2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)
22nd-23rd August 2019 at Institut Teknologi dan Bisnis (ITB) STIKOM Bali, Indonesia

Part of Speech Tagging for Indonesian Language using Bidirectional Long Short-Term Memory

Dellon Handrata
Department of Information Technology
Institut Sains dan Teknologi Terpadu Surabaya
Surabaya, Indonesia
dellon80@gmail.com

Christian Nathaniel Purwanto
Department of Informatics
Institut Sains dan Teknologi Terpadu Surabaya
Surabaya, Indonesia
christian.np@stts.edu

Fransisca Haryanti Chandra
Department of Electronics Engineering
Institut Sains dan Teknologi Terpadu Surabaya
Surabaya, Indonesia
fhc@stts.edu

Joan Santoso
Department of Information Technology
Institut Sains dan Teknologi Terpadu Surabaya
Surabaya, Indonesia
joan@stts.edu

Gunawan
Department of Information Technology
Institut Sains dan Teknologi Terpadu Surabaya
Surabaya, Indonesia
gunawan@stts.edu

***Abstract*—Part of Speech (POS) is a label to distinguish a word based on its grammatical and morphological form. By providing the POS label, we can get the contextual meaning. This label can be used as contextual features for several computational linguistic research — for example, word sense disambiguation, chunking, machine translation, and sequence classification. Our work is done by using bidirectional long short-term memory to do the part of speech tagging task for Bahasa Indonesia. We use deep learning model for Indonesia language POS tagging because deep learning can achieve excellent performance on it. We could reach 96.92% of F1 Score based on our approach.**

Keywords—part of speech tagging, Indonesian language, bidirectional long short-term memory, natural language processing, deep learning

I. INTRODUCTION

A single word can express multiple meanings according to the context of its sentence. The word "apel" in the Indonesian language can be the name of a fruit or an event. Part of speech label can be used to determine the context of the word. It can provide more information for further tasks in natural language processing. However, we still face the fact that there is still a few research and dataset available on POS Tagging in Bahasa Indonesia. Conducting research on this task is a need.

Some natural language tasks take advantage of POS label by digging the contextual feature from each word. Word sense disambiguation [1] uses POS Tagging to keep in mind the context used in the sentence and relationship with adjacent words or phrases in the corpus. Noun phrase chunker [2] also use part of speech label after preprocessing the raw texts. Machine translation [3] take advantage of part of speech label. Sentiment analysis on short texts [4] or even long documents [5] uses part of speech to help the classification of the sentiments. There is another feature called named entity [6] which often predicted together with part of speech because these features are not independent of each other

POS Tagging can be viewed as a sequence classification task which predicts a class label for every time step of the input. Based on this view, we use a recurrent model built on neural networks with a gating system. Bidirectional Long Short Term Memory (Bi-LSTM) is one of the recurrent models that can handle a long sequence of text. It has a unique feature called forget gate to handle its memory from a long sequence. We intend to use deep learning approach to solve this task without a need for any handcrafted rules. We want to contribute our work to help other research on natural language in Bahasa Indonesia by providing part of speech label as a contextual feature from plain text.

II. RELATED WORKS

There was some research on POS tagging in Bahasa Indonesia. Wicaksono et al. [7] proposed Hidden Markov Model for POS tagging. Pisceldo et al. [8] evaluate two probabilistic models, which are Conditional Random Fields and Maximum Entropy. These models rely on machine learning-based model. The other method is by using a rule-based approach. Rashel et al. [9] use a rule-based approach to build the POS Tagger for the Indonesian language. Fu et al. [10] compare several methods to build part of speech corpus for the Indonesian language. Nurwidyantoro et al. [11] use MapReduce to parallelize maximum entropy in the Indonesian part of speech tagging.

Recently, research on other languages also uses machine learning or even the rule-based approach in predicting part of speech labels. Màrquez et al. [12] use the decision tree as their main classifier. Maximum entropy is also tested on the English language by A. Ratnaparkhi [13]. Toutanova et al. [14] use Cyclic Dependency Network to do part of speech tagging. This network reads input in two ways (bidirectional), which are front to end and end to front. Brants [15] use a different approach by using a statistically-based model as the main method. Besides predicting part of speech on a single language, Plank et al. [16] try to predict part of speech in a multilingual environment. Their research uses one of deep learning model, namely bidirectional long short-term memory for predicting part of speech label.

