by traditional grammarians.

Dependency-based parsing allows a more adequate treatment of languages with variable word orders, where discontinuous syntactic constructions are more common than in languages like English [17, 18].

Having a more constrained representation, where the number of nodes is fixed by the input string itself, should enable conceptually simpler and computationally more efficient methods for parsing.

At the same time, it is clear that a more constrained representation is a less expressive representation and that dependency representations are necessarily underspecified with respect to certain aspects of the syntactic structure [19].

In this grammar, there are dependency relations between the words. Each word has a head and a dependent on it.

The following shows an example in which a sentence is interpreted incorrectly if there is no information about the syntactic relations in the sentence.

SBJ
MOS
"علیرضا خوشنود است."

"Alireza is pleased"

In this example, "علیرضا" is a subject (SBJ) for a verb and "خوشنود" is a Mosnad (A property of a noun, an adjective or a pronoun ascribed to the subject of a sentence whose main verb is a predicative verb such as the verb forms derived from any of these Persian infinitives [18] for the verb). "علیرضا" is a specific noun in Persian and "خوشنود" is an adjective that can also be a family name. Since "خوشنود" does not have a dependency relation with "علیرضا" in this sentence, it is not a family name.

As we can see, if we do not have dependency relations of the words in this sentence, we cannot find that here "خوشنود" is not a family name for "علیرضا".The above example shows that by having syntactic information, the correct concept of a sentence can be obtained. Therefore, a syntactic level of Persian language was decided to be used in our research work.

In the followiong, eight designed dependency features are introduced.If the relation between the current word and the head is object.

SBJ
OBJ
"چرا احمد محمود را آزار داد؟"

"Why did Ahmad annoy Mahmood?"

In the example, "احمد" and "محمود" have a subject and object relation with the verb, respectively since "محمود" can indicate a person's name or a family name for "احمد". Here, "محمود" does not indicate a family name for "احمد", so without syntactic representation, we cannot recognize the proper boundary of the noun in the above sentence.

1. If the relation between the current word and the head is Non-Verbal Element (NVE).

NVE
"مریم به سارا اعتمادی نداشت."

"Maryam did not trust Sara."

In the above example, "اعتمادی نداشت" is a compound verb and "اعتمادی" is a none-verbal element for "نداشت".

Without syntactic analyses, maybe it realized that "سارا اعتمادی" is an entity and "اعتمادی" is a family name indicating for "سارا".

2. If the relation between the current and the head is Mosnad (MOS).

MOS
"علیرضا خوشنود است."

"Alireza is pleased"

3. If the head of current word is a location title.

Ezafe Dependent (MOZ)
"بوستان لاله"

"Laleh garden"

4. If the head of the current word is a Person title.

MOZ
"آقای احمدی"

"Mr. Ahmadi"

5. If the word has a child which is a Person title.

Pre-Dependent
"آقا جمال"

"Mr. Jamal"

6. If the word has a head which is a geographical direction?

MOZ MOZ
"شمال عراق و مغرب ایران"

"West of Iran and North of Iraq"

7. Does the word have a head which is in Person gazetteers?

MOZ MOZ MOZ
"آقای علی شجایی طباطبایی"

"Mr. Ali Shojaei Tabatabaei"

In the example, "شجایی" may not be in the person list but "علی" is in the person list and the head of "شجایی", thus "شجایی" can be a continuation of the person's name.

5.4. Feature selection

Among many redundant or irreverent attributes in NLP, choosing good features is a difficult and time-consuming process, especially when we

cannot guess the behavior of the data.
Thus using a parameter for selecting features, simplifies this issue.
Here, our feature selection is based on the IG parameter, which, in turn, helps us to find the best features among all the designed features.
IG measures the amount of information an attribute gives us about the class with entropy defined as:

$$H = -\sum_{i=1}^{k} p_k \log_2 p_k \quad (3)$$

Then the change in entropy, or IG, is defined as:

$$\Delta H = H - \frac{m_i}{m} H_i - \frac{m_R}{m} H_R \quad (4)$$

Where m is the total number of instances, with $\boldsymbol{m_k}$ instances belonging to class k, where k = 1… k.

