Reading Assignment 1

Jingyuan Wang, 260860682

In the paper *Achieving Human Parity on Automatic Chinese to English News Translation*, the authors proposed a new language translating model, defined a way to assess human parity translators, and evaluated their model with both automatic evaluation and source-based human assessment.

The model made rather significant improvements compared to other machine translators by the following three innovations. First of all, the authors used the idea of dual learning, namely the system would first translate the source sentence to the target language and further translate the generated target sentence back to the source language. On basis of this, they implemented a new approach called joint training, which is to iteratively apply the source-to-target and target-to-source learning. Both translation models will be augmented from the process and either supervised or unsupervised training can be used. Another technique they invented is a new agreement regularization term which enhanced an algorithm called deliberation networks where decoding performance can be learned from both left-to-right and right-to-left. Otherwise, with single direction sequence learning, the result of a decoding largely depends on the accuracy of previous results. Finally, since the model performance is rather sensitive to the quality of data, they utilized a data filtering procedure. A cross lingual sentence to vector system is trained and each pair of Chinese and English sentence in the corpus is transformed into vectors. The sentence will be removed if vector representation of both language is not similar to each other. In addition, system combination was used to obtain better performance.

The system is evaluated in two aspects: one is how well did the model accomplish translation tasks, the other is if the performance of the system reached human parity. For the first task, the authors uses BLEU, an automatic translation evaluation approach. It compares the generated translations with several reference sets, and a model would achieve higher score if more n-gram of the candidate overlaps with the reference. Following this rule, different systems, different data filtering algorithm and several combinations were experimented and it is shown that the models with DLDN and ARJT enhanced the system with a large extent. However, when evaluating model's human parity, the paper utilized source-based direct assessment, where bilingual annotators are given candidate translation and its corresponding source data and are asked to give marks towards how well does the translation express the semantic of the original sentence. Afterwards, with normalized scores, different systems are clustered into groups with certain score ratio, so the models and other benchmark translations would be evaluated in a fair condition.

The paper claimed that the translator model has achieved human parity by proving that there is no statistically obvious difference between machine and human translation. As the result indicates, combo system 4,5,6 has equal quality with one of the reference human translation corpus, and thus we can say that human parity has been reached on this Chinese to English translation task on this specific source text set. However, there are several crucial issues. Model performances may vary from task to task, and it is likely that the model reached such effect only with particular source corpora. The news texts they used made the translation task simpler, however, they may not achieve human parity with other corpus. Furthermore, manually scoring is a subjective process and an unstable factor. It is critical if the annotators are professional and redundancy where each sentence is graded by several annotator is enough to alleviate the bias.

Notwithstanding that human parity translation is realized, the problem of machine translation is far away from solved. The main purpose of this paper is to prove that there is no significant difference between human and machine translation. But when looking into the translations, numerous errors were found such as incorrect words and syntax error. Besides, as mentioned above, human parity is only reached on this particular task. No performance is necessarily generalized to other dataset, nor other languages. Finally, since joint training is used in this model, such enhancement is not suitable for real-time translation, while it is a noteworthy problem in machine translation. People still need to work on this topic and ideally implement an error-free, universal, yet semantically correct and beautiful machine translation system.