Related research on other language using Bi-LSTM for the Arabic language [20]. This research using BLSTM approach. The accuracy result using Bi-LSTM Approach is 91%. The other research is about part of speech tagging for unknown words using support vector machine (SVM) [21] achieves accuracy 97.1%. The other related research is using Conditional Random Fields (CRF) for Tamil Part of Speech

Tagging and Chunking [22], this method achieves accuracy of about 89.18%.

III. Research Framework

Fundamental techniques in Natural Language Processing (NLP) include sequence labeling, word modeling, back off, and evaluation. This technique is useful in many fields and tagging gives us a simple context for presenting it. Tagging is a typical step in pipelines following tokenization. The process of classifying word into part of speech and providing appropriate labels is known as part of speech tagging or simply tagging. Part of speech is also known as word class or lexical category. A collection of tags is called a tag set.

A. Dataset Preparation

We use Indonesian part of speech dataset from Universitas Indonesia [9]. This dataset is annotated manually by a human annotator. The data contained in the corpus is obtained from Pan Asia Networking Localization network (PANL10n). This dataset consists of 10.000 sentences which built from 256.683 tokens. This dataset is available online.

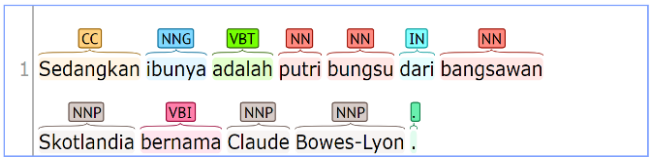

Fig. 1. Dataset Example.

Fig. 1 is the visualization from one example of the dataset. The input is a set of tokens. Word or punctuation are referred to as a token. Every sentence consists of multiple tokens, either word or punctuation mark. It is assumed that there is no relation between tokens. The output is a set of tokens with its POS label. Fig. 1 visualizes the output of the example sentence. We use the BRAT Rapid Annotation Tool [A] for dataset visualization.

B. Part of Speech in Indonesian Language

TABLE I. Indonesian Part of Speech Tag Sets

Tag	Description	Example
CC	Coordinating conjunction	dan, atau
CD	Cardinal number	satu, 15, sepertiga
OD	Ordinal number	ke-4, pertama
DT	Determiner, article	para, si
FW	Foreign word	terms and conditions
IN	Preposition	dengan, di, oleh, untuk
JJ	Adjective	panjang, jauh, bulat
MD	Modal and auxiliary verb	boleh, harus, perlu
NEG	Negation	tidak, belum, jangan
NN	Noun	bawah, sekarang
NNP	Proper noun	Laut Jawa, Indonesia
NND	Classifier, partitive, and measurement noun	orang, ton, helai, lembar
PR	Demonstrative pronoun	ini, situ, sini
PRP	Personal pronoun	saya, kamu, kalian
RB	Adverb	sangat, justru, segera
RP	Particle	-pun, -lah, -kah
SC	Subordinating conjuction	sejak, jika, sebab, maka
SYM	Symbol	IDR + % @
UH	Interjection	ayo, mari, hai
VB	Verb	pergi, bekerja, tidur
WH	Question	siapa, kenapa, di mana
X	Unknown	statemen
Z	Punctuation	‘ “ ? . !

Bahasa Indonesia has five important parts part of speech, namely: verb, adjective, adverb, noun, and verb for function words. Nouns can be divided into subcategories such as countable and uncountable common nouns, common genitive nouns, proper nouns, and some variations of pronouns. Function words can be further divided into several subcategories such as preposition, conjunction, interjection, article, and particle.

