Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 3, 2022, 373-383.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A Transformer-based Approach for Persian Text Chunking

Parsa Kavehzadeh, Mohammad Mahdi Abdollah Pour, and Saeedeh Momtazi*

Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran.

Article Info

Article History:
Received 27 July 2021
Revised 30 October 2021
Accepted 20 February 2022

DOI:10.22044/jadm.2022.11035.2250

Keywords:
Persian Text Chunking, Sequence Labeling, Deep Learning, Contextualized Word Representation.

**Corresponding author: momtazi@aut.ac.ir (S. Momtazi).*

Abstract

Over the last few years, text chunking has taken a significant part in the sequence labeling tasks. Although a large variety of methods have been proposed for shallow parsing in English, most of the proposed approaches for text chunking in the Persian language are based on the simple and traditional concepts. In this paper, we propose using the state-of-the-art transformer-based contextualized models, namely BERT and XLM-RoBERTa, as the major structure of our models. Conditional random field (CRF), a combination of bidirectional long short-term memory (BiLSTM) and CRF, and a simple dense layer are employed after the transformer-based models in order to enhance the model's performance in predicting the chunk labels. Moreover, we provide a new dataset for noun phrase chunking in Persian, which includes the annotated data of Persian news text. Our experiments reveal that XLM-RoBERTa achieves the best performance between all the architectures tried on the proposed dataset. The obtained results also show that using a single CRF layer would yield better results than a dense layer, and even the combination of BiLSTM and CRF.

1. Introduction

Text chunking or shallow parsing is one of the most contextual tasks that has been done in the recent years in the natural language processing (NLP) communities. The major importance of chunking is to provide a smart way of finding meaningful and comprehensive structures in text, which could be very helpful in other NLP areas such as information extraction [35], sentiment analysis [48], and plagiarism detection [40].

Considering that in most cases phrases follow specific patterns, the rule-based methods have been proposed for the task. This approach, however, results in numerous rules and regular expressions for detecting different types of phrases in the text. Moreover, due to the sophisticated linguistic characteristics of phrases in most languages, detecting them just by employing rudimentary rule-based methods is roughly impossible. Consequently, in order to achieve acceptable results, various machine learning methods, including Recurrent Neural Networks (RNNs) and probabilistic models, have been used to handle text chunking. These methods have also been improved using the state-of-the-art text representation models.

Considering the characteristics of Persian as one of the languages with diverse and complicated patterns for different kinds of phrases, we aim to use the state-of-the-art transformer-based models in combination with CRF and BiLSTM-CRF as sequence labeling layers in order to solve the chunking problem. Moreover, having a sufficient and validated dataset is always a critical concern for everyone willing to struggle with a NLP task. Based on our current knowledge, there are a limited number of datasets for text chunking in Persian, which are not publicly available for research in this field. This motivated us to gather a new comprehensive dataset with chunk tags by crawling the Persian news websites and labeling the text semi-automatically for our experiments.

The main contributions of the paper are as follows:

- Providing a new dataset for Persian text chunking semi-automatically
- Using state-of-the-art transformer-based contextualized representation for text chunking
- Comparing different neural architectures beside transformers for the target task.

In the following, we will first point to the related works about shallow parsing for both the Persian and English languages. In Section 3, we will introduce our dataset. Our models and experiments will be elaborated in Sections 4 and 5. Finally, Section 6 summarizes the paper.

2. Related Works

Many novel approaches have been used on English text chunking in the recent years. The contributions of the proposed models in the field are mainly categorized in the text embedding, sequence tagging or learning process. In this section, we first review the state-of-the-art models on text chunking from a general viewpoint, and then focus on Persian text chunking, which is the target of this study. The pre-trained word embeddings like SENNA [7] and GLOVE[1] have been used by Huang, Xu, and Yu [21], L. Liu *et al*. [25], Sogaard and Goldberg [45], and Zhai, Potdar, Xiang, and Zhou [52] in order to enhance the performance of the models in text chunking. The character-level language models have been employed by Akbik, Blythe, and Vollgraf [1], L. Liu *et al*. [25], and byte-level embeddings have been used to improve the accuracy of sequence labeling.

Moreover, for the sequence labeling unit, the traditional probabilistic models like Hidden Markov Model (HMM) [10] and CRF [23] have been widely used with various kinds of RNNs [39] including LSTM [18], BiLSTM [14], and Gated Recurrent Unit (GRU) [5]. Their ability to keep the information during passing through a sentence assists the model to maintain important information in the text in order to detect specific segments. For instance, the BiLSTM-CRF model has been used by Huang *et al*. [21] to handle the sequence labeling tasks including chunking.

Ma *et al*. [27] have used convolutional layers in addition to BiLSTM for sequence chunking. They also leveraged the information of other sequential features like Named Entity Recognition (NER) and Parts Of Speech (POS) tags in order to create embeddings for each word before feeding input to the network.

