SiPOS: A Benchmark Dataset for Sindhi Part-of-Speech Tagging

Wazir Ali and **Zenglin Xu**
SMILE Lab, School of Computer Science and Engineering
University of Electronic Science and Technology of China, Chengdu 611731, China
{aliwazirjam,zenglin}@gmail.com

Jay Kumar
Data Mining Lab,School of Computer Science and Engineering
University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract

In this paper, we introduce the SiPOS dataset for part-of-speech tagging in the low-resource Sindhi language with quality baselines. The dataset consists of more than 293K tokens annotated with sixteen universal part-of-speech categories. Two experienced native annotators annotated the SiPOS using the Doccano text annotation tool with an inter-annotation agreement of 0.872. We exploit the conditional random field, the popular bidirectional long-short-term memory neural model, and self-attention mechanism with various settings to evaluate the proposed dataset. Besides pre-trained GloVe and fastText representation, the character-level representations are incorporated to extract character-level information using the bidirectional long-short-term memory encoder. The high accuracy of 96.25% is achieved with the task-specific joint word-level and character-level representations. The SiPOS dataset is likely to be a significant resource for the low-resource Sindhi language.

1 Introduction

Annotated corpus is an essential resource for developing automatic natural language processing (NLP) systems (Ali et al., 2020). Such language resources (LRs) play a significant role in the digital survival of human languages (Jamro, 2017). The part-of-speech (POS) tagging is a fundamental preprocessing task in various NLP applications (Mahar and Memon, 2010), used to assign appropriate in-context POS tags to each word. One of the main challenges (Britvić, 2018) in POS tagging is ambiguity since one word can take several possible POS labels. Another problem is the unspoken or complex POS or words. Both of these problems are not rare in natural languages. Moreover, there is a lack of benchmark labeled datasets for Sindhi POS tagging. To tackle these challenges, we propose a novel benchmark SiPOS tagset.

Sindhi is a rich and complex morphological language (Rahman, 2010). It is a low-resource language (Ali et al., 2019) which lacks primary LRs for mature computational processing. Sindhi is being written in two famous writing systems of Persian-Arabic, Devanagari, and more recently Roman (Sodhar et al., 2019) is also getting popularity. However, Persian-Arabic is standard script as well as frequently used in literary work, online communication, and journalism. Sindhi POS tagging has been previously investigated in various scripts including Persian-Arabic,Devanagari (Jamro, 2017), and Roman (Sodhar et al., 2019). However, the low-resource Sindhi language lacks a POS labeled dataset for its supervised text classification.

In this paper, we introduce a novel benchmark SiPOS dataset for the low-resource Sindhi language. Due to the scarcity of open-source POS labeled data, two native experienced annotators performed the POS annotation of Sindhi text using the Doccano (Nakayama et al., 2018) text annotation tool. To the best of our knowledge, this is the first attempt to address the Sindhi POS tagging at a large scale by proposing a new gold-standard SiPOS dataset[1] and exploiting conditional random field (CRF), bidirectional long-short-term memory (BiLSTM) network, and self-attention for its evaluation. Our novel contributions are as follows:

- We reveal a novel open-source gold-standard SiPOS dataset for the low-resource Sindhi language. We manually tagged more than 293k tokens of the Sindhi news corpus using the Doccano text annotation tool.
- We compute the inter-annotator agreement and exploit machine learning models of CRF and BiLSTM, self-attention to evaluate the proposed dataset with different settings.

[1]The SiPOS dataset is publicly available @ https://github.com/AliWazir/SiPOS-Dataset

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 22–30,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_004

- Besides pre-trained GloVe, fastText word-level representation, the task-specific character-level word representations are incorporated to extract character-level information using BiLSTM encoder.

