Poster: A Novel Approach for POS Tagging of Pashto Language

Haris Ali Khan[1], Muhammad Junaid Ali[1] and Umm e Hanni[2]
[1]COMSATS UNIVERSITY ISLAMABAD
[2]Department of Computer Science UET Lahore

***Abstract*— Pashto is a language that belongs to the Indo-European family, mostly spoken in South Asian countries, especially in Pakistan and Afghanistan. To build software that enables us to translate Pashto sentences into various languages and building Natural Language Understand (NLU) applications to make interactive Pashto software requires a well-defined corpus and Parts of Speech (POS) tagging approach. Therefore, a well-defined corpus is developed by scraping data from different websites. We have prepared dataset according to the guidelines written for Persian and Arabic languages, as these languages are somehow similar to these languages. Training of POS tagging using BiLSTM with GloVe embedding shows the effectiveness of our proposed approach and achieved 97% accuracy.**

Keywords—Pashto, POS tagging, LSTM, Bidirectional LSTM, Corpus.

I. Introduction

Pashto is an eastern Iranian language mainly spoken in the northwest region of Pakistan and Afghanistan. It is estimated that there are over 45-60 million Pashto speakers worldwide. Due to this vast amount of Pashto speakers worldwide, it is necessary to build Natural Language Processing (NLP) based software for native speakers. Therefore, a well-defined corpus is developed, where some words are manually written, and some of them are scraped from different Pashto dictionary websites such as thepashto.com, yorku.ca, etc. Parts of Speech tagging (POS tagging) is a process in which we have to assign parts of speech to each word in a sentence not based on its meaning but also on context as well, because of the different meanings of the same word in different sentences.

Tagging words in sentences manually is a very time-consuming process. But due to advancement in NLP, this process is computed automatically by POS tagging algorithms [1]. Automatic POS tagging is not just an approach to assign tags in given sentences, but it could also aid in removing ambiguous words within a sentence. For example, given two English sentences have different parts of speech for one word. Aslam's conduct is always ethical. Hina conducts herself in a professional manner. In the first sentence the word 'conduct' is used as a noun, whereas in the second sentence, the word 'conduct' is used as a verb. Also, the noun is further subdivided into common or proper, singular or plural. Before developing a POS tagger, a well-defined tagset and dictionary is necessary.

A limited number of researchers have conducted research related to POS tagging of Pashto. A first-ever rule-based approach for POS tagging in Pashto language has been done by Rabbi et al. in [2]. They also built a tagset for POS tagging in Pashto language [3]. After getting motivated from this, we developed a tag set of over 18000 Pashto words and for POS tagging, and Bidirectional LSTM is used in this study. The rest of the paper is divided as follows: In section 2 we discussed the related work done in Pashto language, then in section 3 our proposed scheme is explained. Results and experiments are discussed in section 4 and Conclusion and future work are discussed in section 5.

II. Related Work

A. *Related Work in the Pashto Language*

A finite number of research work has been published by researchers related to natural language processing in Pashto language, Rabi et al. developed a first ever tagset for Pashto language. They follow the EAGLES guidelines, which were written for the languages of European Union. They used a tagset of 215 tags [3]. A first ever rule based tagset also built in [2]. They used 54 tags and developed a simple algorithm for POS tagging. Their POS tagger shows better results when number of words in lexicon and rules increases. They achieve 88% accuracy with 120 rules and 100,000 tagged words.

Persian is also called the sister language of Pashto and has very much lexical similarity with Pashto. Khan et al. used Bijankhan Corpus (A freely available Pashto corpus which was manually tagged by Tehran University in Iran) to tag Pashto. They also used Hidden Markov Model trigram tagger (HMM TnT) for tagging the dataset. They achieved 70.84% accuracy.