6. Evaluation

To evaluate this project, and to estimate the accuracy in performance of our predictive model in practice, the ten-fold cross-validation was used. Cross-validation averages the measures of fitness in prediction to derive a more accurate estimation of model prediction performance. Thus our dataset is randomly partitioned into 10 equal sizes. Only one of the sub-samples is used testing the model, the nine others are used for training.

Table۵ . ESEM results (%).

Right Match	Person	Location	Organization	Total
Precision	89.11	88.55	75.79	86.86
Recall	82.83	85.14	62.36	80.29
F-measure	82.83	86.79	68.36	83.44

Table۶ . ASEM results (%).

Right Match	Person	Location	Organization	Total
Precision	91.85	89.98	83.20	89.78
Recall	85.38	86.51	68.46	82.99
F-measure	88.49	88.19	75.05	86.24

Table۷ . A comparison between the ESEM results (%)

Method	Precision	Recall	F-measure
HMM-based NER	81.20	36.42	50.28
Unsupervised	72.79	62.94	67.51
Dictionary-based using Local Filters	84.86	71.40	72.70
Rule-based	72	76	73.94
Semi-supervised	79	70	74
Izafe	83	81	81.9
BiLSTM-CRF	-	-	77.45
Our approach	86.86	80.29	83.44
S&M CRF (English NER)	-	-	90.2

Table۸ . A comparison between the ASEM results (%)

Method	Precision	Recall	F-measure
Dictionary-based using local filters	88.95	79.65	82.73
Our approach	89.78	82.99	86.24

This process is repeated for ten times in such a way that each one of the 10 sub-samples is used in turn as the validation data. Finally, we average the ten results to produce a single estimation.

6.1. Evaluation parameters

The proposed method used three evaluation parameters including Precision, Recall and F-measure.
Precision tells us how accurate our method is, in other words, how many of the predicted NEs are correct. Recall calculates the number of actual NEs captured by our model in the labeling process and F-measure investigates the balance between

$$\Pr ecision = \frac{number - of - correctly - recognized - entites}{number - of - recognized - entites} \quad (5)$$

$$Recall = \frac{number - of - correctly - recognized - entites}{number - of - entites - in - the - test - set} \quad (6)$$

$$f - measure = \frac{2 * precision * recall}{precision + recall} \quad (7)$$

precision and recall. These parameters are calculated by the following relations:
In the evaluation of our system, the following metrics are used:

- Exact string evaluation metric (ESEM)

The exact boundaries of the named-entities are considered. Thus in this case, a complete recognition of the named-entity and a correct identification of i type is desired. The following shows an example of this metric.

Organization : { سازمان سنجش آموزش کشور }

• Approximate string evaluation metric (ASEM)

Persian is a head-initial language. Since the Persian transcription is right to left, the head stands in the right. Thus the right boundary of a nominal group should be considered.

In this case, recognizing the right boundary type is desired. The following shows an example of this metric.

Person :مسعود } شجاعی طباطبایی }

Organization :سازمان سنجش } آموزش کشور }

Tables 5 and 6 show the exact match and right match evaluation results, respectively, and table 7 compares the exact string search in our approach with Izafe [7], HMM-based, rule-based [3], dictionary-based using local filters [3], un-supervised [5], and semi-supervised [6] and deep-based [10] NER. In table 8, we compared the approximate string search in our approach with dictionary-based using local filters NER.

As we can see, in comparison to the reported works, we achieved a higher performance by training CRF with rich linguistic features.

7. Conclusion and future work

In this work, we considered the designing proper syntactic and morphological features for the Persian language, which enabled us to improve the capability of the CRF machine learning algorithm in recognizing NEs in a Persian text.

