ABSTRACT

LANGUAGE RELATEDNESS IN MULTI-SOURCE NEURAL MACHINE TRANSLATION WITH MISSING DATA

Devni Karmelita Prasetyo

15/377119/PA/16445

Machine translation is tasked to translate from source language to target language. There are three major approaches for this task, Rule-based Machine Translation, Statistical Machine Translation, and Neural Machine Translation. The recent machine translation is using the neural networks approaches, but this approaches need a large amount of data to deliver a decent translation result, and not all languages have a large amount of data complete multilingual corpus.

This research try maximize the language combination in input of multi-source NMT in low resource settings. The multi-source NMT system task is to translate one source language to target language with the help of another language. Generally, machine translation need a complete multilingual corpus for input, but this NMT can process incomplete multilingual corpus as the input. The data used is an actual incomplete multilingual corpus from TED Talks transcription.

The results show that multi-source NMT is generally performing better than single-source NMT with average BLEU gains is 1.2. On the source side the number of missing sentences in the helper language give more impact compared to the relatedness between source language and helper language in translation quality. The best BLEU gains in two-to-one NMT setting is 3.2 from {En, Nl}-to-De, it showed German perform best with Dutch, which is very closely related to German, as a helper language. In three-to-one NMT setting, the best BLEU gaines is from {En, Uk, Be}-to-Ru and {En, Es, Uk}-to-Ru task with 5.6 BLEU gaines, both models have Russian as target language and Ukrainian as helper language, which is very closely related to Russian.

Keywords: machine translation, neural networks, multi-source NMT, and language relatedness

INTISARI

KETERKAITAN BAHASA DALAM MULTI-SUMBER MESIN TRANSLASI SYARAF TIRUAN DENGAN DATA YANG HILANG

Devni Karmelita Prasetyo

15/377119/PA/16445

Mesin translasi memiliki peran untuk menerjemahkan dari sumber bahasa ke target bahasa. Di dalam mesin translasi ada 3 pendekatan yang sering digunakan, Mesin Translasi Statistik, Mesin Translasi Berbasis Aturan, dan Mesin Translasi Syaraf Tiruan. Dalam beberapa waktu terakhir, mesin translasi syaraf tiruanlah yang paling sering digunakan, namun pendekatan ini membutuhkan jumlah data yang sangat besar, dan tidak semua bahasa memiliki *corpus* multi-sumber dalam jumlah yang besar.

Penelitian ini mencoba untuk memaksimalkan kombinasi bahasa dalam masukan dari multi-sumber MTS. Multi-sumber MTS memiliki peran untuk menerjemahkan dari satu sumber bahasa ke target bahasa dengan bantuan dari bahasa lain. Secara umum, mesin translasi membutuhkan jumlah *corpus* multibahasa lengkap yang sangat besar sebagai masukan., tapi MTS ini bisa memproses *corpus* multibahasa yang tidak lengkap. Data yang digunakan dalam penelitian ini adalah *corpus* multibahasa yang tidak lengkap yang bersumber dari TED Talks.

Hasil dari penelitian ini menunjukan bahwa multi-sumber MTS meraih skor BLEU yang lebih bagus dibandingkan satu-sumber MTS secara keseluruhan. Pada bagian sumber bahasa, jumlah kalimat yang hilang akan memberikan pengaruh lebih daripada relasi antara sumber bahasa dan bahasa pembantu dalam akurasi translasi. Penambahan BLEU terbaik dari pengaturan two-to-one NMT adalah 3.2 dari {En, Nl}-to-De, disini dapat dilihat bahwa Bahasa Jerman memperoleh hasil terbaik dengan bantuan dari Bahasa Belanda yang memiliki relasi yang sangat dekat. Sedangkan pada setelan three-to-one NMT, penambahan BLEU terbaik dicapai oleh {En, Uk, Be}-to-Ru dan {En, Es, Uk}-to-Ru dengan penambahan score 5.6, kedua model tersebut memiliki target Bahasa Rusia dan di bantu dengan Bahasa Ukraina yang dekat relasinya dengan target bahasa.

Keywords: mesin translasi, jaringan syaraf tiruan, multi-sumber MTS, and keterkaitan bahasa