2 Related Work

The labeling of natural language text with POS tags can be a complicated task, requiring much effort, even for trained annotators (Rane et al., 2020). A large number of LRs are publicly available for high-resource languages such as English (Marcus and Marcinkiewicz), Chinese, Indian languages (Baskaran Sankaran and Subbarao, 2008; Khan et al., 2019) and others (Petrov et al., 2012). Unlike rich-resourced languages such as English and Chinese with abundant publicly accessible LRs, Sindhi is relatively low-resource (Ali et al., 2019), which lacks the POS tagged dataset that can be utilized to train a supervised or statistical algorithm.

Previously, Mahar and Memon (2010) proposed a Sindhi POS labeled dataset consists of 33k tokens. Later, Dootio and Wagan (2019) published a new dataset containing 6.8K lexicon, which is insufficient to train a robust supervised classification algorithm. More recently, (Rahman et al., 2020) annotated 100K words by employing a multi-layer annotation model, which comprises different annotation layers like POS, morphological features, dependency structure, and phrase structure. But their dataset is not publicly available. Except for Sindhi Persian-Arabic, the POS tagged datasets in Devanagari (Motlani et al., 2015), and Roman (Sodhar et al., 2019) scripts have also been introduced. The POS tagged corpus of Sindhi-Devanagari consists of 44K tokens, while Sindhi-Roman (Sodhar et al., 2019) only contains 100 sentences. The review of existing work shows that the low-resource Sindhi language lacks a benchmark POS labeled dataset for its supervised text classification.

3 Corpus Acquisition and Annotation

In this section, we illustrate the opted annotation methodology. We utilized the news corpus (Ali et al., 2019) of popular and most circulated Sindhi newspapers of Kawish and Awami-Awaz (see Table 1. Two native graduate students of linguistics were engaged for the annotation purpose using the Doccano (Nakayama et al., 2018) text annotation tool to assign a POS label to each token. The detailed annotation process is illustrated as under:

Table 1: The statistics of news articles utilized for the annotation of SiPOS tagset.

Resource	Articles	Sentences	tokens
Kawish	563	3 769	1 58 145
Awami-Awaz	458	3 015	1 35 539
Total	1 021	6 784	2 93 684

3.1 Preprocessing

Sindhi news corpus contains a certain amount of unwanted data (Ali et al., 2019). Thus, filtering out such data and normalizing it is essential to obtain a more authentic vocabulary for the annotation project. The preprocessing steps consist of the following steps:

- Removal of unwanted multiple punctuation marks from the start and end of the sentences.
- Filtration of noisy data such as non-Sindhi words, special characters, HTML tags, emails, URLs, etc.
- Tokenization to normalize the text, removal of duplicates, and multiple white spaces.

Moreover, it requires human efforts and careful assessment for the consistency in the labeled dataset. Sindhi Persian-Arabic is being written in the right to left direction (Jamro, 2017). An example of a Sindhi sentence is given in Table 2 with language-specific and corresponding universal part-of-speech (UPOS) tags. A Sindhi word comprises one or more clitics or segments (Narejo and Mahar, 2016), typically a stem to which prefix and suffix may be attached. Therefore, the tagging can be done for each clitic in sequence or a word simultaneously. For the annotation, we used Doccano (Nakayama et al., 2018) to assign a POS label to each token. The Doccano is an open-source annotation platform for sequence labeling and machine translation tasks. We engaged two native graduate students of linguistics for the annotation purpose. The annotators also used an online Sindhi Thesaurus portal[2] in case of ambiguity or confusion while deciding a POS label for a token. Moreover, the project supervisor also worked with annotators to monitor annotation quality by following the annotation guidelines (Dipper et al., 2004; Petrov et al., 2012).

[2] http://dic.sindhila.edu.pk/

Table 2: An example of a Sindhi sentence with its corresponding language specific and universal part-of-speech tags. The Roman transliteration of each token is given for the ease of reading.