The final result of the POS Indonesia tag set from the research conducted by Dinakaramani et al. [18] produced 23 POS tags from the Indonesian corpus of more than 250.000 lexical tokens. A summary of 23 POS Tags can be seen from Table I. In the Indonesian table the POS Tag Sets above can be classified into several categories of tags such as noun (NN, NEG, NND, PR, PRP), verb (VB, MD, adjective (RR, RB, JJ adverb (description) and others (Z, X, SYM, UH, SC, CC, CD, OD, DT, FW, IN).

C. Bidirectional Long Short-Term Memory

Bidirectional Long Short-Term Memory (Bi-LSTM) or also known as the Bidirectional Recurrent Neural Network (Bi-RNN) method was introduced to increase the amount of input information that exists in a network. BRNN will be very useful if context input is needed. LSTM is designed to solve gradient problems that disappear and the first to introduce the gating mechanism.

The LSTM architecture explicitly divides the vector into two parts, where one half is treated as a "memory cell" and the other is working memory. Memory cells are designed to preserve memory, and also gradient errors, all the time, and are controlled through differentiable gating components that smoothly stimulate logical gates. At each input state, the gate is used to decide how many new inputs must be written in the cell's memory and how much content the cell's memory must be forgotten. LSTM is currently the most successful type of RNN architecture and is responsible for many sequences of state of the art modeling.

D. Model Architecture

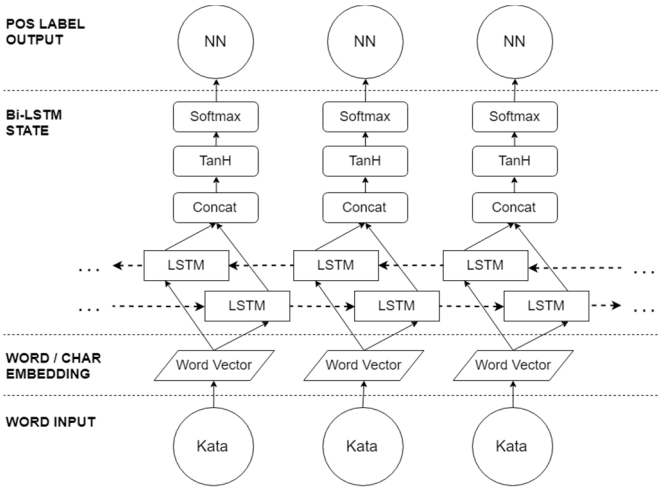

Fig. 2. Model Architecture

In sequential data labeling, Bi-LSTM plays an important role to capture sequential information. By readings from two directions, this model can capture more information than a

single sequence. Illustration of the Bi-LSTM architecture can be seen in Fig. 2. Word input from the illustration above are the words that the label will predict. In the word or char embedding section, use word2vec skip-gram model to present the word to a vector of numbers. Because word2vec cannot present words that have not been given before, a char embedding is performed to present the new word.

Fig. 2 visualize the Bi-LSTM state, which holds the sequence learning process from the previous input embedding results. There are two LSTM cells, namely Forward LSTM and Backward LSTM. The results of these two LSTMs are then skipped and passed the TanH activation function, then a probability distribution per class is generated using the Softmax function. The last process is the POS label output to determine which class is the output of the time step based on its probability distribution. This POS label output is a POS label from the current input word.

We use a word or char embedding to represent a word into a vector as input. This vector is trained first at the training stage. If there is a word tested and not included in the train set, then the word is represented by char embedding. This word or char embedding is trained together with the main model. Bi-LSTM is used to capture feature sequences from forward and backward from a sentence. Fig. 2 shows the combination of each time step to the activation function using TanH and the Softmax as prediction layer. Softmax is used to guess POS word labels using probability distribution.

E. Training Optimizer

The good optimization algorithm can speed up the training phase and save computation cost. Adam (Adaptive Moment Estimation) [19] is used because it is very efficient, uses less memory, is good for noisy gradients, is very suitable for large datasets, and is invariant for diagonal rescale of gradients. Adam is different from SGD (Stochastic Gradient Descent), which has a single learning rate for all weight updates and learning rates that do not change during the training phase.

