Journal of Engineering Science and Technology
Special Issue on AASEC2022, August (2022) 62 - 68

CANONICAL SEGMENTATION FOR JAVANESE-INDONESIAN NEURAL MACHINE TRANSLATION

SRI HARTATI WIJONO[1,2,*], KURNIAWATI AZIZAH[2], WISNU JATMIKO[2]

[1] Department of Informatics, Faculty of Science and Technology,
Sanata Dharma University, Yogyakarta, Indonesia
[2] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
*Corresponding Author: tatik@usd.ac.id

Abstract

Corpus-based Neural Machine Translation (NMT) has achieved remarkable results in many high-resource language pairs and becomes widely used in recent years. However, it generates many out-of-vocabulary (OOV) words in the low-resource parallel corpus, especially for agglutinative language pairs such as Javanese to Indonesian translation. This paper proposes a canonical word segmentation and a linguistic feature tag to be incorporated in a Transformer-based NMT for translating Javanese into Indonesian. The word segmentation is to increase vocabulary frequency of affixed words that rarely appear, while the feature tag is to help the learning process and generates translation output. This research is conducted in two stages. First, we explore some Javanese segmentation approaches using a Transformer-based encoder-decoder to find the best segmentation model. As for the Indonesian language, we use MorphInd to do corpus segmentation. Second, we conduct experiments on NMT by applying canonical segmentation and feature tag resulted in the first stage as the input to the encoder and decoder. Our experiments show that the best canonical segmentation is the one that uses character-level inputs concatenated with feature tags that includes affixes and root words. It achieves an accuracy value of 84.20% of all occurrences and 56.09% of canonical segmentation. This study also reports that it reaches a F1 score of 92.78% and 96.35% for all words and canonical segmentation, respectively. As for the NMT experiments, the results show that the proposed canonical segmentation and affixes/root word feature tag applied to NMT model improves the translation performance. Our best model increases the BLEU score by 3.55 points compared to baseline model using words as inputs. It also increases as much as 2.57 BLEU points compared to baseline model using BPE segmentation.

Keywords: Canonical segmentation, Javanese-Indonesian NMT, Linguistic feature tag, Neural machine translation..

1. Introduction

Corpus-based Machine Translation is a popular approach to overcome the problem of providing rules and linguistic resources that are mandatory in rule-based Machine Translation. With the rapid growth of deep learning, Neural Machine Translation (NMT) [1, 2] has become is a widely used corpus-based approach and has obtained satisfactory results in translating many language pairs. However, a good translation results need a large amount of parallel corpus for the NMT training process. Therefore, it becomes problematic in many parallel corpora that are used to translate low-resource languages, such as Javanese and Indonesian. Limited resource parallel corpus causes many out-of-vocabulary (OOV) to appear. One way to reduce OOV is using word segmentation [3, 4]. For example, Javanese words "nyapu" *(sweep)* and "disaponi" (*be swept)* have the same root word vocabulary, namely "sapu".

Some NMT studies used various word segmentation methods such as Byte-Pair Encoding (BPE), GenSeg based on root, prefix, and postfix [4], and Morfessor [5] . Several studies explored the effectiveness of word segmentation on NMT and showed that word segmentation concerning linguistic properties has a better BLEU value compared to the one without linguistic attributes [4, 6-8].

In agglutinative languages, the word segmentation that considers linguistic properties can be done by splitting words into their constituent morphemes: affixes and root words. Decomposing words into morphemes is called morphological segmentation. Cotterell et al. [9] categorizes morphological segmentation into surface segmentation and canonical segmentation. Unlike surface segmentation, canonical segmentation can get the root word even though there is an allomorph in the word formation process. The difference between both segmentation types can be illustrated using Javanese word "nyaponi" (*sweep*) that has morpheme alteration from root word "sapu":

Canonical segmentation: nyaponi => ny+sapu+oni

Surface segmentation: nyaponi => ny+apo+ni

Our previous study on Javanese canonical segmentation applied a character-level input to a Transformer-based encoder and decoder model [10]. The model used affix characters as a unit to be inputted to the decoder during training process [11] . The results showed an accuracy of 42.7% for affixed words and 21.97% for canonical occurrences with affixes. This canonical segmentation model still needs improvements so that it can be used to increase BLEU score in NMT system for Javanese to Indonesian.

