Beheshti-NER: Persian named entity recognition Using BERT

Ehsan Taher*
NLP Research Laboratory
Shahid Beheshti University
Tehran, Iran
e.taher@mail.sbu.ac.ir

Seyed Abbas Hoseini*
NLP Research Laboratory
Shahid Beheshti University
Tehran, Iran
seyeda.hoseini@mail.sbu.ac.ir

Mehrnoush Shamsfard
NLP Research Laboratory
Shahid Beheshti University
Tehran, Iran
m-shams@sbu.ac.ir

Abstract

Named entity recognition is a natural language processing task to recognize and extract spans of text associated with named entities and classify them in semantic Categories.

Google BERT is a deep bidirectional language model, pre-trained on large corpora that can be fine-tuned to solve many NLP tasks such as question answering, named entity recognition, part of speech tagging and etc. In this paper, we use the pre-trained deep bidirectional network, BERT, to make a model for named entity recognition in Persian.

We also compare the results of our model with the previous state of the art results achieved on Persian NER. Our evaluation metric is CONLL 2003 score in two levels of word and phrase. This model achieved second place in NSURL-2019 task 7 competition which associated with NER for the Persian language. our results in this competition are 83.5 and 88.4 f1 CONLL score respectively in phrase and word level evaluation.

1 Introduction

in this paper we trained our model which is participated in NSURL-2019 task 7 competition (Taghizadeh et al., 2019) which associated with NER for the Persian language.

Named Entity Recognition (NER) is one of the important and basic tasks in natural language processing, assigning different parts of a text to suitable named entity categories.

There are several sets of named entity (NE) categories introduced and used in different NE tagged corpora as their tagsets. For example, Peyma's (Shahshahani et al., 2018) tagset consists of person, organization, location, date, money, percent, and time, while the Arman tagset (Poostchi et al.) contains person, organization, location, facility, product, and event.

NER is one of the key parts of many downstream applications in NLP, such as question answering (Aliod et al., 2006), information retrieval (Guo et al., 2009), and machine translation (Babych and Hartley, 2003). As a result, the performance of NER can affect the quality of a variety of downstream applications. Furthermore, this effect is more obvious in low-resource languages because in these languages due to lack of data and tagged corpora, usually applications are implemented in pipe-line architecture unlike other languages like English which prefer to use End-to-End solutions.

Preparing basic tools in under-resourced languages by high performance can be a good solution to such languages while we counter with lack of data issue for training such tools.

We have trained conditional random field on the top of pre-trained bidirectional transformer BERT. Delvin et al. (Devlin et al., 2019) introduced BERT as a pre-trained Bidirectional Transformer model for language understanding tasks. BERT achieved state of the art results in many tasks like question answering, language inference, and Named entity recognition.(Devlin et al., 2019)

The need for large tagged data is the main problem with the recent supervised methods such as deep learning. Transfer learning can help this problem for under-resourced languages. Word embeddings approach (Mikolov et al., 2013),(Bojanowski et al., 2016), (Joulin et al., 2017) and (Peters et al., 2018) are the first kind of trans-

*Equal contribution.

fer learning solutions. We use word embeddings for supervised tasks after we trained them unsupervised on large raw corpora of texts. By this, they can reduce the need for huge labeled data. They defer by BERT usually in many aspects like the fine-tuning step. After pretraining BERT on large row corpora, we fine-tune it for our specific supervised problem. While BERT tokenizes text by itself, it extracts contextualized embeddings for each token. BERT is pre-trained on 104 languages like Persian, and this is one of the big advantages of this model. Vaswani et al. (Vaswani et al., 2017) introduced transformer architecture and self-attention as an alternative for encoder-decoder recurrent neural networks architecture which could achieve state of the art results in English to German and English to France machine translation problem. Furthermore, the speed for training transformers is much less than recurrent neural networks in encoder-decoder architecture. CRF as a probabilistic model like hidden Markov model makes it possible to extract and consider structural dependencies among tags in data. While Encoders like BERT try to maximize likelihood by selecting best hidden representation while CRF maximizes likelihood by selecting best output tags. We achieved 88.4% CONLL F1 score in word-level and 83.5% CONLL F1 score in phrase-level evaluation on Peyma dataset. We won second place in NSURL-2019 task 7 (Taghizadeh et al., 2019) competition for NER task.

In section 2, we talk about previous methods for NER and solutions like transfer learning to deal with under-resourced languages. Section 3 describes BERT. Section 4 explains our model in more details, discussing the training and test phases. In section 5, we show the achieved results on experiments like evaluating our model on different datasets. Section 6 concludes the paper.

