MorphoBERT: a Persian NER System with BERT and Morphological Analysis

Mahdi Mohseni
Jena University Hospital
University of Jena
Jena, Germany
mahdi.mohseni@uni-jena.de

Amirhossein Tebbifakhr
Fondazione Bruno Kessler
University of Trento
Trento, Italy
atebbifakhr@fbk.eu

Abstract

Named Entity refers to person, organization and location names, and sometimes date, time, money and percent expressions as well. Named entity Recognition (NER) systems are developed to extract these essential information units from a text. Persian is a less-developed language in many natural language processing tasks such as NER. In this paper we present our system, MorphoBERT, submitted to the First Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019)(Taghizadeh et al., 2019). We train the BERT model (Devlin et al., 2019a) on a large volume of Persian texts to get a highly accurate representation of tokens and then we apply a BiLSTM (bidirectional LSTM) on vector representations to label tokens. Persian is a rich language in terms of morphology and word parts may convey grammatical and semantic information. To inform the model of this information we analyze texts morphologically to split the lemma and affix(es) of each word and then we train the model on the analyzed texts. The test data, provided by the organizers, contains in-domain and out-of-domain texts. Our system achieves the first rank among all participated systems with a total high precision, recall and F1 of 87.0, 83.8, 85.4, respectively.

1 Introduction

Named Entity Recognition is a well-known classification topic in the research areas of language processing. NER systems aim to classify tokens of a text into classes such as person, organization and location, which are the most important named entity categories. Numerical expressions such as date, time, percent and monetary values are the other important classes, which are recognized in some systems. Named entities are the essential units of a text because either they convey most important information of the text or the text talks about them.

Various approaches have been used to recognize named entities in a text. Hidden Markov Model (Bikel et al., 1997), Maximum Entropy Model (Borthwick et al., 1998) and Conditional Random Filed (McCallum and Li, 2003) are statistical methods applied to the NER task. Neural network models have been also developed to categorize named entities. Collobert et al. (Collobert et al., 2011) propose a neural model based on a feed-forward architecture that takes into account a window of words around each target word. In (dos Santos and Guimarães, 2015) a convolutional neural network (CNN) is used to extract character-level and word-level embeddings representing contextual and structural word features. Ref. (Chiu and Nichols, 2016) combines a CNN model with a LSTM to utilize the strengths of both models. Lample et al. (Lample et al., 2016) approach the NER task using a hybrid statistical and neural model. In their model, LSTM-CRF, a bidirectional neural model extracts features from a text and CRF labels tokens.

Some research has focused on NER in Persian texts. PersoNER (Poostchi et al., 2016) is a Persian NER system in which a word embedding model and a sequential max-margin classifier are used. In (Poostchi et al., 2018) the LSTM-CRF model developed by (Lample et al., 2016) is applied to Persian texts. Shahshahani et al. (Shahshahani et al., 2019) have recently published a study in which a rule-based system, a CRF model and a LSTM-based system are compared on a newly well-designed Persian NER dataset.

After almost three decades of study, NER is still an open problem, especially for low-resource and under-developed languages such as Persian. The First Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) allocated a track to the Persian NER task (Taghizadeh et al., 2019). This paper presents our system called Mor-

phoBERT submitted to the workshop. In our system, we combine the BERT model (Devlin et al., 2019a) and a BiLSTM model and utilize a morphological analyzer developed for the Persian language. We train the BERT model on a large volume of Persian texts to get a highly accurate representation vector for each token in the input text. The Persian language is a morphologically rich language in which a single lemma may appear in various forms in a text. To allow the model to learn grammatical and semantic roles of lemmas and affixes, we first analyze words and split them into their constitutes. Then we feed them to the BERT model to generate a dense vector representation for them. Afterwards, a BiLSTM network gets the representations generated by BERT and tags tokens with the named entity labels.

Section 2 opens a discussion about the morphology of the Persian language and then explains our morphological analyzer. In Section 3, we describe our Persian NER Model. Section 4 covers resources that we use for training and evaluation of our system and presents the results under various experimental settings.

2 Morphological Analysis

Persian is an agglutinative language in which affixes and clitics attach to the base form of words. Not only verbs are inflected in the Persian language but also nouns and adjectives are highly affected by morphological rules of the language. Other part-of-speeches such as pronouns and adverbs may also get inflected especially in colloquial use. The main word order in Persian is subject-object-verb (SOV). The Persian script has an Arabic root and is written from right to left. In Persian texts short vowels are rarely written. It adds an ambiguity to processing a text as it produces many non-lexical homographs (Bijankhan et al., 2011), inflected words with the same spelling but different meanings and pronunciations.