The work of Sennrich and Haddow [12] showed that adding features to sub words or words in the encoder can help disambiguate the learning process and improve translation results. During the training process Garcia-Martinez et al. [13] added features to the input of the decoder to help generate output. We combine both ideas and add the feature tags to the inputs of both encoder and decoder for canonical segmentation and translation processes.

We collect parallel corpus to examine the translation process from Javanese to Indonesian. This parallel corpus is canonically segmented first to overcome the OOV problem that generally appears on low-resource languages. We use MorphInd [14] for canonical segmentation of Indonesian. As for Javanese we use program

that applies allomorph rule as in Wijono et al. [10] then manually check based on Javanese dictionary. We use this silver standard parallel corpus for training and testing our canonical segmentation and NMT models. We conduct the research in two stages. First, we conduct experiments for Javanese canonical segmentation to improve the results of our previous research [10]. We propose a Transfomer-based canonical segmentation model that uses the affix, root word feature tag concatenated for each input character to the encoder and decoder. Second, the sub words resulted from the canonical segmentation are concatenated with the feature tag as input to the encoder and decoder of the NMT model.

Our proposed canonical segmentation model can improve the accuracy and F1 score over our previous study [10] and baseline [15]. The baseline is character-based encoder-decoder input with no tags. Compared to the previous study, the proposed model increases the accuracy by 28.51 points for all occurrences and 34.21 points for canonical segmentation words. The F1 score also improves by 9.36, and 11.7 for all occurrences and canonical segmentation words, respectively. Compared to the baseline model, our model improves the accuracy by 24,14 points for all occurrences and 56.09 points for canonical segmentation words. It also improves the F1 score by 4.35 and 13.99 for all occurrences and canonical segmentation words, respectively. The experiments also show that NMT using concatenation of canonical segmentation sub-words and feature tag improves BLEU by 3.55 points over word-based and 2.57 points over BPE input.

Overall, the contribution of this research can be summarized in the following points (1) to propose Transformer-based canonical segmentation using concatenation of character and affix, root word feature tags; (2) to propose Transformer-based NMT that integrates canonical segmentation sub-words and affix, root word feature tags; and (3) to provide silver standard parallel corpus for Javanese canonical segmentation and Javanese-Indonesian Machine translation.

2. Material and Methods

This study improves translation results by reducing OOV words in low resource NMT. We linguistically segment words into their constituent morphemes: affixes and root words. Rare words in the form of affixed words are segmented into root words to increase the frequency and make it possible to translate unknown words. For example, “nyaponi”, “disaponi”, “nyapu” (*sweep)* have the same root word, namely “sapu”. The segmentation process uses canonical segmentation because in Javanese and Indonesian many words experience allomorphs in their formation.

We conduct the research in two stages: canonical segmentation and Neural Machine Translation. We start by showing the characteristics of morphological segmentation in Javanese, which can be used for other languages with the same characteristics. Figures 1 and 2 show the architecture of canonical segmentations and canonical segmentation training program, respectively.

3. Results and Discussion

3.1. Canonical segmentation

Table 1 shows that our proposed approach that uses feature tags (PC2) gives a significant improvement compared to the baseline model (BC) that uses a sequence of characters [15] and our previous study (PC1) that uses affixes as a unit [10]. Our

approach that concatenates affixes and root word feature tags to the input characters (PC2) can improve accuracy. The evaluation for all words increases the accuracy by 28.42 points compared to our previous study (PC1). It also gets the highest accuracy compared to the other models.

Overall, our proposed feature tags approach, PC2, gives best accuracy than the model without feature tags, PC1 and BC. The consistent results can also be seen in the F1 score and recall value. Meanwhile, the proposed feature tags of affixes and root words (PC2) achieves the highest precision values, except for non-affixed words. Overall, using affixes and root word features can improve the accuracy, F1, precision, and recall values rather than using a sequence of characters or affix units.

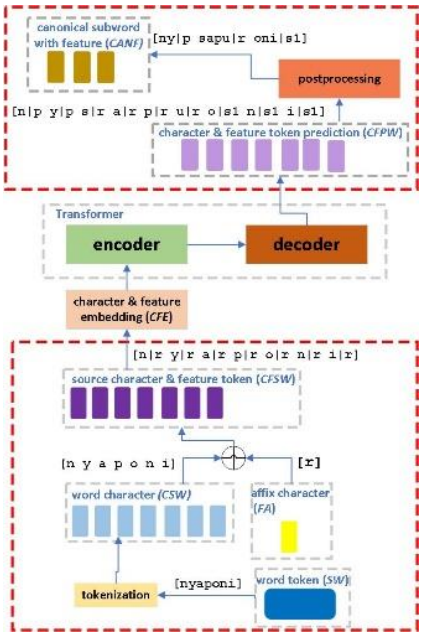

Fig. 1. Architecture of canonical segmentation.