2 Related Work

This paper describes a deep learning method based on word embedding and transfer learning, for named entity recognition in Persian language. Thus, in this section we first discuss some related work on Persian NER, then some recent work on English NER, and then talk about some word embedding models which can be used in NER tasks via transfer learning.

Mortazavi and Shamsfard (Mortazavi and Shamsfard, 2009) used a rule-based system to extract named entities for Persian languages. It was one of the first implementations for NER in Persian while no datasets existed in that time for evaluation. Poostchi et al. (Poostchi et al.) introduced new annotated Persian named-entity recognition dataset named Arman. Arman contains 250,015 tokens and 7,682 sentences. Set of entity categories consists of person, organization (like banks and ministries), location, facility, product, and event. They also trained conditional random field with bidirectional LSTM on this dataset as a baseline. Shahshahani et al. (Shahshahani et al., 2018) introduced new annotated Persian named-entity recognition dataset called Peyma. Peyma contains 7,145 sentences, 302,530 tokens and 41,148 tokens with entity tags collected from 709 documents. Class distribution for both Peyma and Arman datasets are presented respectively in Fig.1 and Fig.2.

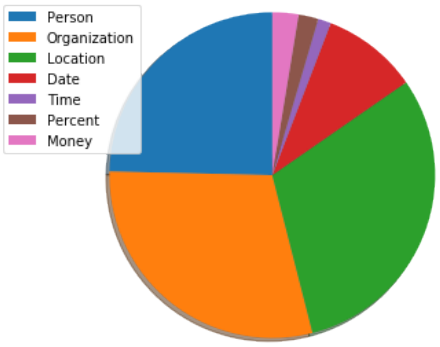

Figure 1: distribution of different classes in Peyma dataset

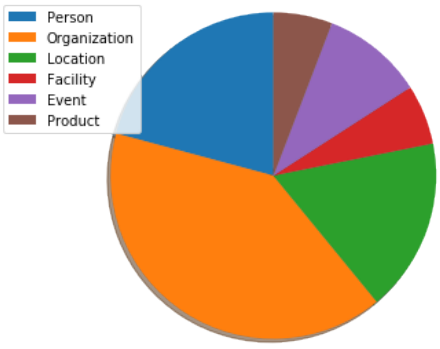

Figure 2: distribution of different classes in Arman dataset

Bokaei and Mahmoudi (Bokaei and Mahmoudi, 2018) trained recurrent and convolutional neural networks with CRF on Arman dataset.

Baevski et al. (Baevski et al., 2019) used a novel method for training bidirectional transformer which could over perform previous work and achieved state of the art result in English NER.

Akbik et al.(Akbik et al., 2018) used contextualized word embeddings extracted from character-level language model to solve the NER problem.

Delvin et al. (Devlin et al., 2019) introduced BERT as a pre-trained bidirectional transformer. They used and evaluated BERT on many tasks, including NER.

Using unsupervised methods can be a promising way because the most important issue for low resource languages is the lack of labeled data while but the access to a large amount of raw texts is more probable and feasible. Today, word embeddings such as Glove (Pennington et al., 2014), word2vec (Mikolov et al., 2013) , and fastText (Joulin et al., 2017) are essential parts of many methods in NLP. These models give continuous representations in n-dimensional space for each word which contain semantic information and features about that word.

Elmo (Peters et al., 2018) introduced deep contextualized word embedding by considering the context of words. Which means words have different embedding in different contexts. Delvin et al. (Devlin et al., 2019) and Radford et al.(Radford et al.) proposed a new method with transfer self-attention blocks without the need to change in architecture for a specific problem. They suggest fine-tuning pre-trained bidirectional transformers for specific problems.

Radford et al. (Radford et al.) introduced a new language model called GPT.2, which could reach 55% F1 score on the CoQA dataset without any labeled data. This approach tries to remove the need for labeled data and gives a general model to solve problems against BERT, which tries just to give a general model.

best performing models before us for NER in Persian are LSTM based models which usually come with CRF and pre-trained non-contextualized embedding layers. these models are evaluated on two common datasets for NER: PEYMA and ARMAN. Bokaei and Mahmoudi (Bokaei and Mahmoudi, 2018) and Shahshahani et al. (Shahshahani et al., 2018) had reported the best results which you can see in Table 3