Algorithm 1. Adam Optimizer
1. # *Hyperparameter Initialization*
2. $\alpha = 0.001$
3. $\beta_1 = 0.9$
4. $\beta_2 = 0.999$
5. $\epsilon = 10^{-8}$
6. $\theta_0 = \text{GetGradient}(0)$
7. $m_0 = 0$
8. $v_0 = 0$
9. $\text{t} = 0$
10.
11. # *Weight Updates*
12. WHILE θ_t NOT CONVERGED DO:
13. $\text{t} = \text{t} + 1$
14. $g_t = \text{GetGradient}(\theta_{t-1})$
15. $m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g_t$
16. $v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * {g_t}^2$
17. $\hat{m}_t = m_t / \left(1 - {\beta_1}^t\right)$
18. $\hat{v}_t = v_t / \left(1 - {\beta_2}^t\right)$
19. $\theta_t = \theta_{t-1} - \alpha * \hat{m}_t / \left(\sqrt{\hat{v}_t} + \epsilon\right)$
20. return θ_t

In Algorithm 1, we can see how Adam optimizer works. Adam is theoretically better than Stochastic Gradient Descent (SGD) in carrying out weight updates. There is momentum in the Adam algorithm weight update formula. Adam combines adaptive learning and momentum. This optimizer calculates an adaptive learning rate for each weight that will be updated. The usefulness of the Adam optimizer is to update the weight of all neurons, including the LSTM. The aim is to use Adam optimizer so that the weights of neurons can converge to reach global optima. The detail of how Adam optimizer works can be seen in Algorithm I.

IV. Experimental Analysis

We split POS dataset from Dinakaramani et al. [18] with the default composition of 70% training data (7.000 sentences) and 30% testing data (3.000 sentences). This dataset contains 256.683 tokens. We use CoNLL Evaluator tool from CoNLL-2000 shared task to calculate the F1 score for all of our experiments. This evaluator can be used specifically for measuring sequential pattern in part of speech. Our experiments mainly conducted on two aspects, which are training optimizer and training size.

The first experiment tried to proof that Adam is slightly better than common optimizer like SGD. Based on our experiments on these two optimizers, we found that the overall performance of Adam is better than using SGD. F1 Score result from Adam is 4.25% much better than the F1 score from SGD. The second experiment tried to give a clear view of the Bi-LSTM ability in handling different training size.

TABLE II. Experimental Results

Iteration	60-40	70-30	80-20
10	90.56	96.72	96.52
20	96.67	96.83	96.61
30	96.57	96.84	96.61
40	96.65	96.81	96.78
50	96.56	96.89	96.78
60	96.67	96.85	96.75
70	96.74	96.93	96.65
80	96.59	96.93	96.82
90	96.72	96.68	96.82
100	96.78	96.92	96.79

Table II contains the experiment results of different composition versus a number of iteration. The horizontal axis states the percentage composition of the train set and test set. The vertical axis states the number of training iterations. For the results of 10 iterations with 70% train set and 30% test set, our model scored 96.72% of F1 score. More iterations mean higher F1 score. However, the increase is not significant in certain iterations. We found that using more training data did not bring linear increasing. This phenomenon happened when the model is too overfitting with the training data.

TABLE III. Indonesian POS Tag set F1 Score Result

No.	Tag	F1 Score
1	**CC**	99.58%
2	**CD**	99.76%
3	**DT**	99.56%
4	**FW**	95.22%
5	**IN**	98.85%
6	**JJ**	97.89%
7	**MD**	99.32%
8	**NEG**	99.22%
9	**NN**	98.95%
10	**NND**	98.08%
11	**NNP**	98.34%
12	**OD**	97.73%
13	**PR**	99.78%

No.	Tag	F1 Score
14	**PRP**	99.91%
15	**RB**	98.15%
16	**RP**	96.30%
17	**SC**	98.29%
18	**SYM**	99.31%
19	**UH**	91.67%
20	**VB**	99.46%
21	**WH**	93.12%
22	**X**	84.00%
23	**Z**	99.89%

Table III shows the result of Indonesian part of speech tag sets that are used for this experiment. Open class word which divided as Verb achieved score of 99.46%, Noun achieved average score of 98.45%, Adjective achieved score of 97.89%, and Adverb achieved score of 98.15%. On the other side, Close class word achieved average score of 97.15%.

