ABSTRAK

Abstrak— Pembelajaran bahasa dengan mesin sekarang ini sudah mencapai pada tahap dimana jumlah data memegang kunci penting pada hasilnya. Berbagai pengembangan dilakukan sehingga masalah terkait kualitas data dapat diatasi, terutama banyak terjadi di era mesin penerjemah saraf tiruan (MPST). Salah satu ide besar untuk mengembangkan pembelajaran lebih lanjut untuk hal ini adalah dengan membangun sistem secara *low-resource*, dimana akan memfokuskan media pembelajar untuk bekerja dan mampu membentuk pengetahuan dengan sumber data dengan kondisi tersebut. Arsitektur MPST sudah umum digunakan sebagai bagian arsitektur sistem *machine translation* sejak populer di WMT16, dan mempunyai kelebihan dapat melakukan berbagai tugas terkait penerjemahan bahasa secara tepat dan mudah. Tersebutlah *toolkit* MarianNMT, Sebuah *toolkit* mesin penerjemah saraf tiruan yang dikembangkan oleh tim *Microsoft Translator* dengan harapan menciptakan *toolkit* yang *resource-friendly* dan dapat mencapai kecepatan *training* dan penerrjemahan yang tinggi serta *support* untuk impementasi pada sistem lokal yang mengandalkan sumber daya *hardware* dalam *device* dengan GPU atau CPU. Dalam kasus penerjemahan Bahasa Inggris ke Indonesia dan menggunakan model “Nematus-Style Shallow RNN” pada MarianNMT, dalam 28 jam mampu untuk menyelesaikan training untuk kedua kasus training dengan korpus yang memiliki baris < 500K kalimat. Pada training digunakan validasi dengan repository FLORES-101, dan membawa dua kasus training dengan korpus berbeda dari Wikimedia untuk memperoleh nilai BLEU (5.2 - 4.7), SpBLEU (8.1 - 7.2) dan QED&TED dengan nilai BLEU (4.0 - 4.3) SpBLEU (6.8 - 6.9) untuk terjemahan dari korpus *dev* dan *devtest* berturut-turut. Menyimpulkan bahwa korpus Wikimedia memiliki kecocokan dengan evaluasi pelatihannya, namun belum cocok untuk melakukan penerjemahan pada bentuk kata yang tak pernah terlihat. Sedangkan pada korpus QED&TED hal tersebut tercapai walau dengan perbandingan skor yang lebih kecil.

Kata Kunci: mesin penerjemah saraf tiruan, sentencepiece, pemrosesan bahasa alami, marianNMT, BLEU, FLORES-101.

ABSTRACT

Abstract— Machine learning now has reached a stage where the amount of data holds the key to the outcome. Various developments were carried out so that problems related to data quality can be situationed, especially on what happened in the era of neural machine translation (NMT). One of the big ideas to develop further learning for this matter is to build a low-resource system, which will focus on learning media to work and be able to form knowledge with data sources under these conditions. The NMT has been commonly used as part of the machine translation system architecture since it was popular in WMT16, and has the advantage of being able to perform various language translation tasks precisely and easily. The toolkit is MarianNMT, a neural machine toolkit developed by the Microsoft Translator team with the hope of creating a resource-friendly toolkit that can achieve training and translation speeds as well as support for implementation on hardware-resourced on-premises systems with GPUs or CPUs. In the case of the English-to-Indonesian translation way and using the "Nematus-Style Shallow RNN" model on MarianNMT, in 28 hours he was able to complete the training for both training cases on corpora having lines < 500K sentences. In the training, validation with the use of the FLORES-101 repository, brought two cases of training on different corpus source from Wikimedia to obtain BLEU scores (5.2 - 4.7), SpBLEU (8.1 - 7.2) and QED&TED with BLEU values (4.0 - 4.3) SpBLEU (6.8 - 6.9) for the translation of the dev and devtest corpus respectively. Conclusion made, that the Wikimedia corpus was suitable for its training evaluation, but not yet suitable for translating unseen word forms. Meanwhile, in the QED&TED corpus, this was achieved even with a smaller score comparison.

Keywords: neural machine translation, sentencepiece, natural language processing, marianNMT, BLEU, FLORES-101.