SPOK-ID: A Deep Learning Model to Identify the Sentence Structure of Indonesian Language

R S Panigas, Y S Purnomo WP\* and A J Santoso

Informatics Department, Faculty of Industrial Technology, Universitas Atma Jaya Yogyakarta, Babarsari 43, Yogyakarta, Indonesia, 55281

\*sigit.purnomo@uajy.ac.id

**Abstract**. Natural Language Processing (NLP) research has overgrown over the last few years. The previous NLP works mainly on high-resource languages such as English, French, Spanish, and Chinese. This research tried to construct an NLP model with a specific task of identifying the sentence structure of the Indonesian languages. All models use the sequence-labelling system used for several high-resource language tasks. The experiments show that: (1) several models that achieve a good result on high-resource language tasks can be used to classify the phrase structure of the Indonesian language, (2) the best 75-cell Bidirectional Long Short-term Memory (BiLSTM) model achieves 81.42% of the F1 score.

1. Introduction

Artificial neural network-based models such as Long Short-term Memory (LSTM) and Convolutional Neural Networks (CNN) have proven effective in NLP-specific tasks. In addition to model availability, the availability of linguistic data sets is essential for sufficient NLP growth. NLP's design also has significant limitations, namely the lack of resources and models focusing on low-resource languages such as Indonesian, Punjabi, and Swahili (1). This limitation means that a model with particular tasks for high-resource languages can have very different results on some particular tasks for low-resource languages.

Research on the use of NLP for Indonesian has been carried out, such as for text classification (2–4), summarizing text or documents (5–7), name-entity recognition (NER) (8,9), and part-of-speech tagging (10). By far, there is no research or experiment yet on the task of defining the "Subyek-Predikat-Obyek-Keterangan" (SPOK) sentence structure in Indonesian using a deep learning model. This research aims to develop NLP models for the task of defining the SPOK sentence structure in Indonesian.

This study tried to use variations in the LSTM model used in the NER model (11). This study has trained and tested the model on a relatively small dataset containing 3627 sentences in Indonesian. Experiments show that: (1) BiLSTM-CRF and LSTM-CRF models can be used to predict sentence structure in the Indonesian language, (2) it can achieve a score of 81.42 percent of F1 points on the test set.

1. Method

## Dataset Collection and Annotation

The first phase of this research is data collection and annotation. For the sentence dataset, this research collects Indonesian language sentences from several online newspapers in Indonesia, such as Kompas, Historia, and Beritagar. The sentence obtained is then manually annotated by all sentence tokens with the correct tag. For example, the S tag means "Subyek", P ("Predikat"), O ("Obyek"), K ("Keterangan"), and Pel ("Pelengkap"). IOB2 coding scheme standard (12) use to convert each tag. For example, from the sentence “Bu Juni membuat mainan dengan kertas” the IOB2 tags will be B-S for token “Bu”, I-S for token “Juni”, B-P for token “membuat”, B-O for token “mainan”, B-K for token “dengan”, and I-K for token “kertas”.

The dataset is still relatively small and contains 3627 sentences. This study uses 70% of the overall dataset for training purposes. To validate the model, it uses 20% of the total data set. 10% of the entire dataset used to assess the performance of the proposed model. Table 1 shows the number of the pattern of each sentence structure. The distribution of the pattern of each sentence structure in the dataset is reasonably well balanced.

Table 1 Number of each sentence structure pattern

| Sentence Structure Pattern | Number of Datasets |
| --- | --- |
| S-P | 603 |
| S-P-O | 604 |
| S-P-Pelengkap | 605 |
| S-P-K | 605 |
| S-P-O-K | 605 |
| S-P-Pelengkap-K | 605 |

## Deep Learning Model Development

In the implementations, all of the LSTMs and BiLSTMs have two layers. Dropout (13) used as a regularizer for each LSTM output layer. The experiment also uses a recurrent connection regularizer as proposed by Gal and Ghahramani (14). The Clip norm was used to resolve the exploding gradient problem by placing a constraint on the gradient law (15). Forget gate bias was set to 1 during initialization as suggested by Jozefowicz et al. (16) to improve LSTM efficiency. For the initialization method, these experiments use Xavier's initialization (17). Conditional Random Fields (CRF) models (18) used as a tag decoder. The architecture of the LSTM-CRF and BiLSTM-CRF showed in Figure 1 and Figure 2.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Figure 1** LSTM*-*CRF model’s architecture |  | **Figure 2** BiLSTM-CRF model’s architecture. |

Until the sentence feeds into the model, all the terms will set to lowercase and convert into an integer representation of the word vocabulary. Terms outside the dictionary represent by UNKNOWN TOKEN, which has index 1. The index set to 0 as "Masking" integer representation since all input sequences set to 0. All input sequences load at a fixed length of 100. The word embedding volume is 300. For the word embedding, these experiments use pre-trained embedding trained in Wikipedia dump text (19) for the Indonesian language.