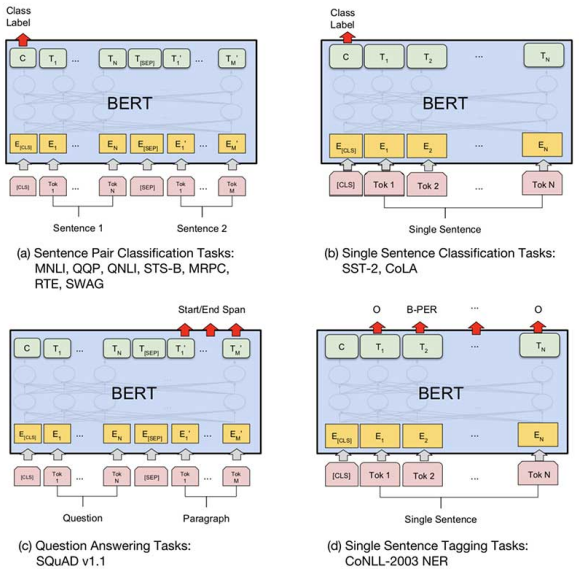

Figure 3: fine-tuning BERT in different tasks (Devlin et al., 2019)

3 BERT

BERT (Bidirectional Encoder Representations from Transformer) is a language model representation based on self-attention blocks. BERT is pre-trained in different language model tasks on raw unlabeled texts. The pre-trained deep bidirectional model with one output layer can reach state-of-the-art results in many tasks such as question answering and Multi-Genre Natural Language Inference. The idea is to have a general architecture which fits many problems and a pre-trained model which minimize the need for labeled data. For example, in Fig. 3 You see how BERT can be used in different tasks like question answering, sentences pair classification, single sentence classification, and single sentence tagging task. While each task has a different format of inputs and outputs. As mentioned before, one of the big advantages of BERT is that it was trained in 104 languages and Persian is one of those. Which motivate us to use it for NER in Persian.

4 Our Proposed Model

In this paper, we propose a method for Persian NER. In this method, we use BERT pre-trained model. As in NER task, we need to assign the most suitable tag to each token, and suitable tokenization is an important step.

While BERT has its tokenization with Byte-Pair encoding and it will assign tags to its extracted tokens, we should take care of this issue. BERT

extracted tokens are always equal to or smaller than our main tokens (that taken from the Step-1 Shamsfard et al., 2010) because BERT takes tokens of our dataset one by one. As a result, we will have intra-tokens which take X tag (meaning don't mention). We trained a conditional random field and fully connected layer after output representation of tokens extracted by BERT. It Is a fine-tuning step to make the entire model ready for NER task. You can see a schema of the model in Fig.4 .

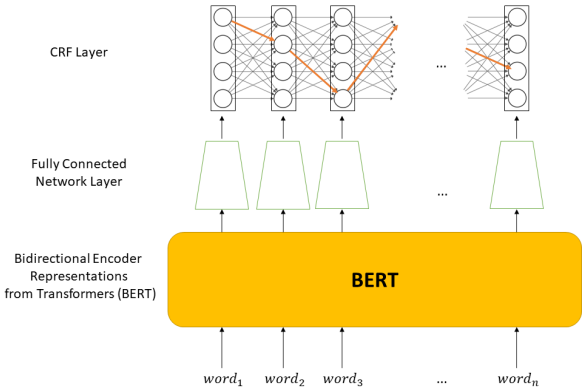

Figure 4: architecture of our trained model

5 Experiments

We have trained and tested our model on two different datasets: Peyma (Shahshahani et al., 2018) and Arman (Poostchi et al.).We split PEYMA dataset into 5 equal subsets (Peyma contains 7145 sentences thus each subset contains 1429 sentences) and use 5-fold cross-validation. we repeated training phase 5 times separately, Each time, one of the 5 subsets is used as the test set and the remaining 4 subsets are put together to form a training set. In all experiments, CONLL F1 score is calculated in two levels: word and phrase as a metric for evaluating the performance of model. Results of our model on Peyma and Arman datasets are given respectively in Table 1 And Table 2.

On Peyma dataset We can reach 90.59% CONLL F1 score in phrase-level and 87.62% F1 score in word level. Best results are seen for Percent class and worst for Time.

On Arman dataset, We reached 79.93% CONLL F1 score in phrase-level and 84.03% F1 score on word-level. Best results are seen for Person class and worst for Event

One of the causes for achieving different results in each class is the amount of named entities in the datasets. As can be seen in Fig.1 and Fig.2, the number of phrases for Time and Event classes are much lower than others.