There are several tenses in the Persian language and each verb is inflected in six different forms according to the person and number of the subject. Persian is a genderless language in which there is no discrimination between male and female, neither in its grammar nor in referring words. Nouns appears in a text as singular or plural. There are a few suffixes which create plural nouns from singulars. Few of these suffixes have been imported from Arabic. There is no definite article in the Persian language. However, indefinite articles have been defined in the language. There is no real possessive pronoun in the language and possession is expressed by adding clitics to a noun or sometimes to an adjective when it accompanies the noun. Fig. 1 shows some sample Persian word and their morphology.

Word	Analysis	Translation
کتابهایش (ketɒ:bhɒ:jaš)	کتاب + ها + ی + ش	his/her books
کتابهای زیبایش (ketɒ:bhɒ:je zibɒ:jaš)	کتاب + ها +ی؛ زیبا + ی + ش	his/her beautiful books
رفته‌ام (rafteam)	رفت + ه + ام	[I] have gone
می‌روم (miravam)	می + رو + م	[I] go
مسئولین (masɒ:lin)	مسئول + ین	authorities

Figure 1: Sample Persian words and their analyses.

Paykare (Bijankhan et al., 2011) is a Persian corpus designed and developed based on the EAGLES guidelines (Leech and Wilson, 1999) to capture the complexity of the Persian morphology. It contains almost 10M words, which have been manually tagged under a hierarchical structure. Although words are categorized into 14 major categories, the tagset consists of 109 distinct tags. A combination of these tags is used to label each word of the corpus. For example, "ketɒ:bhɒ:yaš" (his/her books) has been tagged with "N, COM, PL, 3" which stands for Noun, Common, Plural, 3rd possessive pronoun. The total number of hierarchical tags of words in the corpus rises up to 606 tags. We use this corpus to develop a Persian morphological analyzer.

Developing a morphological analyzer for Persian is very challenging. On the one hand, the Persian morphology is complex and ambiguous and requires an intensive contextual interpretation. On the other hand, some words have a beginning or an ending similar to affixes and clitics that makes the analysis error-prone. If a text could be tagged with the hierarchical tagging system of the Paykare corpus, one can analyze words precisely. But developing a fully automatic Part-of-Speech (POS) tagger with more 600 tags is demanding and a high accuracy is not achieved. We take another more practical approach. Texts in Paykare have been tagged manually, so, they are very accurate and reliable. As the hierarchical tag of a word reveals its structure and the way it has been created, one can develop a system to analyze words of the corpus. However, there are some exceptions which

need to be taken care of differently. For example, borrowed words from Arabic do not follow the Persian morphological rules and may be analyzed wrongly. For the exceptions, we make a list containing words and their correct analyses. The result shows that around 15.5% of the corpus consists of inflected words, which have different lemmas than their original surface forms in the texts.

Some words may have different analysis depending on their contexts. For each word and its major tag, we save the most frequent analysis in a map. For example, the word "ketɒ:bhɒ:yaš" (his/her book) with it major tag, "N", is anlyzed to "Ketɒ:b + hɒ: + y + aš". Tagging a text with only 14 major categories can be accomplished with a high accuracy. For a new text, we label the text with the major POS tags and then search the map to find the analysis of each word. To tag a new text with major POS tags, we use the Persianp toolbox (Mohseni et al., 2018). Since we take only the major tags into account we lose some information and cannot analyze all words correctly. However, the accuracy of the method remains very high. Using this method, only 3% of inflected words in the Paykare corpus are analyzed incorrectly. We use this method to analyze texts before training our NER neural model for the Persian language.

3 Persian NER Model

Our Persian NER system is depicted in Fig. 2. The lower layer is the morphological analyzer. Inflection changes the surface form of words and makes it difficult for a machine learning method to infer the role of words precisely and find out the grammatical and semantic role of lemmas and affixes. To help the model infer this information, words are split into their constitutes in the first layer. The neural part of the model is composed of the BERT model and a BiLSTM which are described below.