The LSTMs and BiLSTMs were designed to have 50 and 75 units, each with a 50-unit time-distributed pre-output layer. The final layer uses the CRF layer. The model trained at 50, 75, 100, 150, 200, and 500 epochs for each model because this study will measure the model output for each epoch. For the optimizer, these experiments use ADAM (20) with a learning rate of 0.001 and a batch size of 64. All of the models evaluated with the F1 score.

1. Result and Discussion

Table 2 and Table 3 shows the model performance of the test set for each observed epoch. All models are tested using a test set containing 364 sentences. BiLSTM achieved the highest score in F1 with two layers and 75 cells trained in 500 epochs. All BiLSTM models have beaten all LSTM models and can produce a state of the art accuracy on this dataset like what Huang *et al.* (11). All models except LSTM with two layers and 75 cells achieved their highest score at the 500th epoch. The highest F1 score that can be reached by LSTM models is 80.07 points. Overall, the models can be used to predict the structure of the Indonesian language sentence.

**Table 2 F1 scores on a test set of each LSTM model for each observed epoch.**

| Model | Epoch | Precision | Recall | F1 |
| --- | --- | --- | --- | --- |
| **LSTM 2 Layer 50 cell** | 50 | 76.71 | 76.20 | 76.45 |
| 75 | 76.77 | 76.48 | 76.42 |
| 100 | 77.92 | 77.27 | 77.59 |
| 150 | 78.78 | 79.01 | 78.89 |
| 200 | 77.72 | 77.87 | 77.79 |
| 500 | 79.33 | 78.50 | **78.92** |
| **LSTM 2 Layer 75 cell** | 50 | 75.16 | 74.98 | 75.07 |
| 75 | 79.08 | 78.90 | 78.99 |
| 100 | 79.56 | 79.18 | 79.37 |
| 150 | 80.19 | 79.61 | 79.90 |
| 200 | 80.05 | 80.08 | **80.07** |
| 500 | 78.87 | 77.98 | 78.42 |

**Table 3 F1 scores on a test set of each BiLSTM model for each observed epoch.**

| Model | Epoch | Precision | Recall | F1 |
| --- | --- | --- | --- | --- |
| **BiLSTM 2** **Layer 50 cell** | 50 | 76.27 | 76.37 | 76.32 |
| 75 | 79.41 | 79.09 | 79.25 |
| 100 | 79.72 | 79.65 | 79.69 |
| 150 | 79.70 | 79.95 | 79.82 |
| 200 | 80.30 | 79.69 | 79.99 |
| 500 | 81.10 | 81.04 | **81.07** |
| **BiLSTM 2 Layer 75 cell** | 50 | 78.34 | 78.22 | 78.18 |
| 75 | 79.07 | 78.92 | 79.00 |
| 100 | 80.88 | 80.83 | 80.86 |
| 150 | 80.90 | 80.91 | 80.91 |
| 200 | 80.43 | 80.23 | 80.33 |
| 500 | 81.36 | 81.62 | **81.42** |

## LSTM Model

Both models often misclassify an I-Pel label with an I-O label (Figure 3). LSTM with two layers and 75 cells misclassified I-Pel with I-O 26% of the time. LSTM's 50-unit cells misclassified I-Pel with I-O 22 percent of the time. Both models appear to have a poor understanding of B-Pel and I-Pel labels because both models correctly classified around 50 percent of these labels compared to other labels. Such two versions have no significant differences between them. Both models have a good understanding of the "Subyek" label and achieved about 90 percent true positive on the B-S label and about 85 percent true positive on the I-S label.



Figure 3 LSTM Model’s Confusion Matrix.

## BiLSTM Model

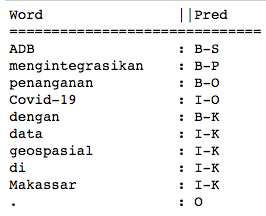
Neither of the two BiLSTM models has a significant difference either (Figure 4). All models are often

misclassified B-Pel and I-Pel with B-O and I-O, but the BiLSTM models have a lower misclassification rate for these labels. Both BiLSTM models, just like all LSTM models, also have a rather poor understanding of predicting B-Pel and I-Pel labels compared to other labels despite BiLSTM, with a lower rate of misclassification compared to LSTM models. Apart from the B-Pel and I-Pel labels, BiLSTM sometimes also misclassifies the I-Pel label with an I-K label with a false-positive rate of about 20 percent.



Figure 4 BiLSTM model’s confusion matrix

Overall, all models have very similar performance and a shortcoming in defining the sentence structure of the Indonesian language task. The most likely cause of misclassification between the tag Pel (“Pelengkap”) and O (“Obyek”) is due to the same structural position right after the mark P (“Predikat”). It presumed that without additional features to differentiate Pel (“Pelengkap”) and O (“Obyek”) names, both models could only rely on word embedding to identify them apart from learning their positional structure. Figure 5 shows the result of the BiLSTM 2 layer model with 75 cells trained on 500 epoch to predict the structure of new sentence that not exist in the dataset “ADB mengintegrasikan penanganan Covid-19 dengan data geospasial di Makasar.”.