A Pashto Corpus is developed using Xaira tool, a tool used for building corpus. They collected data from books, novels, news and research publications. Their corpus contains 1.225 million words. Based on their corpus they built a morphological system of Pashto language [4]. They also built a general purpose corpus for written Pashto [5]. This corpus contains 10,000 words provided by Pashto Academy, University of Peshawar (UoP). A user friendly interface is also built for searching words in corpus.

In previous studies, various machine learning and traditional approaches have been used for POS tagging which uses hand-crafted based features and shows good results. Whereas, our study uses deep learning based approach for POS tagging.

B. *Related Work in other Languages*

Most of the work done in NLP is in languages like Arabic, English, Chinese and German etc. Several POS taggers have been developed for these languages. For instance, Arabic has many neural network based POS taggers [6], [7], [8], for English [9] and for Chinese language [10].

However, from recent years neural networks achieved better results over various traditional machine learning techniques. Most of the methods used for POS tagging in Pashto language are traditional, and vast research area in Pashto language is still pending. Therefore, in this research study we built a well-defined POS tagging of Pashto using Bidirectional LSTM technique.

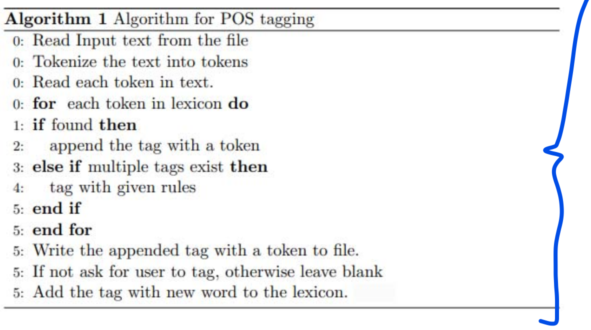
III. Proposed Scheme

A. *Proposed Tagset*

There exist several tag sets for different languages. For English there are many tagsets available i.e ENGLISH PENN TREEBANK, Oxford POS tagset, SUSANNE Corpus and HISTORICAL PENN TREEBANK etc [11][12]. For Arabic Khoja et al. developed a tagset [13] and Andrew Hardie developed a tagset for Urdu [14]. But every language has specific grammar and set of rules, therefore a separate dictionary of words with tags are necessary for this study. We built a dataset as there are no publicly available POS tagging datasets and there is a need for large-scale POS tagging dataset

B. *Building Lexicon*

To build lexicon, Pashto words are scraped from various Pashto dictionary websites and some of them are manually written. For scraping Beautiful Soup (BS) library is used in python and data is pre-processed with pandas and numpy libraries. The proposed lexicon is based on 18000 properly annotated unique words and also can be used for other purposes.



Algorithm 1 Algorithm for POS tagging

```
Read Input text from the file
Tokenize the text into tokens
Read each token in text.
for each token in lexicon do
if found then
    append the tag with a token
else if multiple tags exist then
    tag with given rules
end if
end for
Write the appended tag with a token to file.
If not ask for user to tag, otherwise leave blank
Add the tag with new word to the lexicon.
```

Figure 1 shows the visual representation of proposed algorithm used to perform this study and Algorithm of the said approach is shown in Algorithm 1. The ratio of occurrence of manual tagging is 5-10 % which are manually tagged by us. For manual tagging in case of multiple tags exists we 1) search from Pashto dictionary for given tag and 2) uses rules given in [2]

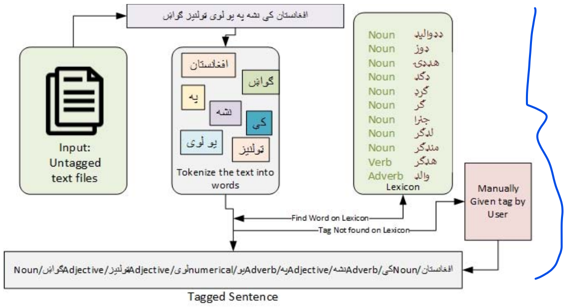
C. *Bi-LSTM RNN for POS Tagging*

From recent years, Recurrent Neural Networks (RNN) has been shown very effective for tagging of sequential data such as speech recognition and non-segmental handwritten characters. The idea of RNN was to connect information with previous information that is why it is very useful for sequential data. But the problem RNN is that it cannot handle huge data due to vanishing and exploiting gradient problems.