3.1 BERT

We use BERT (Devlin et al., 2019b) as a pre-training step. BERT is a language representation model in which bidirectional Transformer (Vaswani et al., 2017) is used in each layer of the model. It is trained by predicting masked words in an input sentence according to the preceding and proceeding words. This model can be trained on large-scale monolingual corpora. One of the advantages of using BERT compared to the word-level approaches such as word2vec

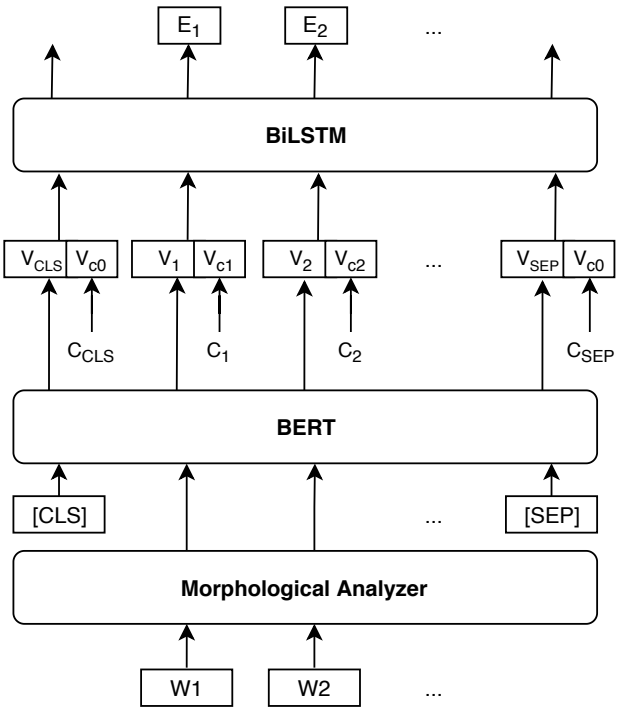

Figure 2: The architecture of the MorphoBERT NER system.

(Mikolov et al., 2013) and Glove (Pennington et al., 2014) is that the representation of each word is not fixed and is influenced by the other words in the sentence.

In our model, we use $\text{BERT}_{\text{BASE}}$, which uses 12 layers of bidirectional Transformer with 12 attention heads and 768 as the hidden size units. We pre-train the model from the scratch. This allows us to use our strategy in analyzing word morphologically and paying attention to the language-specific features. Also, we train the model on Persian monolingual data to have a more accurate model, while the available pre-trained models are multilingual and may have less performance on representing the Persian texts.

As depicted in Fig. 2, the output of the morphological analyzer is delivered to the BERT model. $[CLS]$ and $[SEP]$ are two tokens added by BERT to each input sentence indicating its boundary.

3.2 BiLSTM

We use a bidirectional LSTM to tag the named entities in sentences. As it is shown in Fig. 2, the input of this network is the representation of the sentence obtained from the BERT model. We use a bidirectional LSTM in order to leverage both left and right context to tag tokens. On the top of the BiLSTM, we use a linear model with the Softmax

activation function to get the probability distribution over all tags for each token.

3.3 Word Class Feature

Although neural models are very successful in extracting contextual information from a text, providing explicit features can still improve their performance. In Ref. (Shahshahani et al., 2019) that a LSTM-based model is proposed for Persian NER, feeding a feature representing word clusters enhances the result. This feature is the cluster number of the word, which is given to the model as another input. As Fig. 2 shows, we take the same approach and give the word cluster feature to the BiLSMT network. The representation of words generated by BERT is not fixed, so, we train a word2vec model to get a fix representation for each word. Then we apply a k-means clustering on word vectors. The number of clusters is set to 1500. The distance between instances is computed using cosine similarity. To create the word2vec model and cluster words we use the Gensim library [1]. The cluster number of each word is fed to the model and a cluster number is reserved for unknown words. The cluster numbers have their own embedding vectors, which are learned during training. The size of the vectors is set to 32. The cluster representation and the representation generated by BERT for each token are concatenated into a 800-dimensional vector and is given to the BiLSTM. Our experiments show that adding this feature improves the F1 measure of the system by 0.5%.

4 Experiments

4.1 Unlabeled Text Corpus

To train the BERT model for the Persian language, we collected a large volume of Persian texts consisting of news articles and Wikipedia documents. News articles crawled from 10 online news agencies contain 300M words and the dump of Persian Wikipedia[2] provides texts with more than 75M words. All texts are analyzed with our morphological analyzers and fed to the BERT model. We trained the model with more than 1M steps with the batch size equal to 16. The max sequence length of input sentences is set to 256 and the values of the parameters for masking words is set to the default values i.e. 15%. We use the Adam optimizer with initial learning rate equal to 5×10^{-5} and 10,000 warm-up steps. The vocabulary contains words with frequency more than 80 and its size reaches to 52K tokens.

[1] https://radimrehurek.com/gensim/index.html
[2] https://dumps.wikimedia.org/fawiki/latest/

Table 1: The statistics of the training dataset.

Named Entity	#Entities (phrases)	#Words
Person	12553	21121
Organization	14285	34774
Location	15412	21102
Date	4474	10413
Time	572	1786
Money	1295	4726
Percent	12557	2386
Total	49592	96308

Table 2: The statistics of the test dataset.