The learning unit of chunking has been further expanded by other advanced methods including multi-task learning, which provides the possibility of training a model for different tasks [17, 25, 45].

Semi-supervised learning is another well-known technique that has been used in chunking. In the models proposed by Rei [38] and Clark, Luong, Manning, and Le [6], both the labeled and unlabeled data was used for sequence chunking.

Peters, Ammar, Bhagavatula, and Power [33] have devised a bidirectional architecture, TagLM, to be a language model for following the sequence labeling tasks such as text chunking. Deep transition RNN has been another innovative approach proposed by Y. Liu *et al*. [26] to enhance performance on sequence chunking.

Most suggested methods for text chunking, especially for the Persian language, used models like HMM, genetic algorithm, simple multi-layer neural network, LSTM, and BiLSTM. On the other hand, with the advent of attention-based transformers [49] and novel transformer-based architectures proposed by Delvin, Chang, Lee, and Toutanova [9] and Lample and Conneau [24], notable advances have occurred during the recent years in different NLP tasks such as text generation, NER, and text chunking. Lots of these pre-trained transformer-based models have been trained on multi-lingual datasets, making them ideal to be employed in works related to rare languages [47] or even machine translation tasks [51]. In this article, we will show how we use these models for Persian text chunking.

2.1. Persian chunking

As mentioned earlier, the rule-based methods have been used for chunking, especially in the low-resource languages that suffer from lack of annotated data.

In the proposed model of Mohtaj, Roshanfekr, Zafarian, and Asghari [30], a simple rule-based approach has been used to handle shallow parsing. They defined certain rules to extract Persian phrases from 100 randomly selected sentences from a large corpus.

In another research study, a rule-based approach has been used to create the dataset to be used for

[1] (http://nlp.stanford.edu/projects/glove/)

the task. We also follow this idea to create the dataset, which will be described in Section 4.

Noferesti and Shamsfard [31] have used a rule-based method to create a tagged dataset for the chunking task. They also employed the genetic algorithm to learn and predict phrases. Each gene takes one of the IOB tags as its value.

Kiani, Akhavan, and Shamsfard [22] have proposed a hybrid approach for chunking Persian sentences. First, they employed a rule-based method in order to create labeled data for the chunking task. Afterwards, they fed the labeled data including POS tags of the previous and next tokens as the features to a multi-layer neural network, and used the Fuzzy C-Means clustering algorithm to predict the labels. Shamsfard and Mousavi [41] have designed a rule-based shallow parser in order to extract semantic relations in a sentence. Shamsfard and SadrMousavi [42] have also used a rule-based approach for the semantic role labeling task in the Persian language.

SharifiAtshgah [43] has handled noun phrase segmentation by defining certain regular expressions in order to provide a Persian Treebank. 3452 noun phrases were segmented with an error rate of 7% in their work. Homayoonpour and Salimibadr [19] have used Support Vector Machines (SVMs) and CRF to segment a tiny part of Persian text data. A small corpus including 589 sentences with roughly 6000 words in total was provided as the training set, and they manually generated the test set. In order to predict the boundaries of syntactic groups, they used the target word and its grammatical label, the grammatical roles of two words before and after the target token, two sequences of the target word with the previous and forwarding neighbors, and the main grammatical roles of the words around the target. Ghayoomi [13] has provided the first constituency treebank for Persian based on the head-driven phrase structure grammar. He used regular expressions in the ClaRK [44] system with a bootstrapping approach. Rasooli, Kouhestani, and Moloodi [36] have also introduced a treebank containing 30,000 sentences annotated by syntactic roles and relations. Tabatabayi and HoseinNezhad [46] have proposed a CRF-based model for Persian chunking. For training their model, they extracted Persian chunks from the Persian dependency treebank [36] using regular expressions. Kiani and Shamsfard (1387) have used a neural model for shallow parsing. They used the dataset provided by [41, 42] for training a Multi-Layer Perceptron (MLP) model.

Like Kiani *et al*. [22], Mohseni, Ghofrani, and Faili [29] have used the POS tags of the previous and next works of each particular token as the features of each input token. They used these features plus IOB chunk labels for learning the classifier introduced by Manning and Klein [28] in order to detect noun phrases in the text.

Hosseinnejad, Shekofteh, and Emami Azadi [20] have introduced a text corpus, *A'laam*, in which annotated Persian text data with different noun phrase labels have been provided. Their corpus contains 13 different noun phrase labels over 250,000 words. In order to evaluate a model on their dataset, they trained a CRF-based NER system in the Persian language on their annotated data. The model showed satisfying results on predicting the labels of different noun phrases in corpus. Asgari-Bidhendi [52] also introduced a corpus for NER in Persian language, containing over 200,000 tokens crawled from social media.

Memory-based learning is another method used for shallow parsing tasks in Persian. Ghalibaf, Rahati, and Estaji [12] have extracted features related to the POS tags of previous and next tokens of each word and fed these features to the MBL algorithm to detect the phrases.