Long Short Term Memory (LSTM) is a type of RNN that came to overcome the problem of RNN and have long term dependencies. Due to this property LSTM is commonly used in NLP related problems. Bi-Directional

978-1-7281-7407-5/20/$31.00 ©2020 IEEE
DOI 10.1109/SMART-TECH49988.2020.00068

LSTM-RNN has shown effective performance for sequential data tagging. It shows state of the art results on Penn Treebank dataset for POS tagging [9]. As it is a powerful model for sequential labeling tasks such as POS tagging. Therefore, we choose Bi-LSTM RNN for this task.


افغاستان کی نشه یه یو لوی تولنیز گواښ
Input: Untagged text files
افغاستان
گواښ
یه
نشه
کی
یو لوی
تولنیز
Tokenize the text into words
Noun ددوالید
Noun دوز
Noun هدیی
Noun دګد
Noun ګرد
Noun ګر
Noun جنرا
Noun لدګر
Noun مندګر
Verb هدګر
Adverb والد
Lexicon
Manually Given tag by User
Find Word on Lexicon
Tag Not found on Lexicon
Noun/گواښAdjective/تولنیزAdjective/لویnumerical/یوAdverb/یهAdjective/نشهAdverb/کیNoun/افغاستان
Tagged Sentence

.

Fig. 1: Visual representation of proposed algorithm.

Given a sentence set s={ $S_1,S_2,S_3,....,S_n$ } with tags t={ $t_1,t_2,t_3,...,t_n$ } .Bi LSTM RNN is used to predict the tag probability distribution of each word. Where S_i is the one hot representation of word in a sentence and f(S_i) returns a three-dimensional representation vector for a word. The input vector I_i is represented as:

$$I_i = S_1S_i + S_2S_i$$

Where S_1 and S_2 are Weight matrices connected to two layers. For word representation of tagged corpus, Global Vectors for word representation (GloVe), an unsupervised algorithm is used [15]. Each of the word used in their original form doesn't give semantic meaning. However, defining a word in vector form gives more information. These vectors contain both semantic and syntactic information. GloVe vectors are generated by creating co-occurrence matrix of all words in corpus by Singular Value Decomposition (SVD) to which create k dimension representation of each word. The detailed description of GloVe vector is given in [15].

Our word-tag representation is also converted into a single file using Pickle library. For each word-tag representation our model has to predict a tag. The architecture used for our model is:

- Embedding Layer: This layer converts the input word into unique number to the GloVe vector
- Bi-LSTM Layer: This layer learns the sequence of data so that it creates feature pattern that would be given to fully connected layer for prediction.
- Fully Connected Layer: This layer takes input from Bi-LSTM layer and gives to lower dimension form. The output layer is further given to Softmax function for prediction that gives probability values. The output layer dimension is the number of tags.

To train and update the weights of our architecture, Adam optimization algorithm is used. In this algorithm each parameter has a learning rate and each learning rate is divided by running average of the magnitude of previous gradients.

IV. Experiments and Results

Bi-LSTM-RNN used in our experiments are implemented in Keras library built on Tensorflow, which uses GPU for fast execution [16]. We run experiments on NVIDIA GeForce 930Mx. For training of neural network, Adam optimizer is used.

For building corpus, we took 19 articles from BBC Pashto (A well-known British Broadcasting Service Pashto news site) [17] which consists of 18000 words. For training, size of input layer is 64, hidden layer is 50 and output layer is 14 as we have 14 tags in our corpus. Table 1 shows the impact of accuracy on the number of words in lexicon. By using 5000 words in lexicon and 5 epochs gives 56% accuracy. On 10000 words, the accuracy increase to 77% and using 18000 words it achieves 97% accuracy which shows increasing the numbers of words in lexicon improves the performance in terms of accuracy. The learning curve of training and validation using 18000 words lexicon of accuracy score is shown in Figure 3(a) and loss curves of training and validation set is shown in Figure 3(b).