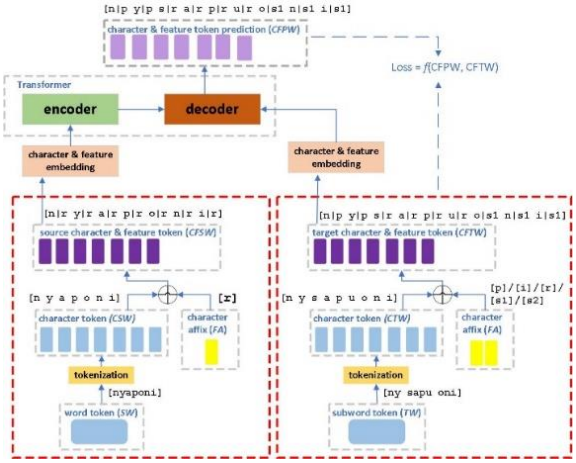

Fig. 2. Canonical segmentation training process.

Table 1. Canonical segmentation experimental results using dataset CCI.

Model		Acc (%)	P (%)	R (%)	F1 (%)
All Word (4336 words)	BC[1]	60.15	99.01	80.46	88.36
	PC1	55.78	92.79	77.01	83.35
	PC2	**84.20**	**99.05**	**87.46**	**92.78**
Affixed Word (994 words)	BC[1]	3.80	99.19	68.16	88.82
	PC1	42.70	95.60	85.84	89.29
	PC2	**69.82**	**99.84**	**97.57**	**98.63**
Canonical Affixed Words (149 words)	BC[1]	0	98.50	72.63	82.63
	PC1	21.97	91.80	80.41	84.92
	PC2	**56.09**	**99.20**	**94.08**	**96.35**
Non-affixed Word (3193 words)	BC[1]	80.72	**98.95**	94.78	96.35
	PC1	60.53	91.70	87.67	88.83
	PC2	**90.27**	98.11	**99.40**	**98.59**

[1] Baseline

3.2. Neural machine translation

The experiment results on translating Javanese to Indonesian are shown in Table 2. Segmenting words in NMT using BPE (BMT2) increases BLEU by 0.98 points compared to NMT without any segmentation (BMT1). However, the NMT still requires a canonical segmentation process to deal with allomorph. It shows that segmentation in the form of affixes and root words (PMT1) increases BLEU by 1.11 and 0.13 points compared to BMT1 and BMT2, respectively. The concatenation of the affix feature tag to the canonical segmentation subword (PMT2) as the input to the encoder and decoder can further improve the BLEU score. It increases by 2.79 points, superior to BMT1. Meanwhile, the affixes and root word feature tag concatenated with subwords from canonical segmentation (PMT3) increased BLEU by 3.55 points compared to BMT1. In conclusion, the best NMT model is PMT3 which uses the affix and root word features tag in the subwords resulting from canonical segmentation.

Table 2. NMT experimental results.

Experiments	BLEU
BMT1[1]: Word	33.96
BMT2[1]: BPE	34.94
PMT1: canonical segmentation	35.07
PMT2: canonical segmentation with affixes feature tag	36.75
PMT3: canonical segmentation with affixes and root word feature tag	**37.51**

[1] Baseline

Since the training process uses low resource parallel data, the translation results cannot be perfect. However, many translation results, as in Fig. 3 show that the proposed method (PMT3) can return affixes and root words that the other models cannot. For example, from the search results on the corpus training, the words “nggadhekake” *(pawn)* and “nampani” *(receive)* only appear once, but the root words “gadhe” and “tampa” appear many times, either in the form of root words or other words with affixes. In case (a), PMT1 and PMT2 cannot translate the Javanese root word “gadhe” into Indonesian root word “gadai”, while PMT3 can. In case (b), the Javanese root word “tampa” can still be translated into Indonesian root word “terima” by PMT1 but gets the wrong suffix, while PMT2 and PMT3 can translate correctly. These experiments show that affix and root feature tags in canonically segmented data get better translation results.