3. Models

In this section, we introduce our models used for Persian text chunking. In the recent years, the models containing the CRF structure have been noticeably employed in order to handle different sequence labeling tasks such as chunking in the English language. More common RNNs like LSTM, GRU, and BiLSTM have also been combined with CRF in order to improve the performance by capturing contextual information from the input sequences. However, by introducing the contextualized transformer-based models [49], the state-of-the-art works and ideas have been proposed in the last couple of years for different NLP tasks including sequence labeling and text chunking. We aim to employ these pre-trained transformer-based models to handle text chunking in the Persian language. As we must use the models that support the non-English languages, we used the ParsBERT, multilingualBERT (mBERT), and XLM-RoBERTa models, which were trained on multiple languages, as our candidates. A large version of each architecture was used to create an equivalent situation for the models. For each model, we tried three different scenarios for the last layer of the model. We used CRF, BiLSTM-

CRF, and a simple dense layer beside those aforementioned pre-trained models in order to evaluate the performance of different architectures. First, we will introduce the transformer-based models we use and their characteristics. Then we will elaborate the CRF and BiLSTM-CRF layers used in our models in the next following part.

3.1. Text representation with BERT family

Delvin *et al*. [9] have firstly introduced a bidirectional transformer-based model, called BERT. BERT combines the three different vocabulary, segment, and position embeddings to create the input embeddings before feeding input to the deep architecture. Two major objectives are considered for the pre-training phase of BERT. The first one is masked language modeling, in which the model should consider the context of unmasked tokens in a sequence to predict the masked word. Another pre-training target for BERT is next sentence prediction. In this task, the model is forced to determine whether two sequential sentences are related to each other or not. The pre-trained bidirectional transformer-based architecture of BERT has been used in various NLP tasks, including question answering [34], sentiment analysis [50], and sequence labeling [26]. Furthermore, there are multilingual and Persian versions of BERT trained on the text data from different languages. We use mBERT[2] and ParsBERT for our text chunking task on the Persian language.

Considering the shortcomings of mBERT in the non-English languages, the XLM model has been proposed by Lample and Conneau [24]. The objectives of the XLM model are masked language modeling, which forces the model to detect masked tokens in a sentence, next token prediction, requiring the model to predict the next token by giving it the previous ones, and translation language modeling in which the model tries to predict masked tokens by considering the unmasked words in the sentences from different languages.

Later, XLM-RoBERTa was introduced by Conneau *et al*. [8]. They employed the previous XLM model [24] to be trained on a huge amount of data from 100 different languages. In order to provide such a big amount of data, they created a dataset called CommonCrawl entailing two terabytes of textual content from 100 different languages. The XLM-RoBERTa's pre-training goal is masked language modeling and predicting masked tokens in a sequence. This multilingual bidirectional transformer-based model is another architecture we use for Persian text chunking.

3.2. Sequence labeling with CRF and BiLSTM-CRF

A combination of CRF with various types of RNNs such as LSTM and BiLSTM has been used in different works on text chunking. The CRF layer helps the model to consider the labels of other tokens in a sentence, while predicting the label of a particular word. In this way, the model learns the correlations between different labels alongside the probability of the labels in a specific position. CRF employs two major matrices to handle this. T is a $l \times l$ matrix that is the probabilities of transitions between the existing labels (l is the number of labels), and P is a $n \times n$ matrix showing the probability of each label in each position in a sequence (n is the length of a sequence). Equation 1 shows the calculation of the score for an input sequence $x = \{x_1, \ldots, x_n\}$ and a label sequence $y = \{y_1, \ldots, y_n\}$. The objective for detecting the best label sequence is presented in Equation 2. Y represents all the possible label sequences.

$$s(x, y) = \sum_{i=0}^{n} T_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$$

$$E = -s(y) + log \sum_{\tilde{y} \in Y} e^{s(\tilde{y})}$$

RNNs have also been used in combination with CRF in order to assist the model to consider the context of sentences and tokens before feeding them to CRF. BiLSTM is one of the most common types of RNNs harnessed in different NLP tasks. The architecture helps the model to carry the important information of tokens without the problem of vanishing gradient resulting from passing through a long sentence. The bidirectional architecture of BiLSTM assists the model to capture the information from both the left and right contexts of each token, which causes a significant enhancement in the performance of the model in sequence labeling tasks including chunking. We use both the CRF and BiLSTM-CRF architectures in order to observe the effect of these models on the performance of the contextualized pre-trained models in Persian text chunking.