The features such as word-based, entity-based, hybrid, and syntactic features were designed, and among them, features with big IG were selected. Then CRFsuite was trained using the manually NE annotated Persian syntactic dependency Treebank, prepared in the Noor Islamic Science Computer Research Center. Evaluation of the work with standard parameters showed an 86.86% precision and an 80.29% recall for the exact string search and an 89.78% precision and an 82.99% recall for the approximate string search. The final results were compared with the existing rule-based, dictionary-based, and machine-learning-based systems, and it was found that the designed syntactic and morphological features exhibited good performances.

The drawback of our work is the lack of semantic features. If a word like "Iran" that has several meanings and can be various entities in different contexts ("Iran" can be organization, location, and person entities) exists in our corpus in different contexts, our system can recognize the type of the entity properly. However, a word that does not appear in different contexts in our corpus, may rarely be recognized properly. For example, the word "افسانه" possesses two meanings: name of women and fabulous. If in a given text, 'افسانه' means fabulous, our system may recognize it as the name of a person.

In the future works, some semantic features can be added to our system, which avoid the misdiagnosis or non-recognition of Persian NE's. Moreover, using WordNet may solve the problem of words like "افسانه" that have different meanings in various contexts. Some research works [20] have used semantic role labels for recognizing named entities. As our Treebank also has semantic role labels, we can use them to improve our results. Furthermore, many possible shortcomings of the model could be rectified by increasing the amount of data. We can also add other tags to our treebank and use our approach in other applications [22].

Acknowledgments

This project was funded by Computer Research Center of Islamic Sciences (CRCIS). We appreciate the colleagues who helped us in this project: Morteza Rezaei-Sharifabadi, Dr. Mahdi Behniafar, Dr. Azadeh Mirzaei, and the programmers who helped us in implementing the project, Ahmad Eftekhari, Shirin Taghipour, Yousof Alizadeh, and Pooya Alaei; and also our colleagues who helped us with checking the annotation of the corpus. We also wish to thank Majid Jafar Tafreshi for his useful grammatical comments on the paper.

References

[1] Lafferty J., McCallum, A. & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML, pp. 282-289.

[2] Finkel, J. R., Grenager, T. & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics, pp. 363-370.

[3] Shamsfard, M. & Mortazavi, P. S. (2009). Named entity recognition in Persian texts. 15th International conference of Iranian computer community.

[4] Khormuji, M. K. & Bazrafkan, M. (2014). Persian named entity recognition based with local filters. International Journal of Computer Applications, vol. 100, no. 4.

[5] Seraj, R. M., Jabbari, F. & Khadivi, S. (2014, September). A novel unsupervised method for named-entity identification in resource-poor languages using

bilingual corpus. 7'Th International Symposium on Telecommunications (IST), IEEE, pp. 519-523.

[6] Zafarian, A., Rokni, A., Khadivi, S. & Ghiasifard, S. (2015, March). Semi-supervised learning for named entity recognition using weakly labeled training data. In 2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP) (pp. 129-135). IEEE.

[7] Poostchi, H., Zare Borzeshi, E., Abdous, M. & Piccardi, M. (2016, December). PersoNER: Persian named-entity recognition, International Conference on Computational Linguistics-COLING.

[8] Abdoos, m. & Minaei, B. B. (2018). Improving named entity recognition using Izafe in Farsi. Signal and data processing journal, vol. 14, no. 4, pp. 3.

[9] Bijankhan, M., Sheykhzadegan, J., Bahrani, M. & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. Language resources and evaluation, vol. 45, no. 2, pp. 143-164.

[10] Poostchi, H., Borzeshi, E. Z. & Piccardi, M. (2018, May). Bilstm-crf for persian named-entity recognition armanpersonercorpus: The first entity-annotated persian dataset. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[11] Rasooli, M. S., Kouhestani, M. & Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 306-314.

[12] Mirzaei Azadeh, Moloodi AmirSaeid, (2016). Persian Proposition Bank. 10th Language Resources and Evaluation Conference (Lrec).

[13] Mirzaei, A. & Safari, P. (2018). Persian Discourse Treebank and coreference corpus. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC).

[14] Kashefi, O., Nasri, M. & Kanani, K. (2010). Towards Automatic Persian Spell Checking. Tehran, Iran: SCICT.

