ABSTRAK

Nama	:	Sri Hartati Wijono
Program Studi	:	Doktor Ilmu Komputer
Judul	:	*Canonical Segmentation* Untuk Meningkatkan Hasil Terjemahan Mesin bahasa Jawa – bahasa Indonesia
Pembimbing	:	Prof. Dr. Eng. Wisnu Jatmiko, S.T., M.Kom.

Neural Machine Translation (NMT) mencapai hasil yang memuaskan untuk *high-resource parallel corpus,* sehingga NMT banyak digunakan. NMT memiliki kelemahan untuk *low-resource parallel corpus* karena memunculkan banyak kata *out of vocabulary*. Penelitian pengembangan NMT bahasa Jawa – bahasa Indonesia merupakan penelitian dengan data *low-resource* dan berupa bahasa aglutinatif. Penelitian mengusulkan *subword* hasil *canonical segmentation* digabung dengan *tag* fitur linguistik sebagai masukan ke *encoder-decoder Transformer* pada proses *training*. Subword hasil segmentasi untuk meningkatkan frekuensi kosa kata. Tag fitur berupa afiks dan *root word* untuk membantu proses pembelajaran *encoder* dan membentuk keluaran *decoder* saat proses *training*. Penelitian ini terbagi menjadi dua tahap. Tahap pertama, penelitian melakukan proses *canonical segmentation* bahasa Jawa menggunakan *encoder-decoder* berbasis *Transformer*. Eksperimen mengusulkan tiga tipe masukan ke *encoder-decoder* pada proses *training.* Tipe pertama adalah unit afiks dan sederetan karakter *root word.* Kedua dan ketiga adalah sederetan karakter afiks dan *root word* yang di gabung dengan *tag* fitur berupa afiks dan *root word* atau urutan *root word.* Tahap kedua, penelitian mengembangkan NMT berbasis *Transformer* dengan usulan masukan berupa *subword* bahasa Jawa hasil *canonical segmentation* tahap satu. Subword di gabung dengan *tag* fitur berupa afiks dan *root word*. Subword bahasa Indonesia menggunakan hasil dari *MorphInd*. Karena *corpus* paralel *canonical segmentation* bahasa Jawa belum tersedia, maka penelitian ini mengembangkan algoritma untuk membuat corpus tersebut. Hasil penelitian menunjukkan bahwa *canonical segmentation* terbaik adalah menggunakan masukan karakter di gabung dengan *tag* fitur berupa afiks dan *root word*. Tipe masukan terbaik menghasilkan akurasi segmentasi sebesar 84,29% untuk semua kata dan 56,09% untuk kata berimbuhan *canonical*. Nilai F1 yang dihasilkan 92,71% untuk semua kata dan 96,62% untuk kata berafiks *canonical.* Metode NMT terbaik adalah yang menerapkan masukan *subword* hasil segmentasi di gabung dengan *tag* fitur berupa afiks dan *root word.* Metode terbaik meningkatkan nilai BLEU sebesar +3,55 poin dibandingkan penggunaan kata dan meningkat +2,57 poin dibandingkan penggunaan *subword* BPE. Disimpulkan, *canonical segmentation* berbasis *neural network* dengan masukan sederetan karakter digabung *tag* fitur afiks dan *root word* pada proses *training* lebih baik dibanding masukan sederetan karakter. Subword hasil *canonical segmentation* digabung dengan *tag* fitur berupa afiks dan *root word* sebagai masukan ke *encoder* dan *decoder* saat *training* dapat meningkatkan nilai BLEU pada *low resource* NMT.

Kata kunci : *low-resource Neural Machine Translation, canonical segmentation, terjemahan mesin bahasa Jawa- bahasa Indonesia, tag fitur*

ABSTRACT

Nama	:	Sri Hartati Wijono
Program Studi	:	Doctor of Computer Science
Judul	:	*Canonical Segmentation* to Improve Machine Translation Javanese – Indonesian
Pembimbing	:	Prof. Dr. Eng. Wisnu Jatmiko, S.T., M.Kom.

Neural Machine Translation (NMT) achieves satisfactory results for high-resource parallel corpus, so NMT widely used. NMT has a weakness for low-resource parallel corpus because it raises many out-of-vocabulary words. Research developes Javanese – Indonesian NMT and both are low-resource and agglutinative languages. The research proposes that the canonical segmentation subwords are concatenated with linguistic feature tags as input to the Transformer encoder-decoder in the training process. Subword segmentation is to increase the frequency of vocabulary. Feature tags are in the form of affixes and root words to help the encoder learn the process, and decoder generates output during the training process. This research is divided into two stages. In the first stage, the research conduct the process of Javanese canonical segmentation using a Transformer-based encoder-decoder. The research proposes three types of input to the encoder-decoder in the training process. The first type is an affix unit and a sequence of root word characters. The second and third are a sequence of affixes and root words concatenated with feature tags in the form of affixes and root words or sequences of root words. The second stage, the research used a Transformer-based NMT with the proposed input in the form of Javanese subwords resulting from the first stage of canonical segmentation. Subwords are combined with feature tags in the form of affixes and root words. Indonesian subwords use results from MorphInd. Because the corpus parallel canonical segmentation for the Javanese is not yet available, this research develops an algorithm to create the corpus. The study reports that the best canonical segmentation is the one that uses character input concatenated with feature tags that include affixes and root words. It achieves a segmentation accuracy value of 84,29% of all occurrences and 56.09% of canonical affixed words. This study reaches F1 score of 92,71% of all occurrences and 96,62% of canonical affixed words. The best NMT experiment is the one that applies subword input as a result of segmentation concatenated with feature tags in the form of affixes and root words. This study increased the BLEU score by +3.55 points compared to using the word and +2.57 points compared to using BPE subwords. In conclusion, neural network-based canonical segmentation with a sequence of input characters concatenated with affix and root word feature tags in the training process is better than the input of a sequence of characters. Subwords resulting from canonical segmentation concatenated with feature tags in the form of affixes and root words as input to the encoder and decoder during training can increase BLEU score at low resource NMT.

Keyword : *low-resource Neural Machine Translation, canonical segmentation, Javanese-Indonesian Neural Machine Translation, feature tag*