[2] https://github.com/google-research/bert/blob/master/multilingual.md

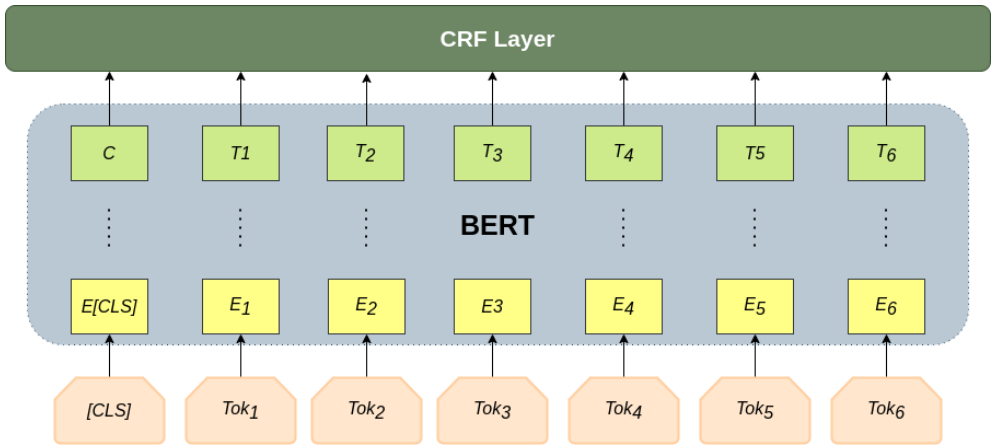

Figure 1. Feeding input tokens to the input stage of BERT to achieve input embeddings. After processing the input embeddings by transformer-based architecture, the output of BERT is passed through a CRF layer in order to obtain the final labels.

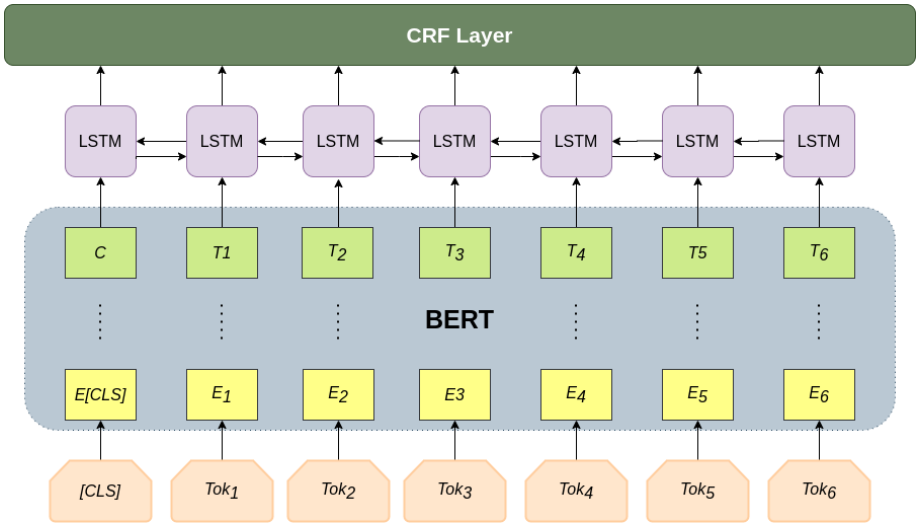

Figure 2. After Feeding input tokens to BERT, outputs of the contextualized model are passed to a bidirectional LSTM layer and the following CRF layer.

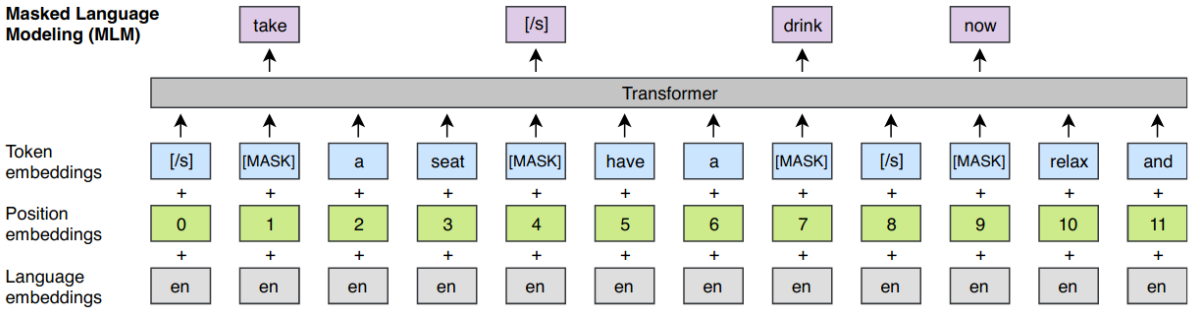

Figure 3. Procedure of masked language modeling objective in pre-training phase of the XLM-RoBERTa model.

In Figures 1 and 2, the BERT-CRF and BERT-BiLSTM-CRF models are shown. In BERT-CRF, the outputs of BERT are directly fed to a CRF layer to predict the labels. In BERT-BiLSTM-CRF, the outputs of the transformer-based model are first fed to a BiLSTM architecture, and then the outputs of BiLSTM are fed to the CRF layer. Figure 3 presents the masked language modeling objective of XLM-RoBERTa during pre-training. This model is used as the BERT model within the architectures shown in Figures 1 and 2 with CRF and BiLSTM-CRF, respectively. In the next

section, we will elaborate on the performance of the two transformer-based models in combination with CRF, BiLSTM-CRF, as well as a simple dense layer in chunking.