Table 1. Effect of using different words in lexicon on accuracy

Words in lexicon	Number of Epochs	Accuracy
5000	5	58%
10000	5	77%
18000	5	97%

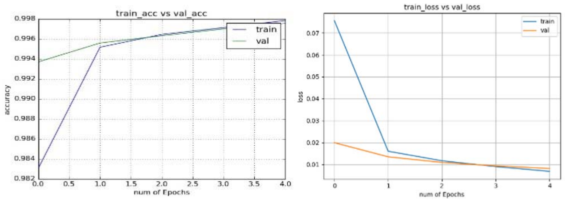


Fig. 3: (a) Training and validation accuracy on 5 epochs (b) Training and validation loss on 5 epochs

Conclusion

To make effective NLP based applications and parser for Pashto Language, POS tagging is an effective task. POS tagging plays an important role in NLP for every language. This paper presents a neural network based approach for POS tagging in Pashto language which shows effective performance. By expanding the size of lexicon and adding multiple tag sets and rules, we achieve better results. The future work also includes topic modeling of Pashto language, building a parser for Pashto language using neural network and trying larger number of corpus.

References

[1] M. F. Hasan, N. UzZaman, and M. Khan, "Comparison of unigram, bigram, hmm and brill's pos tagging approaches for some south Asian languages," 2007.

[2] I. Rabbi, A. Khan, and R. Ali, "Rule-based part of speech tagging for pashto language," in Conference on Language and Technology, Lahore, Pakistan, 2009.

[3] I. Rabbi, M. A. Khan, and R. Ali, "Developing a tagset for pashto part of speech tagging," in Electrical Engineering, 2008. ICEE 2008. Second International Conference on. IEEE, 2008, pp. 1–6.

[4] M. A. Khan and F. T. Zuhra, "A corpus-based study of pashto."

[5] "A general-purpose monitor corpus of written pashto." Citeseer.

[6] J. H. Yousif and T. Sembok, "Recurrent neural approach based Arabic part-of-speech tagging," in proceedings of International Conference on Computer and Communication Engineering (ICCCE'06), vol. 2, 2006, pp. 9–11.

[7] Y. Belinkov and J. Glass, "Arabic diacritization with recurrent neural networks," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2281–2285.

[8] K. Darwish, H. Mubarak, A. Abdelali, and M. Eldesouki, "Arabic pos tagging: Don't abandon feature engineering just yet," in Proceedings of the Third Arabic Natural Language Processing Workshop, 2017, pp. 130–137.

[9] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Part-of-speech tagging with bidirectional long short-term memory recurrent neural network," arXiv preprint arXiv:1510.06168, 2015.

[10] Y. Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, "Character-based joint segmentation and pos tagging for chinese using bidirectional rnn-crf," arXiv preprint arXiv:1704.01314, 2017.

[11] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," Computational linguistics, vol. 19, no. 2, pp. 313–330, 1993.

[12] A. Taylor, M. Marcus, and B. Santorini, "The penn treebank: an overview," in Treebanks. Springer, 2003, pp. 5–22.

[13] S. Khoja, R. Garside, and G. Knowles, "A tagset for the morphosyntactic tagging of arabic," Proceedings of the Corpus Linguistics. Lancaster University (UK), vol. 13, 2001.

[14] A. Hardie, "The computational analysis of morphosyntactic categories in urdu," Ph.D. dissertation, Lancaster University, 2004.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[16] F. Chollet et al., "Keras: The python deep learning library," Astrophysics Source Code Library, 2018.

[17] BBC News پښتو. 2020. *BBC News*. [online] Available at: <https://www.bbc.com/pashto> [Accessed 8 March 2020].