	No. Words
In-domain	68063
Out-of-domain	76463
Total	144526

4.2 NER Dataset

The organizers of the Persian NER task in NSURL 2019 have provided a training dataset and the final assessment on the test dataset is blind. The main features of the dataset have been described in (Shahshahani et al., 2019). The provided dataset has a similar structure to the CoNLL format in which each line contains one single word and its label separated by a $< TAB >$. The format of labels are IOB. The dataset contains almost 900K words from which about 50K are named entities. 7 types of entities tagged in the dataset are *person*, *organization*, *location*, *date*, *time*, *money* and *percent*. Table 1 presents the number of entities and the number of words in entity phrases.

The test dataset contains in-domain and out-of-domain texts. Table 2 show the size of the each part. Since the evaluation on the test dataset is blind we do not know the number of named entities in the dataset.

Table 3: The detailed results of MorphoBERT on provided dataset using 5-fold cross validation at the phrase-level for both subtasks.

Subtask	P	R	F1
3-class	87.2	89.2	88.2
7-class	86.2	88.5	87.4

4.3 Results

Once the BERT model is trained with the unlabeled text corpus, its output, the representation vectors of input tokens, is supplied to the BiLSTM network. As previously mentioned, word clusters are also given to the BiLSTM network as an extra features. We apply the same optimization method here as we did for training BERT. We don't fix the parameters of BERT allowing them to be fine-tuned. The number of epochs and the batch size are set to 10 and 32, respectively.

In the Persian NER task of NSURL 2019 (Taghizadeh et al., 2019), two subtasks have been defined. The first one is 3-class Persian NER in which 3 major named entities, *person*, *organization* and *location* are detected. The second subtask, called 7-class Persian NER, takes all types of entities in the dataset into account.

We first report the performance of our Persian NER system, MorphoBERT, on the provided dataset with 5-fold cross validation. Table 3 shows the results of system for both 3-class and 7-class subtasks at the phrase-level.

Table 4 presents the detailed results for all named entities. The evaluation at the word-level, which is obviously higher than the phrase-level, is presented in 5. In the table 'B-' and 'I-', corresponding to the IOB format, indicate respectively the beginning word and the inside word(s) of a named entity. The performance of the system on 3 main classes of *person*, *organization* and *location* is very high. *Percent* and *money* are phrased in a text in a relatively low number of predefined templates and they can be classified with a high precision and recall. In *date* and *time* the performance is lower. This is because of two reasons. First, in the dataset the number of instances for these two types of entities are low, so, the system cannot learn these classes very well. Second, according to the guideline of the dataset temporal phrases are labeled as entities when they are not generic and can be exactly specified knowing the the production time of the document. Therefore, it is very challenging for the system to discriminate between generic and specific temporal expressions. Comparing Table 4 and 5 shows that the results of the system at the word-level is higher as it is expected. It also states that it is more challenging to detect the correct boundary of some entities such as *location*. *I-Location* shows the inside word(s) of location entities. There is about 10% difference in performance between *B-Location* and *I-Location* tags. In Persian many location names are multiword and sometimes they cannot be inferred very well from pre-known instances. Using a rich gazetteer can alleviate this problem.

The organizers of the task evaluated the participated system in both subtasks on the test dataset.

Table 4: The detailed results of MorphoBERT on the provided dataset using 5-fold cross validation at the phrase-level.

Named Entity	P	R	F1
Person	91.5	91.4	91.5
Organization	94.2	88.0	90.9
Location	88.3	90.2	89.3
Date	77.1	82.0	79.5
Time	66.5	75.4	70.7
Money	89.9	93.1	91.5
Percent	94.2	88.0	90.9
Total	86.2	88.5	87.4

Table 5: The results of MorphoBERT on provided dataset using 5-fold cross validation at the word-level.

Named Entity	P	R	F1
B-Person	93.9	92.9	93.4
I-Person	94.1	94.2	94.1
B-Organization	87.3	89.6	88.4
I-Organization	91.8	89.2	90.5
B-Location	91.0	91.7	91.4
I-Location	84.5	77.0	80.6
B-Date	82.8	84.5	83.6
I-Date	87.5	86.9	87.2
B-Time	77.0	79.9	78.4
I-Time	80.8	85.2	82.7
B-Money	94.2	96.3	95.2
I-Money	96.8	97.5	97.2
B-Percent	95.9	89.0	92.2
I-Percent	97.7	95.9	96.8
Total	90.5	89.8	90.2