4. Dataset

In order to handle Persian text chunking, we require a reliable dataset for our experiments. There were some works in other related papers that sampled some sentences from large Persian corpora like Dadegan [36], Hamshahri [3], and Bijankhan [4] for text chunking; however, there is no specific dataset for Persian text chunking yet. The lack of available datasets for Persian text chunking motivated us to create a dataset based on real Persian text existing on the Web, which is part of the contribution of this research work. The news' text gathered by crawling different Persian news websites was used to create the dataset. We applied a POS tagger on the unprocessed text to obtain POS labels, which could be very useful in text chunking. Since the phrases follow certain patterns in text, designing certain rules on POS tagged data has been used in some works [15, 30, 32] in order to detect simple phrases in the text. However, since detecting noun phrases in Persian is challenging in most cases, at the final stage, a language expert should manually check the labels that the rule-based algorithm has tagged. Our dataset also follows the IOB scheme [37] containing three major tags "O", out of phrases, "B-NP", beginning of noun phrases, and "I-NP", inside of the noun phrases. Other details of our dataset are shown in Table 1. In order to be more precise, the following steps summarize the procedure of generating the dataset:

- Crawling the pages of the Persian news agencies and converting them to raw Persian text
- Applying a POS tagger on the raw data to achieve the POS labeled text
- Designing general rules on POS tags to predict simple noun phrases in the text. Among the several rules that we tried, the one below was the best in accuracy to detect phrases:

$$(T?\,Z?\,((U|A)+J?)*(N|O)+(J?\,(U|A)+)*)+Z?$$

- The meaning of the tags used in the above regular expression is shown in Table 2.
- Checking the automatically tagged text by a language expert to handle exceptions and complicated phrases.

Some examples of our dataset are as follows:

(1) ?u sepas goft tedăde besiăr kami dăstăn dărim
he then said number very little story have
B-NP O O B-NP I-NP I-NP I-NP O
'He then said that they have a little number of stories'

(2) in avalin vazife har fardi ast
this first duty every person is
B-NP B-NP I-NP I-NP I-NP O
'This is every person's first duty'

(3) šerkathăye dolati băyad sahăme ǩod ră beforošand
firms governmental should stock themselves - sell
B-NP I-NP O B-NP I-NP O O
'Governmental firms should sell their stocks'

(4) tağirāte nerǩe arz āmele aslie navasāne ğeimate talā ast
changes rate currency cause major fluctuations cost gold is
B-NP I-NP I-NP B-NP I-NP I-NP I-NP I-NP O
'The changes in currency rate is the major cause for the fluctuations in cost of gold'

Table 1. Dataset statistics

Measure	Number
# of sentences	3091
# of tokens	93058
# of labels	3
# of NP tokens	56888
# of NP phrases	17711
Avg. length of sentences	30.10

One of the most challenging parts of Persian noun phrase chunking is detecting the phrases that contain various clauses. Adjective clauses and different types of clauses are very common subsets of bigger noun phrases in the Persian

language, making the prediction of noun phrases difficult. For instance, a challenging example of our dataset is presented in Example 5. In this example, the whole text is a single noun phrase that should be predicted by the model. This indicates that the task requires an advanced architecture to recognize such structures in the text.

Table 2. POS tags in dataset.

POS tag	Meaning
A	Adjective
B	Abbreviation
D	Adverb
E	Preposition
I	Interjection
J	Conjunction
L	Classifier
N	Noun
O	Punctuation
P	Post-position
S	Residual
T	Determiner
U	Number
V	Verb
Z	Pronoun

(5) săzmănhăyi ke darǩăste pardăǩte in tashilăt be karkonăne ǩod ră dărand ...
organizations that request payment these facilities to employers their - have ...
'The organizations that want to endow these facilities to their employers ...'

As mentioned earlier, we only focused on the noun phrases, and left the other phrases for a future work. The main reason for focusing on the noun phrases is that detecting noun phrases is widely used in various NLP tasks such as keyword extraction and query analysis in search engines. In most cases, queries consist of noun phrases, and finding the dependencies between the words within a noun phrase helps to enhance the performance of the engine by extracting queries' target and intent. For instance, consider the query in Example 6.

(6) Query:

nazarăte namăyandegăne majles darbăre hamegirie viruse coronă dar šahrhăye roostăyi
opinions representatives parliament about pandemic virus corona in cities rural
'The opinions of parliament representatives about coronavirus pandemic in rural cities'

The above example could be a common query searched through the search engines. Obviously, there are three noun phrases in this long query:

(7) Phrase 1:

nazarăte namăyandegăne majles
opinions representatives parliament
'The opinions of parliament representatives'

(8) Phrase 2:

hamegirie viruse coronă
pandemic virus corona
'Coronavirus pandemic'

(9) Phrase 3:

šahrhăye roostăyi
cities rural
'Rural cities'

These phrases contain the main purpose of the original query, which means that detecting noun phrases in this query would split it into smaller parts without losing much information. In other words, if the search engine can match one of the above noun phrases of the query with a document, we can say that there is a relation between the document and the input query; however, if a sub-part of the noun phrase (e.g. the word "representatives") matches the document, there is no guarantee to find a reasonable relation between the query and the document.