V. CONCLUSION

Bidirectional LSTM is used to predict the POS tags for the words for the Indonesian language. Using Adam optimizer can speed up the training phase and save on computing costs. The usefulness of the Adam Algorithm is to update the weight of all neurons including the Bi-LSTM. The aim is to use Adam optimizer so that the weights of neurons can converge to reach global optima. The results of the performance of the Part of Speech labeling model using Bidirectional LSTM were able to achieve an F1 score of 96.92%, which are open class about 98.49% and closed class about 97.15%.

REFERENCES

[1] P. Alva and V. Hegde, "Hidden markov model for POS tagging in word sense disambiguation," in *International Conference on Computational Systems and Information Systems for Sustainable Solutions*, 2016, pp. 279-284.

[2] K. Chen and H. Chen, "Extracting noun phrases from large-scale texts: a hybrid approach and its automatic evaluation," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 2011, pp. 234-241.

[3] N. Ueffing and H. Ney, "Using POS information for statistical machine translation into morphologically rich languages," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 2003, vol. 1, pp. 347-354.

[4] W. Khong, L. Soon, H. Goh, and S. Haw, "Leveraging part-of-speech tagging for sentiment analysis in short texts and regular texts," in *Short Texts and Regular Texts: 8th Joint International Conference*, 2018, pp. 182-197.

[5] C. Nicholls and F. Song, "Improving sentiment analysis with part-of-speech weighting," in *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, 2009, pp. 1592-1597.

[6] G. Móra and V. Vincze, "Joint part-of-speech tagging and named entity recognition using factor graphs," in Sojka P., Horák A., Kopeček I., Pala K. (eds) Text, Speech and Dialogue. TSD 2012. Lecture Notes in Computer Science, vol. 7499, pp. 232-239, 2012.

[7] A. Farizki Wicaksono and A. Purwarianti, "HMM based part-of-speech tagger for bahasa Indonesia," in *Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop*, 2010.

[8] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic Part of Speech Tagging for Bahasa Indonesia," 2009.

[9] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian rule-based part-of-speech tagger," in *International Conference on Asian Language Processing (IALP)*, 2014.

[10] S. Fu, N. Lin, G. Zhu, and S. Jiang, "Towards Indonesian part-of-speech tagging: corpus and models," in *International Conference on Language Resources and Evaluation*, 2018.

[11] A. Nurwidyantoro and E. Winarko, "Parallelization of maximum entropy POS tagging for bahasa Indonesia with mapreduce," in *International Journal of Computer Science Issues (IJCSI)*, vol. 9, issue 4, no. 2, 2012.

[12] L. Màrquez and H. Rodriguez, "Part-of-speech tagging using decision trees," in *Proceedings of the 10th European Conference on Machine Learning*, pp. 25-36, 1998.

[13] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Conference on Empirical Methods in Natural Language Processing*, pp. 133-142, 1996.

[14] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL 2003*, pp. 173-180, 2003.

[15] T. Brants, "TnT - A Statistical Part-of-Speech Tagger," in *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pp. 224-231, 2000.

[16] B. Plank, A. Søgaard, and Y. Goldberg, "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 412–418, 2016.

[17] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: a web-based tool for NLP-assisted text annotation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102-107, 2012.

[18] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," in *International Conference on Asian Language Processing (IALP)*, 2014.

[19] D. P. Kingma, J. Ba, "Adam: a method for stochastic optimization", in *3rd International Conference for Learning Representations*, 2015.

[20] R. Alharbi, W. Magdy, K. Darwish, A. A. Ali, and H. Mubarak, "Part-of-Speech Tagging for Arabic Gulf Dialect Using Bi-LSTM".

[21] T. Nakagawa, T. Kudoh, and Y. Matsumoto, "Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines".

[22] S. L. Pandian and T. V. Geetha, "CRF Models for Tamil Part of Speech Tagging and Chunking.