As you see in Table 3 in both word and phrase levels, our model outperform other NER approaches for the Persian language. Unfortunately previous works reported their results just on one of the datasets. Shahshahani et al.(Shahshahani et al., 2018) reported their results just in word level evaluation on Peyma dataset.Table 3 shows that our results are 10 percent better than Shahshahani and colleagues on the same platform. On the other hand Bokaei and Mahmoudi (Bokaei and Mahmoudi, 2018) reported their results on Arman dataset Which is lower than ours in both word and phrase levels according to Table 3.

	Arman		Peyma	
	word	phrase	word	phrase
Bokaei and Mahmoudi (Bokaei and Mahmoudi, 2018)	81.50	76.79	-	-
Shahshahani et al.(Shahshahani et al., 2018)	-	-	80.0	-
Beheshti-NER (Our Model)	**84.03**	**79.93**	**90.59**	**87.62**

Table 3: comparing results of our trained model with others

The results of NSURL task-7 competition is reported in two levels of evaluation, namely word and phrase levels for two subtasks: A) NER for 3- classes (Person, Location, Organization) and B) NER for all classes. for the competition, we have trained our model on PEYMA corpus in addition to another corpus which was prepared by Iran Telecommunication Research Center (ITRC). The organizers also used two kinds of in-domain and out-domain test data. Our model won second place in all of these evaluation types.

Tables 4, 5, 6, 7 and 8 show the results of evaluation reported by competition for all teams which participated in the challenge. Our method is mentioned as Beheshti-NER-1[1]. Table 4 and 5 show the results for subtask A. according to the tables, we reached to 84.0% and 87.9% F1 score respectively for phrase and word level evaluations.

[1]Code is available at https://github.com/sEhsanTaher/Beheshti-NER

	Date		Location		Money		Organization		Percent		Person		Time		
	B-	I-	B-	I-	B-	I-	B-	I-	B-	I-	B-	I-	B-	I-	all classes
word-f1	84.83	88.44	91.60	82.39	95.78	97.59	89.07	90.29	94.97	97.13	93.17	94.25	83.50	86.48	90.59
phrase-f1	80.33		89.75		92.54		84.80		93.57		90.69		73.78		87.62

Table 1: results of our model on Peyma dataset. Two kinds of evaluation is used, namely word and phrase level. in word level evaluation B- assigns to first token of phrase and I- is for middle and last tokens.

	Event		Faculty		Location		Organization		Person		Product		
	B-	I-	B-	I-	B-	I-	B-	I-	B-	I-	B-	I-	all classes
word-f1	72.39	78.58	76.49	78.77	82.53	78.96	81.12	87.51	92.81	94.83	68.56	71.34	84.03
phrase-f1	58.45		69.53		80.73		78.01		91.46		62.97		79.93

Table 2: results of our model on Arman dataset.

	Team	Test Data 1								
		In Domain			Out Domain			Total		
		P	R	F1	P	R	F1	P	R	F1
1	MorphoBERT	88.7	85.5	87.1	86.3	83.8	85	87.3	84.5	85.9
2	Beheshti-NER-1	85.3	84.4	84.8	84.4	82.6	83.5	84.8	83.3	84
3	Team-3	87.4	77.2	82	87.4	73.4	79.8	87.4	75	80.7
4	ICTRC-NLPGroup	87.5	76	81.3	86.2	69.6	77	86.8	72.3	78.9
5	UT-NLP-IR	75.3	68.9	72	72.3	60.7	66	73.6	64.1	68.5
6	SpeechTrans	41.5	39.5	40.5	43.1	38.7	40.8	42.4	39	40.6
7	Baseline	32.2	45.8	37.8	32.8	39.1	35.7	32.5	41.9	36.6

Table 4: Phrase-level evaluation for subtask A: 3-classes

	Team	Test Data 1								
		In Domain			Out Domain			Total		
		P	R	F1	P	R	F1	P	R	F1
1	MorphoBERT	92.5	86.7	89.5	91.5	84	87.6	92.1	85.2	88.5
2	Beheshti-NER-1	90.5	87.2	88.8	89.7	85	87.3	90.1	85.8	87.9
3	Team-3	89.2	79.5	84.1	89.5	74.7	81.4	89.3	76.9	82.7
4	ICTRC-NLPGroup	90.1	78.2	83.7	88.7	70.2	78.4	89.4	73.5	80.7
5	UT-NLP-IR	87.3	71.9	78.9	86.4	61.1	71.6	86.9	65.7	74.8
6	SpeechTrans	66.8	38.3	48.7	66.2	35.2	46	66.6	36.4	47
7	Baseline	46.2	42.6	44.3	45.2	35.1	39.5	45.9	38.4	41.8

Table 5: Word-level evaluation for subtask A: 3-classes

results for subtask B is given in Table 6 and 7. we can achieve 83.5% and 88.4% F1 score respectively for phrase and word level evaluation.