.	آهي	ويندو	ڪرايو	عمل	سان	سختي	تي	ميڊيا	اليڪٽرانڪ	۽	پرنٽ	**Sentence**
SYM	**AUX**	**VERB**	**VERB**	**NOUN**	**ADP**	**NOUN**	**ADP**	**NOUN**	**NOUN**	**CONJ**	**NOUN**	**UPOS**
PUNCT	**AUX**	**VB**	**VB**	**NN**	**ADP**	**NN**	**ADP**	**NN**	**NN**	**CONJ**	**NN**	**Tag**
بيهڪ جي نشاني	فعل معاون	فعل	فعل	اسم	حرفِ جر	اسم	حرفِ جر	اسم	اسم	حرف جملو	اسم	**Sindhi POS**

3.2 Consistency Evaluation

To ensure annotation consistency, we measure the inter-annotator agreement to investigate the consistency in which annotators agreed to the tags. To measure inter-annotator agreement, we chose to use Cohen's Kappa (Cohen, 1960). Cohen's Kappa measures the inter-annotator agreement between two annotators. Since we have two annotators, we compute this measure for POS tag pairs that show agreement between two annotators, which leads to two results. The inter-annotator agreement comes out to be 0.872 with a confidence percentile of 95%. Cohen's Kappa value shows that the dataset is of acceptable quality.

3.3 SiPOS Dataset

The SiPOS has been annotated using the news corpus (Ali et al., 2019) of Kawish and Awami-Awaz Sindhi newspapers. Sindhi grammar (Oad, 2012) give Sindhi POS of nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, numerals, articles, and interjections. The dataset consists of more than 293k tokens annotated with sixteen Sindhi POS and UPOS categories, respectively. The complete statistics of the utilized news corpus in the annotation is given in Table 1. The detailed label distribution in the SiPOS is given in Table 3.

4 Evaluation Methods

We evaluate the SiPOS for the consistency in the dataset by computing the inter-annotator agreement using Cohen's Kappa (Cohen, 1960) coefficient. We evaluate the proposed SiPOS dataset by exploiting CRF, BiLSTM and integrating CRF and self-attention in the BiLSTM network for strong baselines. Moreover, pre-trained GloVe, fastText word representations, and task-specific character-level, and joint $Word_{Character}$ level representations are incorporated to extract word-level and character-level information using the BiLSTM encoder.

4.1 Conditional Random Field

We initially evaluate the SiPOS dataset using a CRF (Lafferty et al., 2001), widely used in sequence classification (Sutton et al., 2012) tasks. The CRF is useful to consider the relationship between labels and jointly decode the most suitable chain of labels for a given input sentence (Huang et al., 2015).

4.2 Representation Learning

Representation learning aims to capture the useful semantic, syntactic, and morphological information (Santos and Zadrozny, 2014) in NLP tasks (Bojanowski et al., 2017). We use pre-trained GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017) word-level representations, character-level representations as well as joint character-level and word-level $Word_{Character}$ representations (Shao et al., 2017; Matteson et al., 2018) to extract the word-level features. Pre-trained word representations enable neural models to exploit the raw textual data larger than annotated data. Then, we jointly learn the task-specific character-level word representations (Liu et al., 2018) using the BiLSTM network. The task-specific contextual representations include the POS-based knowledge.

4.2.1 GloVe

The GloVe (Pennington et al., 2014) is a word representation model that relies on two methods of global word-to-word co-occurrence statistics and local context window. We obtain the pre-trained GloVe representation by training on the large corpus of more than 61 million words (Ali et al., 2019). We train GloVe with AdaGrad by choosing the context window of 5 and the 300-dimensional word representations. We filter out Sindhi stop words (Ali et al., 2019) in the preprocessing step.

4.2.2 fastText

The fastText (Bojanowski et al., 2017) is similar to Word2vec (Mikolov et al., 2013). It uses subword

Table 3: Complete statistics of SiPOS dataset with the number of POS in each label. The highest proportion in the POS labels is noun, followed by preposition and verb.