[15] McDonald, R., Pereira, F., Ribarov, K. & Hajič, J. (2005, October). Non-projective dependency parsing using spanning tree algorithm. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 523-530.

[16] Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields. Available: http://chokkan. Org/.

[17] Mel'čuk, I. (1988). Dependency syntax: Theory and practice, state university of New York press. Arabic Generation in the Framework of the Universal Networking Language, vol. 209.

[18] Covington, M. A. (1990). Technical Correspondence: Parsing discontinuous constituents in dependency grammar. Computational linguistics, vol. 16, no. 4.

[19] Nivre, J. (2005). Dependency grammar and dependency parsing. MSI report, pp. 1-32.

[20] Betina Antony J, G. Suryanarayanan Mahalakshmi (2015). Content-based Information Retrieval by Named Entity Recognition and Verb Semantic Role Labelling. Journal of Universal Computer Science, vol. 21, no. 13.

[21] Nourian A., Rasooli M. S., Imany M., Faili H. (2015). On the Importance of Ezafe Construction in Persian Parsing. ACL, pp. 877-882.

[22] Akkasi, A., Varoglu E. (2019). Improvement of Chemical Named Entity Recognition through Sentence-based Random Under-sampling and Classifier Combination. Journal of AI and Data Mining, vol.7, no. 2, pp. 311-319.

نشریه هوش مصنوعی و داده کاوی

روشی نوآورانه مبتنی بر میدان تصادفی شرطی برای شناسایی موجودیت‌های اسمی نامدار زبان فارسی

لیلا جعفر تفرشی[۱] و فاطمه سلطان زاده[۲،۱*]

[۱] **معاونت تهران، مرکز تحقیقات کامپیوتری علوم اسلامی نور، تهران، ایران.**

[۲] **گروه زبانشناسی همگانی، دانشگاه علامه طباطبایی، تهران، ایران.**

ارسال۲۰۱۹/۰۵/۱۳ ؛ بازنگری ۲۰۱۹/۱۰/۰۹ ؛ پذیرش ۲۰۱۹/۱۲/۱۲

چکیده:

تشخیص موجودیت اسمی نامدار نوعی تکنیک استخراج اطلاعات است که موجودیت‌های اسمی نامدار را در متن شناسایی می‌کند. سه روش اصلی مبتنی بر قاعده، یادگیری ماشین و ترکیبی از آن ها به طور معمول برای استخراج موجودیت‌های اسمی نامدار استفاده می‌شوند. روشهای قاعده‌مند در صورت استفاده از ویژگی‌های مناسب، کارایی خوبی در زبان فارسی دارند. در این پژوهش برای استخراج موجودیت اسمی نامدار با کمک الگوریتم میدان تصادفی شرطی، ویژگی‌های زبانی بسیاری طراحی شد و از بین آنها ویژگی‌های مناسب بر پایه دستور وابستگی همراه با ویژگی‌های صرفی و همچنین ویژگی‌هایی مستقل از زبانی خاص برای فاز آموزش استفاده شد. در این پیاده سازی، برای آموزش مدل از الگوریتم میدان تصادفی شرطی شده است. برای ارزیابی مدل از پیکره وابستگی نحوی زبان فارسی با حدود ۳۰۰۰۰ جمله تهیه شده در مرکز تحقیقات کامپیوتری علوم اسلامی نور، استفاده شده است. این دادگان درخت نحوی، برچسب موجودیتهای اسمی نامدار از جمله شخص، سازمان و مکان را دارا هستند. نتایج این پژوهش نشان می‌دهد که کارایی روش پیشنهادی با ۸۶,۸۶٪ دقت، ۸۰,۲۹٪ بازخوانی و ۸۳,۴۴٪ میانگین هارمونیک دقت و بازخوانی، بهتر از کارهای پیشین است.

کلمات کلیدی: پردازش زبان طبیعی، شناسایی موجودیت اسمی نامدار، میدان تصادفی شرطی، دستور وابستگی.