**Figure 5 Result of the Prediction using the Best Model**

1. Conclusion

This study reported model performance on the identification of the Indonesian sentence structure task. The two-layer, 50-unit BiLSTM achieved the highest F1 score compared to the other model. Overall, all models have produced a good result in the identification of the sentence structure of the Indonesian

language task. One recognized disadvantage of the design is its poor performance in separating Pel (“Pelengkap”) and O (“Obyek”) marks.

To improve the model for future studies, we recommend changing the IOB2 coding scheme to the BILOU coding scheme because BILOU significantly outperforms the widely adopted IOB2 or BIO (21) to increase the quality of the model. Using another model such as the attention-based model (22) that have achieved state-of-the-art results on some NLP tasks may help improve performance in identifying the sentence structure of the Indonesian language task.

**Acknowledgments**

The authors gratefully thank the Informatics Department, Universitas Atma Jaya Yogyakarta, who support this research by providing the computing resources (https://grid.uajy.ac.id).

References

1. Hirschberg J, Manning CD. Advances in natural language processing. Science (80- ). 2015;349(6245):261–6.

2. Saputra AC, Sitepu AB, Stanley, Yohanes Sigit PWP, Sarto Aji Tetuko PG, Nugroho GC. The Classification of the Movie Genre based on Synopsis of the Indonesian Film. Proceeding - 2019 Int Conf Artif Intell Inf Technol ICAIIT 2019. 2019;201–4.

3. Yovellia Londo GL, Kartawijaya DH, Ivariyani HT, Yohanes Sigit PWP, Muhammad Rafi AP, Ariyandi D. A Study of Text Classification for Indonesian News Article. Proceeding - 2019 Int Conf Artif Intell Inf Technol ICAIIT 2019. 2019;205–8.

4. Rusli A, Young JC, Iswari NMS. Identifying fake news in indonesian via supervised binary text classification. Proc - 2020 IEEE Int Conf Ind 40, Artif Intell Commun Technol IAICT 2020. 2020;86–90.

5. Severina V, Khodra ML. Multidocument Abstractive Summarization using Abstract Meaning Representation for Indonesian Language. In: 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA). IEEE; 2019. p. 130–3.

6. Jiwanggi MA, Adriani M. Topic Summarization of Microblog Document in Bahasa Indonesia using the Phrase Reinforcement Algorithm. Procedia Comput Sci [Internet]. 2016;81(May):229–36. Available from: http://dx.doi.org/10.1016/j.procs.2016.04.054

7. Widjanarko A, Kusumaningrum R, Surarso B. Multi document summarization for the Indonesian language based on latent dirichlet allocation and significance sentence. 2018 Int Conf Inf Commun Technol ICOIACT 2018. 2018;2018-Janua:520–3.

8. Wintaka DC, Bijaksana MA, Asror I. Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF. Procedia Comput Sci [Internet]. 2019;157:221–8. Available from: https://doi.org/10.1016/j.procs.2019.08.161

9. Gunawan W, Suhartono D, Purnomo F, Ongko A. Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs. Procedia Comput Sci [Internet]. 2018;135:425–32. Available from: https://doi.org/10.1016/j.procs.2018.08.193

10. Yuwana RS, Yuliani AR, Pardede HF. On part of speech tagger for Indonesian language. Proc - 2017 2nd Int Conf Inf Technol Inf Syst Electr Eng ICITISEE 2017. 2018;2018-Janua:369–72.

11. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. 2015; Available from: http://arxiv.org/abs/1508.01991

12. Tjong EF, Sang K, Veenstra J. Representing text chunks. In: Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics. 1999. p. 173–9.

13. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929–58.

14. Gal Y, Ghahramani Z. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016. p. 1027–35.

15. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on Machine Learning. 2013. p. 1310–8.

16. Jozefowicz R, Zaremba W, Sutskever I. An Empirical Exploration of Recurrent Network Architectures. In: Proceedings of the 32nd International Conference on Machine Learning. 2015. p. 2342–50.

17. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) [Internet]. 2010. p. 249–56. Available from: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

18. Lafferty J, Mccallum A, Pereira F. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. 2001. p. 282–9.

19. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Trans Assoc Comput Linguist. 2017;5:135–46.

20. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR) [Internet]. 2015. Available from: http://arxiv.org/abs/1412.6980

21. Lev Ratinov, Roth D. Design Challenges and Misconceptions in Named Entity Recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009) [Internet]. Association for Computational Linguistics; 2009. p. 147–155. Available from: http://www.usaidbest.org/docs/Burundi\_2013\_Report\_Final\_508.pdf

22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. p. 6000–10.