Count	SPOS	UPOS	Tag	POS type
65,611	اسم	NOUN	NN	Noun
54,810	حرفِ جر	ADP	ADP	Preposition
41,882	فعل	VERB	VB	Verb
21,849	صفت	ADJ	ADJ	Adjective
20,627	ضمير اشارو	DET	DET	Determiner
20,281	نشاني	SYM	PUNCT	Punctuation, symbol
19,823	ظرف	ADV	ADV	Adverb
11,546	اسم خاص	PROPN	NNP	Proper noun
11,015	حرف جملو	CONJ	CONJ	Conjunction
10,553	فعل معاون	AUX	AUX	Auxiliary verb
7,878	ضمير	PRON	PRON	Pronoun
4,888	صفت عددي	NUM	NUM	Numerical adjective
1,424	اُڌارولفظ	-	FOW	Borrowed words
1,091	نامعلوم	X	UNK	Unknown
282	حرف جملو شرطيہ	SCON	SCON	Subordinating conjunction
124	حرف ندا	INTJ	INTJ	Interjection

information in the prediction model to obtain word representations. We train fastText on recently proposed unlabelled Sindhi corpus (Ali et al., 2019) of more than 61 million words. In training, we use the recommended sub-sampling (Bojanowski et al., 2017), negative sampling, the minimum and maximum length of character ngrams (Grave et al., 2018), minimum word count, learning rate, 300-dimensional representations, and default context window size.

4.2.3 Character-level Word Representations

The character-level representations have an advantage in handling the out-Of-the-vocabulary problem because they can learn almost all character representations from even small or moderate corpus (Jia and Ma, 2019). In other words, these representations are good at inferring unseen words and sharing information about morpheme-level regularities. The BiLSTM network learns the character-level representations of words and associates them with usual word representations to perform POS tagging. We employ task-oriented strategy (Liu et al., 2018) for character-level and joint $Word_{Character}$ level representations learned through BiLSTM network (Shao et al., 2017; Matteson et al., 2018) which are different from pre-trained word representations. The BiLSTM is good at capturing prefixes and suffixes from the given input text (Zhang et al., 2018). It consists of interconnected bidirectional forward $\overrightarrow{LSTM}$ and backward $\overleftarrow{LSTM}$ hidden layers, which efficiently encode the contextual information.

4.3 Neural POS Taggers

4.3.1 BiLSTM

The BiLSTM network (Schuster and Paliwal, 1997) has been broadly used in a variety of sequence labelling tasks (Huang et al., 2015; Ma and Hovy, 2016; Peters et al., 2017) including POS tagging (Kann et al., 2018). In this work, we evaluate the SiPOS dataset using the BiLSTM network. The model consists of representations layer, BiLSTM encoder, and softmax for each position in the final layer. The bidirectional layers extract character-level, word-level features and then adopt a random initialization method to transform words into representations. The BiLSTM word-level (pre-train) model is the same as BiLSTM (word-level) but adopts GloVe and fastText for representations. Sim-

ilar to pre-trained GloVe, fastText, the dimension of word representations was set to 300 to initialize the representation layer, and a context window size of 5 was selected. The dimension of character-level representation was set to 50 in character-level models.

4.3.2 Extensions of Neural Model

The CRF and BiLSTM network form strong baselines. Then, we add more variants such as CRF decoder, self-attention, and character-level feature representations. The CRF is employed on the top of the neural models (Huang et al., 2015; Ma and Hovy, 2016; Shen et al., 2018) as a decoder. Moreover, the self-attention (Vaswani et al., 2017) is used above the encoder layers (Shen et al., 2018) to boost the model performance by focusing on tokens with more meaning that contribute to label prediction. Thus, we integrate self-attention for performance analysis into the BiLSTM and BiLSTM-CRF models to deeply capture semantic information and lexical features. In the word-level models, a sequence of words is given as an input where each word is represented in the word representation. However, the character-level models (Shao et al., 2017) consider each sentence as a sequence of characters (Matteson et al., 2018), and outputs a label distribution for each character, then concatenated to word representations. The character-level representations use the one-dimensional neural network or any other model to find the numeric representations of words by looking at their character-level compositions. Our final BiLSTM-Attention-CRF model relies upon joint task-specific character-level and word-level representations, BiLSTM encoder, self-attention, and CRF as depicted in Figure 1.