5. Experiments and Results

5.1. Experimental setup

First, in order to determine our test and validation sets, we randomly selected 10% of our data for the

test and 10% for the validation set. The remaining 80% is left for training. In order to implement the fine-tuning of the contextualized models as well as CRF and BiLSTM-CRF, we used PyTorch[3] and transformers package[4] in Python. We used two different versions of BERT, ParsBERT [11], and multilingual-base[5] with 168 million parameters. We also employed XLM-RoBERTa-large with 550 million parameters. In this way, a specific TokenClassification module is already implemented for these pre-trained models. We used the outputs of the TokenClassification module in order to feed them into the CRF and BiLSTM-CRF modules. In addition, the sklearn-crf[6] model was used to implement the simple CRF model as our baseline. We set the learning rate to 1e-5. We also set the probability of dropout to 0.3 and 0.4 and batch-size equal to 16 and 32. Max length of input sentences was set to 110. In the next section, we explain the results of various models and compare their performance in different situations in combination with CRF, BiLSTM-CRF, and dense layers.

5.2. Results

In order to evaluate our models, we use F1 score[7] as our major evaluation metric, which is the most common metric used for measuring the sequence labeling models' performance. Table 3 presents the performance of the proposed models on our dataset. In the first step of experiments, we considered the word2vec model [16] plus the CRF model as the baseline. Since the main research studies on Persian chunking including the proposed models by Hosseinnejad *et al*. [20], Homayoonpour and Salimibadr [10], as well as Tabatabayi and HoseinNezhad [46] are based on the CRF model, this part of experiments provides a comparative analysis with one of the available models on Persian chunking.

In the next step of experiments, we use the contextualized representation. In order to make the table more readable, we separated the parts based on the architecture of the last layer of the model. In the CRF model, mBERT achieved an F1 score of 72.90. ParsBERT outperformed mBERT, and it achieved 76.33. Due to the huge data on which XLM-RoBERTa is pre-trained, its performance is the best among all other models and gained the F1 score of 79.86.

We also tried the combination of BiLSTM-CRF with transformer-based models. In this case, mBERT achieved the 72.34 F1 score, and the score of ParsBERT was 75.70. As expected, XLM-RoBERTa is better than its counterparts again, and achieved a 79.34 F1 score.

As it can be seen from the tabulated results, since the transformer-based contextualized models have more advanced architecture than the previous RNNs like LSTM and BiLSTM, adding BiLSTM to the last layer of our models does not assist the model to catch more context from the input sequence. In all cases, the models with just a CRF layer outperform their counterparts with BiLSTM-CRF, indicating that adding BiLSTM could not improve the performance of the models in Persian text chunking.

In the next step, we evaluated the impact of the CRF layer on the output. To this aim, we compared the performance of CRF with a model that consists of a simple dense layer. Due to the ability of CRF in considering the correlations and dependencies between chunk labels, it outperforms the model with just a simple dense layer without CRF. Removing the CRF layer results in 70.62 and 75.12 F1 score for mBERT and ParsBERT, respectively, and XLM-RoBERTa achieved the 77.29 F1 score.

6. Conclusion and Future Work

In this paper, we introduced a new dataset for noun phrase chunking in the Persian language. The state-of-the-art transformer-based models such as different versions of BERT as well as XLM-RoBERTa that are trained on multiple languages including Persian were used as the major part of our models' architecture. In order to enhance the model's performance on detecting labels, a CRF layer was added to the end of the model, enabling the model to catch the dependencies between chunk tags. We also compared the impact of CRF, BiLSTM-CRF, and simple dense layer as the final layer on the models' performance.

In the future, the Persian noun phrase chunking could be easily extended to detect other types of phrases in the Persian language. The significant ability of deep contextualized models opens a room for future working on various NLP tasks that work based on sequence labeling in low-resource languages including Persian.

[3] https://pytorch.org/
[4] https://huggingface.co/
[5] https://github.com/google-research/bert/blob/master/multilingual.md
[6] https://sklearn-crfsuite.readthedocs.io/en/latest/
[7] https://github.com/chakki-works/seqeval

Table 3. Comparing F1 scores on our dataset

Model	F1 score
Baseline	
Word2vec + CRF	71.9
Contextualized representation + Dense layer	
ParsBERT	75.12
multilingual BERT	70.62
XLM-RoBERTa	77.29
Contextualized representation + CRF	
ParsBERT	76.23
multilingual BERT	72.90
XLM-RoBERTa	**79.86**
Contextualized representation + BiLSTM-CRF	
ParsBERT	75.70
multilingual BERT	72.34
XLM-RoBERTa	79.34

7. References

[1] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In Proceedings of the 27th international conference on computational linguistics, 2018. (pp. 1638–1649).

