Improving English to Arabic Machine Translation

Wael Abid , Younes Bensouda Mourri
CS 224N | Stanford University

Problem

English to Arabic translation is not very well explored in the literature.The potential bottleneck behind such research is the scarcity of researchers who understand the linguistic structure of the Arabic language very well. Arabic is a morphologically rich language and usually combines pronouns, conjugation, and gender in one word. For example, the word ولمدرستها (walimadrasatiha) is one word. However, each letter represents a word. The prefix و (wa) corresponds to and, the letter ل (li) corresponds to the word for, مدرست (madrasa) means school, and the suffix ها (ha) corresponds to the gender pronoun 'her'. Hence, even when computing the BLEU score, one very small suffix could easily lower the overall results although the other three subwords are right.

Data / Task

We compiled our data from different sources and domains so that the model doesn't learn a specific language or writing style, and that it learns both formal and less formal Arabic (Modern Standard Arabic, not colloquial). We used the Arab-Acquis data (Habash 2017) which presents data from the European Parliament proceedings totalling over 600,000 words. We combine that with the AMARA Corpus (Guzman 2013) which consists of TED talks parallel data totalling nearly 2.6M words. In addition to this we add 1M words of movie subtitles data from Open Subtitles 2016 (Lison 2016).

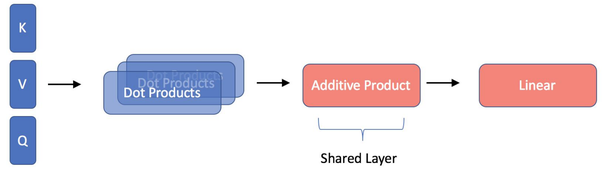
Approach

We worked on both the modeling and architectural side of the problem. For the modeling part, we viewed that it's best to train and test on morphology-aware tokenization that considers the following: (a) morphological and syntactic feature matching, (b) stem matching between the source language, English and the target language, Arabic.. Based on Arabic linguistic intuition, we check the matching of a subset of 5 morphological features: (i) POS tag, (ii) gender (iii) number (iv) person (v) definiteness. We also explored pre-trained embeddings as a tool to improve performance.

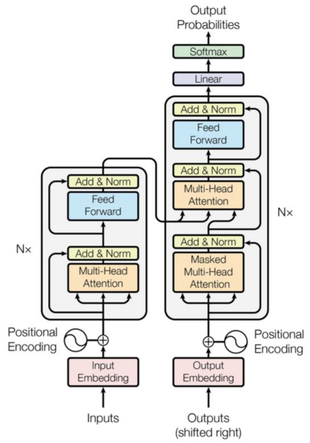
We used a transformer model as described in "Attention is All You Need" (Vaswani 2017) from OpenNMT as our vanilla Transformer model. After a challenging amount of pre-processing and preparation of the data pipeline, we ran the model to get a baseline score. Each unit has a multi-headed attention followed by some normalization layer and a skip connection. The output is then followed by a Feed-forward and another normalization layer. This is considered one unit and there are N of these in the encoder. For the architectural side of it, we modified the multi-headed self attention self attention because we believed that at the point before the projection we can have a shared layer for two main reasons which stem from the logic of the following analogy:

- If you can see a part of a human body, you can assume that there is a body in that location, so you don't need to learn it all over again.
- If you have already seen a body before and you saw a slightly different body, you can generalize. Not each body needs its own model.

The same logic applies to Arabic words, prefixes, suffixes and other subwords. A suffix corresponding to the gender pronoun 'her' is the same in every word so there's no need to reinvent the wheel every time. Our modification in the multi-head attention is as follows:



Where K, V, and Q are the keys, values, and queries. We then implement the following model.



We then use morphology aware tokenization of the Arabic words, and pretrained embeddings.

Results

Model	BLEU
Baseline character- based	28.46
Baseline Transformer	30.73
All Tokenize Model	33.79
All Tokenized + Pre-trained Embbeddings Moels	36.53
Multi-Head Shared Attention	35.18

Analysis

The reason why we think morpholgy-aware tokenization works is the following: as previously stated, Arabic is a morphologically rich language, and this makes Arabic Neural Machine Translation difficult to approach as it increases the number of out-of-vocabulary tokens which means that it consequently worsens the issue of data sparsity. One other problem is that it makes it difficult to have parallel words in the two different languages (in our case it's Arabic and English) which makes it more difficult to get good translations. Tokenization solves this problem and makes for a more parallel data between English and Arabic.
We add to this pre-trained embeddings which improve our results very well, and we believe this is due to the fact that these pre-trained embeddings have been part of a training process where they were trained on very large datasets which makes them encapsulate a lot of the meaning and context of those words that are otherwise not frequent in our low-resource pair.

The Transformer outperforms the character-based model because it applies a self-attention mechanism which directly models relationships between all words in a sentence, regardless of their respective position and this makes it possible to preserve long-range dependencies in a sentence and make good predictions of the next word. We modified the architecture of the transformer model because we believe that the repeated patterns in the language would translate into repeated patterns in the multi-headed self attention that could be learned simultaneously.
We can see the benefits of such distributed attention, but we believe that it is overkill to implement it all the way, especially that it only slightly reduces the BLEU score. This means that we can effectively reduce the number of parameters by [heads*(model dimension - 1] parameters and still keep a similar performance. As a result we do speed up the training and reduce model complexity. Hence, we consider that our hypothesis of a shared layer after the key/value/query product is effective.

Conclusion and Future Work

We present English to Arabic Neural Machine Translation on a rich and diverse dataset including a first extended result on English to Arabic NMT using Transformers. We show the importance of morphology-aware preprocessing and pretrained embeddings and how they contribute to a better translation. We change the transformer model architecture to keep track of shared patterns and show that our architecture still works. For the next steps, we would like to design a new metric for the arabic language instead of the BLEU score.

References

Ashish Vaswani, et. al. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors. Curran Associates, Inc., 2017.
Ahmed El Kholy and Nizar Habash. Orthographic and morphological processing for english–arabic statistical machine translation.Machine Translation 26(1):25–45, Mar 2012.
Guillaume Klein, Yoon Kim, Yuntian Deng, Josep Maria Crego, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. In ACL, 2017.
Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575. Association for Computational Linguistics, 2016