Javanese	ning ora nganti ndadak nggadhekake blangkon barang [nggadhekake] => [ng + gadhe + kake] *(pawn)*
Indonesian	tapi tidak sampai menggadaikan blangkon [menggadaikan] => [meng + gadai + kan]
BMT1	tapi tidak perlu perjanjian
BMT2	tapi tidak perlu blangkon
PMT1	tapi tidak pake kipper
PMT3	tapi tidak akan menjual jasa untuk menggadaikan blangkon

(a)

Javanese	nanging pihak sing nampani isih nduweni janggelan [nampani] => [n + tampa + ni] *(receive)*
Indonesian	tapi pihak yang menerima masih mempunyai keraguan [menerima] => [meN + terima]
BMT1	tapi pihak yang gemar merokok bisa mengurangi
BMT2	tapi pihak yang menakutkan itu tidak baik
PMT1	tapi pihak yang diterima dengan dua puluh
PMT2	tapi pihak yang menerima sedikit saja
PMT3	namun pihak yang menerima lah

(b)

Fig. 3. Example of translated words that rarely appear in corpus training.

4. Conclusion

This study analyses Javanese to Indonesian Neural Machine Translation with a parallel corpus of limited resources that generates many out-of-vocabulary (OOV) words for low resource parallel data. We solve this problem by applying canonical segmentation to each Javanese and Indonesian word. In addition, to help the learning process and generate output in the training process, we combine feature tags in the form of affixes and root words for each input to the encoder and decoder. We use MorphInd for canonical segmentation for Indonesian. We develop resources for the Java language as they are unavailable. Canonical segmentation experiments show that adding feature tags in the form of affixes and root words combined with each character input to the encoder and decoder performs much better than character sets or affix units in previous studies. The Neural Machine Translation experiment shows that using canonical segmentation subwords combined with affix and root feature tags as input to the encoder and decoder can improve translation performance compared to the basic model using words and BPE. Our experiment also shows that the proposed method can produce translated words that the baseline model cannot. For future studies, we consider applying more linguistic information to NMT. We can also explore transfer learning from a corpus of high-resource languages with similar morphology and syntax.

References

1. Bahdanau, D.; Cho, K.; and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, 1-15.
2. Sutskever, I.; Vinyals, O.; and Le, Q.V. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, 3104-3112.
3. Sennrich, R.; Haddow, B.; and Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 1715-1725.

4. Zuters, J.; and Strazds, G. (2019). Subword segmentation for machine translation based on grouping words by potential roots. *Baltic Journal of Modern Computing*, 7(4), 500-509.
5. Creutz, M.J.P; and Lagus, K.H. (2006). Morfessor in the morpho challenge. *Proc. PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, Venice, Italy.
6. Huck, M.; Riess, S.; and Fraser, A. (2017). Target-side word segmentation strategies for neural machine translation. *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, 56-67.
7. Banerjee, T.; and Bhattacharyya, P. (2018). Meaningless yet meaningful: Morphology grounded subword-level NMT. *Proceedings of the Second Workshop on Subword/Character Level Models*, New Orleans, USA.
8. Chimalamarri, S.; and Sitaram, D. (2021). Linguistically enhanced word segmentation for better neural machine translation of low resource agglutinative languages. *International Journal of Speech Technology*, 24(4), 1047-1053.
9. Cotterell, R.; Vieira, T.; and Schütze, H. (2016). A joint model of orthography and morphological segmentation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, 664-669.
10. Wijono, S.H.; Alhamidi, M.R.; Hilman, M.H.; and Jatmiko, W. (2021). Canonical segmentation using affix characters as a unit on transformer for javanese language. *Proceedings of the 2021 6th International Workshop on Big Data and Information Security (IWBIS)*, Jakarta, Indonesia, 67-72.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; and Polosukhin, I. (2017). Attention is all you need. *Proceedings of Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA.
12. Sennrich, R.; and Haddow, B. (2016). Linguistic input features improve neural machine translation. *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, 83-91.
13. García-Martínez, M.; Barrault, L.; and Bougares, F. (2016). Factored neural machine translation architectures. *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington DC, 1-9.
14. Larasati, S.D.; Kuboň, V.; and Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. *Proceedings of the Second International Workshop on Systems and Frameworks for Computational Morphology, SFCM 2011*, Zurich, Switzerland, 119-129.
15. Moeng, T.; Reay, S.; Daniels, A.; and Buys, J. (2022). Canonical and surface morphological segmentation for nguni languages. *Proceedings of the Second Southern African Conference, SACAIR 2021*, Durban, South Africa, 125-139.