Assignment 5 – Paper dissection

Vu, Tu et al. (2018). "Sentence Simplification with Memory-Augmented Neural Networks". In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 79–85. DOI: 10.18653/v1/N18-2013.

The paper I chose is about sentence simplification. Simplifying a text means to reduce its structural and lexical complexity. As the authors state in their introduction, there has been success for this task when using sequence to sequence neural models. These models that can be used to simplify a text, encoder-decoder architectures, were originally designed for machine translation. This is due to the similarity of the task. Why this is similar is not explained in the paper. However, it does make sense as standard (complex) language and simplified language can in fact be considered two different languages, even if the 'base' language is the same. The syntax and the lexicon of two variants of the same language might differ a lot. Their starting point is that they pinpoint a problem with currently used neural models for text simplification: Common practice is to use either LSTM (long short-term memory) or GRU architectures (gated recurrent unit). Both of these architectures are able to memorize dependencies across sequences. However, this memory might not be large enough for simplification tasks as the original sentences that need to be simplified can be very long and complicated. The authors therefore propose to use a memory augmented RNN architecture which is called Neural Semantic Encoder (NSE). Their goal is to go beyond normally used models to perform simplification. An NSE produces a memory matrix during each encoding time step. This means that for each word in an input sequence, this matrix is updated. In the beginning for the first word it contains just the initial word vectors. Those word vectors are then updated with the information gained while processing the sequence. This architecture allows to have unrestricted access to all words in a sequence at any time. The model can therefore look at the other words while encoding one.

Their experiment consists of two parts: First, they want to try different architectures for the encoder of their model and evaluate which is the best. Then, they test the model on three different datasets to evaluate its overall performance. Each dataset consists of complex-simple sentence pairs in English. There are two basic architectures that they use. One is a LSTM-only model that uses two LSTM layers as encoder and another two as decoder. The other model has

an NSE encoder and again two LSTM layers as decoder. They used different hyperparameter settings which I will not describe here. They used five other existing models for simplification to compare their models to. They evaluated their model using BLEU (a metric from MT that is based on n-gram counts) and SARI (a metric specifically made for text simplification that compares the reference, the prediction and the input sequence) and human evaluation. They conclude that their NSE memory-augmented model works quite well in simplifying long and complex sentences.

Personally, I find it very interesting that extending architectures like RNN can help getting better results for those tasks. However, I find it a little suspicious if the conclusion just states that the "results of both automatic and human evaluation on different datasets show that our model is capable of significantly reducing the reading difficulty of the input, while performing well in terms of grammaticality and meaning preservation." I'd wish to have a more in-depth analysis of why it works well as language is very complex and depends on many different factors. Especially, as their second evaluation method is human evaluation of a subset of randomly chosen simplified sentences for each model. Generally, I think this is perfectly fine, but they mention once that there are only three people evaluating the sentences of which only one was a non-native English speaker (i.e. an actual member of the target group of simplification apart from people with disabilities). Maybe this is common practice and works well but I find this a little strange to judge the performance of a model. I understand that human evaluation is expensive and very difficult to get. At least they could have asked people with reading disabilities or language learners to read the texts to get a better feedback if the target group understands the texts. And if I'm correctly informed there are also metrics that evaluate the readability of a text. I would have found that interesting to see how the model perform on producing well readable text.

Further resources

I was not quite sure about the difference between encoder-decoder and sequence to sequence. The following article helped me:
https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346