Canonical Segmentation Untuk Meningkatkan Hasil Terjemahan Mesin bahasa Jawa – bahasa Indonesia = Canonical Segmentation to Improve Machine Translation Javanese Indonesian

Sri Hartati Wijono, author

Abstrak

Terjemahan mesin adalah program komputer yang menerjemahkan kata dari satu bahasa ke bahasa lain. Neural Machine Translation (NMT) merupakan salah satu jenis terjemahan mesin yang menggunakan hasil pelatihan corpus paralel untuk menerjemahkan kata. Proses NMT dengan pelatihan menggunakan corpus paralel dalam jumlah besar (high resource) dapat memberikan hasil terjemahan sangat baik. Tetapi proses NMT yang dilatih menggunakan corpus paralel dalam jumlah kecil (low-resource) tidak mampu memberikan penerjemahan kata dengan baik akibat adanya out-of-vocabulary (OOV). Salah satu cara mengurangi OOV pada low-resourse NMT adalah melatih NMT menggunakan subword dari hasil segmentasi kata. Canonical segmentation dipilih untuk mengsegmentasi kata bahasa Jawa dan bahasa Indonesia menjadi subword afiks dan subword root word yang mengalami alomorf. Hal ini dikarenakan kedua hasil subword tersebut memiliki makna linguistik yang dapat digunakan untuk mengurangi OOV. Proses canonical segmentation tersebut dilakukan menggunakan encoder-decoder Transformer dengan memanipulasi masukannya sebagai usulan dari penelitian. Penelitian ini juga mengembangkan algoritma untuk membuat dataset canonical segmentation bahasa Jawa yang digunakan untuk melatih Transformer. Manipulasi masukan Transformer tersebut berupa penggunaan tag fitur afiks dan root word atau tag fitur afiks dan urutan root word yang digabungkan ke setiap karakter masukan untuk membantu proses pembelajaran Transformer. Manipulasi usulan ini menghasilkan akurasi segmentasi sebesar 84,29% untuk semua kata, 69,82% untuk kata berimbuhan dan 56,09% untuk kata berimbuhan canonical. Nilai F1 yang dihasilkan 92,89% untuk semua kata, 98,69% untuk kata berimbuhan dan 96,81% untuk kata berimbuhan canonical. Subword hasil proses segmentasi ini selanjutnya digabung dengan tag fitur berupa afiks dan root word untuk menguji low-resource NMT. Metode ini dapat eningkatkan nilai BLEU sebesar +3,55 poin dibandingkan penggunaan kata tanpa segmentasi dan meningkat +2,57 poin dibandingkan penggunaan subword BPE yang banyak dipakai saat ini.

.....Machine translation is a machine that translates words from one language to another. Neural Machine Translation (NMT) is a type of machine translation that uses the results of parallel corpus training to translate words. The NMT process with training using a large number of the parallel corpus (high resource) can give excellent translation results. But the NMT process, which was trained using a parallel corpus in small numbers (low resources), could not provide good word translation due to out-of-vocabulary (OOV). One way to reduce OOV in low-resource NMT is to train NMT using subwords from word segmentation results. Canonical segmentation was chosen to segment Javanese and Indonesian words into affix and root word subwords that experience allomorphism. This segmentation method was chosen because the two subword results have linguistic meanings that can be used to reduce OOV. The canonical segmentation process is conducted using Transformer encoder-decoder by manipulating the input as a research proposal. This research also develops an algorithm to create a corpus parallel canonical segmentation in the Java language used to train Transformers. Manipulating the Transformer input uses affix and root word feature

tags or affix and root word sequences concatenated with each input character to help the Transformer learning process. This proposed manipulation produces a segmentation accuracy of 84.29% for all words, 69.82% for affixed words and 56.09% for canonical affixed words. The resulting F1 value is 92.89% for all words, 98.69% for affixed words and 96.81% for canonical affixed words. The subwords resulting from the segmentation process are then combined with feature tags in the form of affixes and root words to test low-resource NMT. This method can increase the BLEU value by +3.55 points compared to using words without segmentation and +2.57 points compared to using BPE subwords which are widely used today.