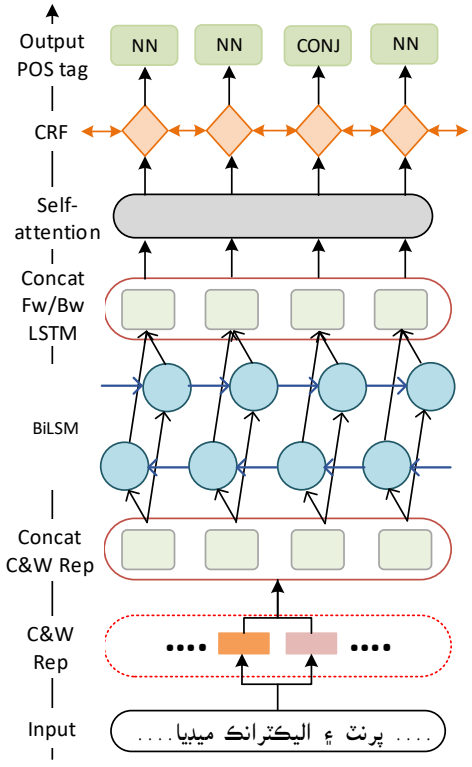

Figure 1: The overall architecture of the BiLSTM-Attention-CRF model. Task-specific joint character-level and word-level representations are regarded as input to the BiLSTM encoder. Then encoder captures the contextual features. The output of the BiLSTM encoder layer is fed into the self-attention layer before decoding through a hidden layer. Finally, we employ CRF to yield the output tag sequence. Concat denotes the concatenation operation; Rep is the representation learning, C&W is the character-level and word-level representations, while Fw&Bw represents the forward, backward LSTM layers.

4.4 Experimental Setting

Since our baseline models and extended neural baselines only rely on labeled training data, no external resources are used. The SiPOS dataset is split into train, validation, and test sets. We report the token level POS tagging accuracy on the large classes in a test collection. All the experiments were conducted on GTX $1080Ti$ Nvidia GPUs using TensorFlow (Abadi et al., 2016).

4.5 Training

We evaluated several hyperparameter configurations (Reimers and Gurevych, 2017) and picked optimal parameters that work well on the SiPOS dataset. The optimization is performed using Adam (Kingma and Ba, 2015) with a learning rate of 0.01 and decay of 0.8, respectively. The BiLSTM hidden layer has 200 units for each direction. A dense layer follows each hidden unit. We project input features by utilizing the dense layer. The batch size was set to 32 for all experiments, except for the CRF, set to 16. A dropout (Srivastava et al., 2014) of 0.25% was applied at the unit representation layer for all the reported neural models with the best performance. Epoch count was set to 50 with early-stopping (Caruana et al., 2001) after five epochs with no improvement in the validation set. We use identical hyperparameters for all neural models.

4.6 Results and Analysis

In this section, we present the results of CRF, BiLSTM baselines, and extended BiLSTM-CRF, BiLSTM-Attention, BiLSTM-Attention-CRF neural POS taggers by employing four representation learning approaches. We conducted several experiments to validate the SiPOS dataset to determine the quality and whether the neural models rely upon the pre-trained word-level, character-level, joint representations, or otherwise due to setting up more model parameters. All the neural models are used to predict the labels in the SiPOS dataset. We initialize pre-trained GloVe, fastText word representations, task-specific character-level, and joint representations learned through BiLSTM encode for the neural models. The results of the CRF and four neural POS taggers are presented in Table 4 with POS tagging accuracy using GloVe word representations. Table 5 shows the results of fastText, the performance of the neural models is depicted in Table 6 using character-level representations learning, and Table 7 presents the results of joint representation learning. In contrast to GloVe, the character-level representations yield better performance. However, the neural POS taggers surpass GloVe and character-level representations with fastText. It is because of the representation learning at the character level with the subword model. The presented results also demonstrate that the CRF is dominant over softmax in neural models. Moreover, it is also important to note that self-attention has enhanced the accuracy across all the experiments. Furthermore, joint word-level and character-level representations are dominant over pre-trained word representations and task-specific character-level representations. However, the performance of the character-level neural models is close to the fastText. It is imperative to mention that joint character-level word representations surpass both pre-trained and character-level representation learning. Our final BiLSTM-Attention-CRF model yields the best accuracy using all types of representation learning compared to BiLSTM, BiLSTM-attention, and BiLSTM-CRF. It surpasses all the models by yielding an accuracy of 96.25% with task-specific joint character-level and word-level representations. The empirical results demonstrate the slight performance difference between pre-trained word-level, task-specific character-level, and joint representations. Thus, it can be determined in the performance comparison that representation learning greatly impacts the performance of the neural network models. Conclusively, it can be observed that the BiLSTM-Attention-CRF with joint representation learning gave us the most significant improvement in the overall accuracy over pre-trained GloVe, fastText, and task-specific character-level representations.