	Team	Test Data 1								
		In Domain			Out Domain			Total		
		P	R	F1	P	R	F1	P	R	F1
1	MorphoBERT	88.4	84.8	86.6	86	83.1	84.5	87	83.8	85.4
2	Beheshti-NER-1	84.8	83.6	84.2	83.9	82	83	84.3	82.7	83.5
3	Team-3	87.4	77.3	82	87.3	72.8	79.4	87.3	74.7	80.5
4	ICTRC-NLPGroup	87	76.1	81.2	86.2	70.2	77.4	86.5	72.7	79
5	UT-NLP-IR	77.3	70.2	73.6	74.1	61.9	67.5	75.5	65.4	70.1
6	SpeechTrans	38	34.5	36.2	38.9	33.6	36	38.5	34	36.1
7	Baseline	32.8	45.7	38.2	32	38.1	34.8	32.4	41.3	36.3

Table 6: Phrase-level evaluation for subtask B: 7-classes

	Team	Test Data 1								
		In Domain			Out Domain			Total		
		P	R	F1	P	R	F1	P	R	F1
1	MorphoBERT	94	89.1	91.5	91.8	85.7	88.6	92.8	87.1	89.9
2	Beheshti-NER-1	91.4	87.3	89.3	89.7	85.7	87.7	90.4	86.5	88.4
3	Team-3	91.3	84.1	87.5	90.9	77.9	83.9	91.1	80.7	85.5
4	ICTRC-NLPGroup	89.2	83.1	86.1	89.8	76.5	82.6	89.7	79.4	84.2
5	UT-NLP-IR	92.7	79.3	85.4	91.1	68.4	78.1	91.9	73.1	81.4
6	SpeechTrans	76.1	32.9	45.9	74.9	30.3	43.2	75.7	31.5	44.5
7	Baseline	50.6	47.8	49.2	42.6	35.1	38.5	46.5	40.9	43.5

Table 7: Word-level evaluation for subtask B: 7-classes

details of evaluation for each class in subtask B is given in Table 7. as you see all teams have higher scores in Percent class and the worst score for many teams is for Time class.

	Team	Test Data 1							
		PER	ORG	LOC	DAT	TIM	MON	PCT	Total F1
1	MorphoBERT	90.4	80.3	87.1	78.9	71	93.6	96.8	85.4
2	Beheshti-NER-1	81.8	80.8	88	77.8	75.8	85.1	91.6	83.5
3	Team-3	79.9	77.2	83.9	74.7	64.3	92.1	97.4	80.5
4	ICTRC-NLPGroup	76.2	75.93	82.8	76	67.1	91.3	93.6	79
5	UT-NLP-IR	63.4	58.8	78.2	76.1	69.1	84.5	93.5	70.1
6	SpeechTrans	24.3	23.5	63.1	12	4.1	0.3	0.7	36.1
7	Baseline	23.5	38.1	44.2	41.6	30.3	13.7	36.6	36.3

Table 8: Details of phrase-level evaluation for subtask B: 7-classes

6 Conclucion

in this work we fine-tuned the pre-trained BERT model with a CRF layer in NER task for Persian language. our trained model achieved best results compared to the previous ones and ranked as the second team in NSURL competition. this work present BERT as a good transfer learning solution for solving low resource problems.

results show that our model could outperform previous methods with a dramatic difference. the reason for this could be using a big pre-trained model, BERT, which achieved state of the art results in English and proved to perform well with a less amount of data for training.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Diego Mollá Aliod, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *ALTA*.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven Pretraining of Self-attention Networks. *arXiv:1903.07785 [cs]*. ArXiv: 1903.07785.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*. ArXiv: 1607.04606.

Mohammad Hadi Bokaei and Maryam Mahmoudi. 2018. Improved Deep Persian Named Entity Recognition. In *2018 9th International Symposium on Telecommunications (IST)*, pages 381–386, Tehran, Iran. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 267–274, New York, NY, USA. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

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*.

P. S. Mortazavi and M. Shamsfard. 2009. Named entity recognition in persian texts. 15th National CSI Computer Conference.

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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. ELMO: Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. BiLSTM-CRF for Persian Named-Entity Recognition. page 5.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2018. PEYMA: A Tagged Corpus for Persian Named Entities. *arXiv:1801.09936 [cs]*. ArXiv: 1801.09936.

Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010. STeP-1: A set of fundamental tools for Persian text processing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.