Table 6: The results of MorphoBERT on the test dataset at the phrase-level. (In: in-domain, Out: out-of-domain)

	3-class Subtask			7-class Subtask		
	P	**R**	**F1**	**P**	**R**	**F1**
In	88.7	85.5	87.1	88.4	84.8	86.6
Out	86.3	83.8	85.0	86.0	83.1	84.5
Total	87.3	84.5	85.9	87.0	83.8	85.4

Table 7: The results of MorphoBERT on the test dataset at the word-level. (In: in-domain, Out: out-of-domain)

	3-class Subtask			7-class Subtask		
	P	**R**	**F1**	**P**	**R**	**F1**
In	92.5	86.7	89.5	94.0	89.1	91.5
Out	91.5	84.0	87.6	91.8	85.7	88.6
Total	92.1	85.2	88.5	92.8	87.1	89.9

Our system, MorphoBERT, gained the first rank among the participated teams in all evaluation measures, in both tasks, and in in-domain and out-of-domain data.

Table 6 and 7 present the results of our system at the phrase-level and word-level, respectively. Comparing the results of the system on the in-domain test data with results of the system on the provided dataset (Table 3) shows that the precision remains high but the recall decreases. This reveals that the coverage of texts in the in-domain part of the test dataset is slightly different from the provided dataset, though the domain is the same. In the out-of-domain data, the decrease in the precision is negligible. However, the recall declines more seriously, evidently because of the difference of named entities covered in different domains.

We do not have access to the gold labels of the dataset. However, in order to have a more comprehensive analysis, we present the detailed results of MorphoBERT on the test dataset reported by the organizers. As Table 8 shows, the most decrease happens in *organization* and it is more than 10%. This shows that the test dataset contains organizations which are not observed in the training dataset. Regarding this fact that more than half of the test dataset consists of out-of-domain text, one can concludes that they come mostly from out-of-domain texts. It is not surprising that if the domain changes, the text refers to different organization names. Other named entities such as *person*, *date* and *time* however experience less changes.

Table 8: The detailed results of MorphoBERT on the test dataset at the phrase-level.

Named Entity	**F1**
Person	90.4
Organization	80.3
Location	87.1
Date	78.9
Time	71.0
Money	93.6
Percent	96.8
Total	85.4

5 Conclusions

We participated in the Persian NER task of NSURL 2019 with our system called MorphoBERT. Our system achieved the first rank in all settings among the participated teams. The system benefited from the BERT model and a Persian morphological analyzer. The assessment on the test dataset was blind. The task had two subtasks, 3-class and 7-class subtasks, and the system was evaluated on the in-domain as well as out-of-domain data. On the in-domain test data the total performance of the system is comparable with the system trained on the provided dataset and it changes slightly. Differentiating between generic temporal expressions with specific ones was a big challenge for the system and as a result the system gained the lowest results in the *time* and *data* classes. Another reason for getting a lower performance in these two classes was the low number of instances in the training dataset. Utilizing a statistical or even a rule-based system might be helpful here. Results showed that on out-of-domain texts the recall of the NER system decreases more, especially in detecting *organization*. This gives us a hint to focus on this challenge for future work. It is also worth focusing on the morphological analyzer. Our current morphological analyzer is not highly accurate in low-frequent and unknown words. Developing a high precise Persian morphological analyzer can be beneficial for many tasks, especially if there are no enough resources available to train data-voracious neural systems.

References

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a persian written corpus: Peykare. *Language Resources and Evaluation*, 45(2):143–164.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC, USA. Association for Computational Linguistics.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. NYU: Description of the MENE named entity system as used in MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Geoffrey Leech and Andrew Wilson. 1999. Standards for tagsets. In *Syntactic wordclass tagging*, pages 55–80. Springer.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mahdi Mohseni, Javad Ghofrani, and Heshaam Faili. 2018. Persianp: A persian text processing toolbox. In *Computational Linguistics and Intelligent Text Processing*, pages 75–87, Cham. Springer International Publishing.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian named-entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Cícero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China. Association for Computational Linguistics.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2019. PEYMA: A tagged corpus for Persian named entities. *Journal of Signal and Data Processing, Vol. 16, Issue 1, 06-2019*.

Nasrin Taghizadeh, Zeinab Borhani-fard, Melika Golestani-Pour, Mojgan Farhoodi, Maryam Mahmoudi, Masoumeh Azimzadeh, and Heshaam Faili. 2019. NSURL-2019 task 7: Named entity recognition (ner) in farsi. In *Proceedings of the first International Workshop on NLP Solutions for Under Resourced Languages*, NSURL '19, Trento, Italy.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.