Table 4: POS tagging accuracy (%) using CRF initial baseline and neural models on pre-trained GloVe word representations. Bold font denotes the best results on the GloVe.

Model	Accuracy
CRF	90.34
BiLSTM	92.73
BiLSTM-CRF	93.26
BiLSTM-Attention	93.58
BiLSTM-Attention-CRF	**93.89**

Table 5: POS tagging accuracy (%) on the pre-trained fastText word representations. Bold font denotes the best results on the fastText.

Model	Accuracy
BiLSTM	94.21
BiLSTM-CRF	94.74
BiLSTM-Attention	94.89
BiLSTM-Attention-CRF	**95.32**

Table 6: POS tagging accuracy (%) on the task-specific character-level representations. Bold font highlights the best results.

Model	Accuracy
BiLSTM	93.86
BiLSTM-CRF	94.25
BiLSTM-Attention	94.43
BiLSTM-Attention-CRF	**95.19**

Table 7: POS tagging accuracy (%) on neural models using task-specific joint character-level and word-level representations. Bold font shows the best results across all the experiments.

Model	Accuracy
BiLSTM	94.37
BiLSTM-CRF	94.78
BiLSTM-Attention	**95.49**
BiLSTM-Attention-CRF	**96.25**

5 Conclusion and Future Work

In this paper, we propose a novel benchmark SiPOS dataset for the low-resource Sindhi language. The dataset consists of more than 293K tokens, annotated using the Doccano text annotation tool. We exploit CRF as the initial baseline and BiLSTM model by incorporating CRF as a decoder and self-attention mechanism for the performance gain. The BiLSTM-Attention-CRF model yields a high accuracy of 96.25% with the joint task-specific character-level and word-level representations. The proposed open-source SiPOS dataset will be a sophisticated addition to the resources for the Sindhi language. In the future, we intend to pre-train bidirectional encoder representations from transformers (BERT) and generative pre-trained transformer (GPT) language models for Sindhi text classification.

Acknowledgements

We thank to the anonymous reviewers for concrete suggestions. This work was funded by the National Key R&D Program of China (No. 2018YFB1005100 & No. 2018YFB1005104).

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th $USENIX$ Symposium on Operating Systems Design and Implementation ($OSDI$) 16)*, pages 265–283.

Wazir Ali, Jay Kumar, Junyu Lu, and Zenglin Xu. 2019. Word embedding based new corpus for low-resourced language: Sindhi. *arXiv preprint arXiv:1911.12579*.

Wazir Ali, Junyu Lu, and Zenglin Xu. 2020. SiNER: A large dataset for Sindhi named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2946–2954. European Language Resources Association.

Monojit Choudhury Tanmoy Bhattacharya Pushpak Bhattacharyya Girish Nath Jha S. Rajendran K. Saravanan L. Sobha Baskaran Sankaran, Kalika Bali and K.V. Subbarao. 2008. A common parts-of-speech tagset framework for Indian languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tihana Britvić. 2018. *Semi-supervised neural part-of-speech tagging*. Ph.D. thesis, University of Zagreb. Faculty of Science. Department of Mathematics.