[2] A. Akhundov, D. Trautmann, and G. Groh. Sequence labeling: A practical approach, 2018. arXiv preprint arXiv:1808.03926.

[3] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. Hamshahri: A standard persian text collection. Knowledge-Based Systems. 2009. 22(5), 382–387.

[4] M. Bijankhan, J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi, Lessons from building a Persian written corpus: Peykare. Language resources and evaluation, 2011, 45(2), 143–164.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. arXiv preprint arXiv:1406.1078.

[6] K. Clark, M. Luong, C. D. Manning, and Q. V. Le, Semi-supervised sequence modeling with cross-view training. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (Eds.), Proceedings of the 2018 conference on empirical methods in natural language processing, brussels, belgium, october 31-November 4, 2018 (pp. 1914–1925). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/d18-1217.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE), 2011, 2493–2537.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wen- zek, F. Guzmán, . . . V. Stoyanov, Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th annual meeting of the association for computational linguistics, July 2020. (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.747.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio (Eds.), 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) (pp. 4171–4186). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/n19-1423.

[10] S. R. Eddy, Hidden Markov models. Current opinion instructural biology, 1996. 6(3), 361–365.

[11] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, Parsbert: Transformer-based model for Persian language understanding, 2020. arXiv preprint arXiv:2005.12515.

[12] A. K. Ghalibaf, S. Rahati, and A. Estaji, Shallow semantic parsing of Persian sentences. In Proceedings of the 23rd pacific Asia conference on language, information and computation, 2009. volume 1 (pp. 150–159).

[13] M. Ghayoomi, Bootstrapping the development of an HPSG-based treebank for Persian. Linguistic Issues in Language Technology, 2012, 7(1), 1–13.

[14] A. Graves, and J. Schmidhuber, Framewise phoneme classification with bidirectional lstm and other neural network architectures. 2005. Neural networks, 18(5-6), 602–610.

[15] C. Grover, and R. Tobin, Rule-based chunking and reusability, 2006. In Lrec (pp. 873–878).

[16] A. Hadifar, and S. Momtazi, The impact of corpus domainon word representation: a study on Persian word embeddings. Language Resources and Evaluation, 2018. 52(4), 997–1019.

[17] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher, A joint many-task model: Growing a neural network for multiple NLP tasks. In M. Palmer, R. Hwa, and S. Riedel (Eds.), Proceedings of the 2017 conference on empirical methods in natural language processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 (pp. 1923–1933). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/d17-1206.

[18] S. Hochreiter, and J. Schmidhuber, Long short-term memory. Neural computation, 1997. 9(8), 1735–1780.

[19] M. Homayoonpour, and A. Salimibadr, Determining the boundaries and syntactic phrases in Persian text. In Journal of signal and data processing. 2013.

[20] S. Hosseinnejad, Y. Shekofteh, and T. A. Emami Azadi, A'laam corpus: A standard corpus of named entity for Persian language. Signal and Data Processing, 14(3). 2017.

[21] Z. Huang, W. Xu, and K. Yu, Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991. Retrieved from http://arxiv.org/abs/1508.01991. 2015.

[22] S. Kiani, T. Akhavan, and M. Shamsfard, Developing a Persian chunker using a hybrid approach. In 2009 international multiconference on computer science and information technology (pp. 227–234). 2009.

[23] J. Lafferty, A. McCallum, and F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[24] G. Lample, and A. Conneau, Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291. 2019.

[25] L. Liu, J. Shang, F. F. Xu, X. Ren, H. Gui, J. Peng, and J. Han, Empower sequence labeling with task-aware neural language model. CoRR, abs/1709.04109. Retrieved from http:// arxiv.org/abs/1709.04109. 2017.

[26] Y. Liu, F. Meng, J. Zhang, J. Xu, Y. Chen, and J. Zhou, GCDT: A global context enhanced deep transition architecture for sequence labeling. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 2431–2441). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1233. July 2019.

[27] C. Ma, H. Zheng, P. Xie, C. Li, L. Li, and L. Si, Dm-nlp at semeval-2018 task 8: neural sequence labeling with linguistic features. In Proceedings of the 12th international workshop on semantic evaluation (pp. 707–711). 2018.

[28] C. Manning, and D. Klein, Optimization, maxent models, and conditional estimation without magic. In Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology: Tutorials-volume 5 (pp. 8–8). 2003.

[29] M. Mohseni, J. Ghofrani, and H. Faili, Persianp: a Persian text processing toolbox. In International conference on intelligent text processing and computational linguistics (pp. 75–87). 2016.