Rich Caruana, Steve Lawrence, and C Lee Giles. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pages 402–408.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Stefanie Dipper, Michael Götze, and Stavros Skopeteas. 2004. Towards user-adaptive annotation guidelines. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 23–30.

Mazhar Ali Dootio and Asim Imdad Wagan. 2019. Syntactic parsing and supervised analysis of Sindhi text. *Journal of King Saud University-Computer and Information Sciences*, 31(1):105–112.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Language Resources and Evaluation Conference*, CONF.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Wazir Ali Jamro. 2017. Sindhi language processing: A survey. In *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, pages 1–8. IEEE.

Yaozong Jia and Xiaopan Ma. 2019. Attention in character-based BiLSTM-CRF for Chinese named entity recognition. In *In Proceedings of the 4th International Conference on Mathematics and Artificial Intelligence*, pages 1–4.

Katharina Kann, Johannes Bjerva, Isabelle Augenstein, Barbara Plank, and Anders Søgaard. 2018. Character-level supervision for low-resource pos tagging. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 1–11.

Wahab Khan, Ali Daud, Khairullah Khan, Jamal Abdul Nasir, Mohammed Basheri, Naif Aljohani, and Fahd Saleh Alotaibi. 2019. Part of speech tagging in Urdu: Comparison of machine and deep learning approaches. *IEEE Access*, 7:38918–38936.

D Kingma and J Ba. 2015. Adam: A method for stochastic optimization. *San Diego*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Qian Liu, He-Yan Huang, Yang Gao, Xiaochi Wei, Yuxin Tian, and Luyang Liu. 2018. Task-oriented word embedding for text classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2023–2032.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Javed Ahmed Mahar and Ghulam Qadir Memon. 2010. Sindhi part of speech tagging system using wordnet. *International Journal of Computer Theory and Engineering*, 2(4):538.

Mitchell P Marcus and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19(2).

Andrew Matteson, Chanhee Lee, Youngbum Kim, and Heui-Seok Lim. 2018. Rich character-level information for korean morphological analysis and part-of-speech tagging. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2482–2492.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Raveesh Motlani, Harsh Lalwani, Manish Shrivastava, and Dipti Misra Sharma. 2015. Developing part-of-speech tagger for a resource poor language: Sindhi. In *Proceedings of 7th Conference on Language and Technology, Poznan, Poland*.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human.

Waqar Ali Narejo and Javed Ahmed Mahar. 2016. Morphology: Sindhi morphological analysis for natural language processing applications. In *2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pages 27–31. IEEE.

JD Oad. 2012. *Implementing GF resource grammar for Sindhi language. Doctor dissertation*. Ph.D. thesis, M. Sc. thesis, Chalmers University of Technology, Gothenburg, Sweden.

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

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1756–1765.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096.

Mutee U Rahman. 2010. Towards Sindhi corpus construction. In *Proceedings of the Conference on Language and Technology*.

Mutee U Rahman, Tafseer Ahmed, and Muhammad Shaheer Memon. 2020. Development of annotated corpus resources of Sindhi. *LANGUAGE & TECHNOLOGY*, page 49.

Gajanan Rane, Nilesh Joshi, Geetanjali Rane, Hanumant Redkar, Malhar Kulkarni, and Pushpak Bhattacharyya. 2020. Part-of-speech annotation challenges in marathi. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 1–6.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st international conference on machine learning (ICML-14)*, pages 1818–1826.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 173–183.

T Shen, T Zhou, G Long, J Jiang, and C Zhang. 2018. Bi-directional block self-attention for fast and memory-efficient sequence modeling. In *Proceedings of the International Conference on Representation Learning*.

Irum Naz Sodhar, Akhtar Hussain Jalbani, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. 2019. Parts of speech tagging of Romanized Sindhi text by applying rule based model. *IJCSNS*, 19(11):91.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Meishan Zhang, Nan Yu, and Guohong Fu. 2018. A simple and effective neural model for joint word segmentation and pos tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1528–1538.