[30] S. Mohtaj, B. Roshanfekr, A. Zafarian, and H. Asghari, Parsivar: A language processing toolkit for Persian. In Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).

[31] S. Noferesti, and M. Shamsfard, A rule-based model and genetic algorithm combination for Persian text chunking. Int. J. Comput. Their Appl., 21(2), 133–140. 2014.

[32] S.-B. Park, and B.-T. Zhang, Text chunking by combining hand-crafted rules and memory-based learning. In Proceedings of the 41st annual meeting of the association for computational linguistics (pp. 497–504). 2003.

[33] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108. 2017.

[34] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer, Bert with history answer embedding for conversational question answering. In Proceedings of the 42nd international acm sigir conference on research and development in information retrieval (pp.1133–1136). 2019.

[35] A. Ramponi, R. van der Goot, R. Lombardo, and B. Plank, Biomedical event extraction as sequence labeling. In Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp) (pp. 5357–5367). 2020.

[36] M. S. Rasooli, M. Kouhestani, and A. Moloodi, Development of a Persian syntactic dependency treebank. In Proceedings of the 2013 conference of the

North American chapter of the association for computational linguistics: Human language technologies (pp. 306–314). 2013.

[37] A. Ratnaparkhi, A linear observed time statistical parser based on maximum entropy models. In C. Cardie and R. M. Weischedel (Eds.), Second conference on empirical methods in natural language processing, EMNLP 1997, providence, ri, USA, august 1-2, 1997. ACL. Retrieved from https://www.aclweb.org/anthology/W97-0301/

[38] M. Rei, Semi-supervised multitask learning for sequence labeling. In R. Barzilay and M. Kan (Eds.), Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, Vancouver, Canada, July 30-August 4, volume 1: Long papers.
(pp. 2121–2130). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/P17-1194.

[39] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors. nature, 323(6088), 533–536. 1986.

[40] S. K. Saha, and A. Prakash, Experiments on document chunking and query formation for plagiarism source retrieval. In Notebook for pan at clef 2014 (p. 990-996). September 2014.

[41] M. Shamsfard, and M. S. Mousavi, Thematic role extraction using shallow parsing. International Journal of Computational Intelligence, 4(2), 126–132. 2008.

[42] M. Shamsfard, and M. SadrMousavi, A rule-based semantic role labeling approach for Persian sentences. In Proc. of 2nd computational approach to Arabic script language. 2007.

[43] M. SharifiAtshgah, Semi-automatic development of Persian treebank. In PhD dissertation dep. of letters, Tehran uni. 2009.

[44] K. Simov, Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, and A. Kiryakov, Clark-an xml-based system for corpora development. In Proc. of the corpus linguistics 2001 conference (pp. 558–560). 2001.

[45] A. Søgaard, and Y. Goldberg, Deep multi-task learning with low level tasks supervised at lower layers. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 231–235). 2016.

[46] S. Tabatabayi, and S. HoseinNezhad, Finding the boundaries and syntactic phrases by using the corpus generated by dependency treebank. In Proceedings of the 3rd national conference on computational linguistics. 2014.

[47] E. Taher, S. A. Hoseini, and M. Shamsfard, Beheshti-NER: Persian named entity recognition using BERT. In Proceedings of the first international workshop on NLP solutions for under resourced languages (NSURL 2019) co-located with ICNLSP 2019-short papers (pp. 37–42). Trento, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/ anthology/2019.nsurl-1.6. 2019, 11–12 September.

[48] C. Thompson, USF: Chunking for aspect-term identification and polarity classification. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014) (pp.790–795). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S14-2140. August 2014.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . . . I. Polosukhin, Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008). 2017.

[50] H. Xu, B. Liu, L. Shu, and P. S. Yu, BERT post-training for review reading comprehension and aspect-based sentiment analysis.
In J. Burstein, C. Doran, and T. Solorio (Eds.), 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) (pp. 2324–2335). Association for Computational Linguis-tics. Retrieved from https://doi.org/10.18653/v1/n19-1242.

[51] J. Yang, M. Wang, H. Zhou, C. Zhao, W. Zhang, Y. Yu, and L. Li, Towards making the most of BERT in neural machine translation. In the thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020 (pp. 9378–9385). AAAI Press. Retrieved from https://aaai.org/ojs/ index.php/AAAI/article/view/6479

[52] Zhai, F., Potdar, S., Xiang, B., and Zhou, B. (2017). Neural models for sequence chunking. In S. P. Singh and S. Markovitch (Eds.), Proceedings of the thirty-first AAAI conference on artificial intelligence, February 4-9, 2017, San Francisco, California, USA (pp. 3365–3371). AAAI Press. Retrieved from http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14776.

[52] M. Asgari-Bidhendi, B. Janfada, O. R. Roshani Talab, and B. Minaei-Bidgoli, ParsNER-Social: A Corpus for Named Entity Recognition in Persian Social Media Texts. *Journal of AI and Data Mining*, *9*(2), 2021, 181-192.