Danish-English Neural Machine Translation with Contextual Embedding

Mammon Habib   Fouzi Takelait
Department of Computer Science
University of Massachusetts Lowell
{fouzi_takelait, mamoon_habib}@student.uml.edu

Abstract

Machine translation (MT) models and systems are recently being steadily improved with the evolution of deep learning. However, research on MT in low-resource languages such as Danish is still very limited. In our work, we utilize the Europal benchmark for Danish-RoBERTa model trained from scratch. Our findings show that fine-tuning Danish-RoBERTa model did not learn well due to limited downstream usage of Danish language. Furthermore, experimental results assessed through BLEU metric show that contextual embedding significantly improves the quality of Danish–English NMT using our own encoder-decoder model. The code for this project is available on GitHub[1].

1 Introduction

Neural MT has recently been dominating machine translation paradigms where this kind of MT attempts to build and train a large neural network that read a sentence and outputs a correct translation (Bahdanau et al., 2014). These MT systems are based on the encoder-decoder model in which the encoder reads and encodes the source sentence into a fixed-length vector while the decoder produces a translation output from the encoded vector (Bahdanau et al., 2014; Cho et al., 2014).

The leading models are large both with respect to the number of parameters and the size of the training data used to build the model; this correlation between size and performance has been demonstrated by (Kaplan et al., 2020). Turning to the Danish language, there is no such truly large-scale models available. Whereas, there are around 300 Germanic models available in the Hugging Face Transformers model repository[2]. Most of these are translation models, but there is already a significant number of monolingual models available in the Germanic languages; at the time of writing there are around 21 danish-english available in Hugging Face Transformers model repository[3] .

However, none of these Germanic languages, including Danish language, models are even close to the currently leading English models in parameter size or training data used. As such, we can expect that their relative performance in comparison with the leading English models is significantly worse.

In this project we aimed to use two transformer models for translating text from Danish to English language. The first transformer consists of our encoder and decoder model initialized from scratch while the second model includes a pre-trained component that we use as an encoder and combine it with our own decoder and compare the performances of the two models. For the pre-trained component of our encoder, we chose a Danish-RoBERTa model.

2 Data

In order to be able to use comparable data in our NMT models, we used Danish-English dataset; A Parallel Corpus for Statistical Machine Translation, consisting of data downloaded from statmt.org/europarl. The Europarl parallel corpus is extracted from the proceedings of the European Parliament. It includes versions in 21 European language. The goal of the extraction and processing was to generate sentence aligned text for statistical machine translation systems.

To further improve the quality of the data we apply preprocessing step to filter out any proceedings containing noise. This results in a balanced set of 1.8 M sentences, with 2,000 samples for testing.

[1] github.com/ftakelait/da-en-machine-translation

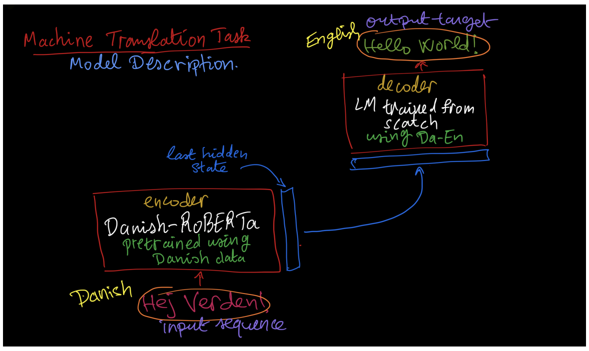
[2] huggingface.co/models

[3] huggingface.co/models?language=en,da

3 Models

In order to fairly select a representative pre-trained model for our considered Danish language, we opt for the most popular native model according to Hugging Face. This corresponds to a BERT-Base model, hence it is represented by a Language Model of identical architecture.

We used Danish Roberta Base; a danish pre-trained Roberta base model that was pretrained on the danish mC4 dataset during the flax community week. This project was organized by Dansk Data Science Community (DDSC). A random sample from the entire mC4 dataset with below 63% chance will be danish language ($< 100M$ sentences) and over 63 % chance will be english language ($> 100M$ sentences), of which it has near poor quality for danish language(Kreutzer et al., 2021).



4 Related Work

Previous work related to Danish-English Machine Translation included using Marian; a neural machine translation framework written in C++ and trained on the OPUS dataset using OPUS-mt-train. This model achieved a BLEU score of 65.1. In Another work related to Danish-English translation, Microsoft trained two transformer models using an encoder-decoder initialized from scratch and using a pre-trained XLM-Roberta model as an encoder and combined it with a decoder that was initialized from scratch. The BLEU for using a transformer and a pre-trained component were 58.73 and 67.02 respectively.

5 Experiments

We train and evaluate each of the two models on the Danish-English Europarl data, using the hyperparameters listed in Table 1. Following common practice, we preprocessed the dataset using BPE tokenizers to alleviate the Out-of-Vocabulary problem for both the languages when training the model from scratch (Sennrich et al., 2015). Similarly, in the case of the pre-trained component we used a pre-trained tokenizer on the Danish input language. Both the models were trained on a single NVIDIA V100 GPU. Training the language model includes first using the encoder that uses a Self-Attention mechanism to generate embeddings for the input. This representation of the input is combined with the representation of the output, through a cross-attention mechanism that directly links the words in the target language with the corresponding words of attention in the source language to make predictions for the output's translation. For the pre-trained Danish-RoBERTa model we fine tune the parameters of the encoder and follow the same methodology as we did for the model initialized from scratch.

6 Results

We used BLEU score metric to evaluate the performance of the two systems. BLEU is the most common method for assessing the accuracy of MT systems. It indicates how similar the candidate text is to the reference text. The range of BLEU is from 0 to 100. A higher BLEU score means that the ML system has higher accuracy. For the baseline model a BLEU score of **34.6** was achieved. While for the RoBerta Model we were able to achieve a BLEU score of only **3.8**.

7 Discussion

After training the model from scratch we were able to achieve a decent BLEU score as shown in Table. However, using the pre-trained Danish-RoBERTa model we were not able to achieve a decent BLEU score and hence the output would produce random translations of the original input. We suspect one of the reasons why the Danish-RoBERTa model could not perform well was because the dataset that the model was trained on included lesser number of training examples from the Danish language that were also not of a higher quality compared to the other languages available in mc4.

Additionally, we also created a web-app where we deployed both our models. This interface would take as input the Danish language and produce an English translated output for both the models. An example of the translation is shown in Section 8.

Parameters	**Value**
train_epochs	1
optimizer	ADAMW
learning_rate	3e-4
batch_size	32
max_seq_length	128
hidden_size	768
num_heads	8

Table 1: Training hyperparameters for the Danish-English machine translation experiments.

Parallel Corpus (L1-L2)	Sentences	L1 Words	English Words
Danish-English	1,968,800	44,654,417	48,574,988

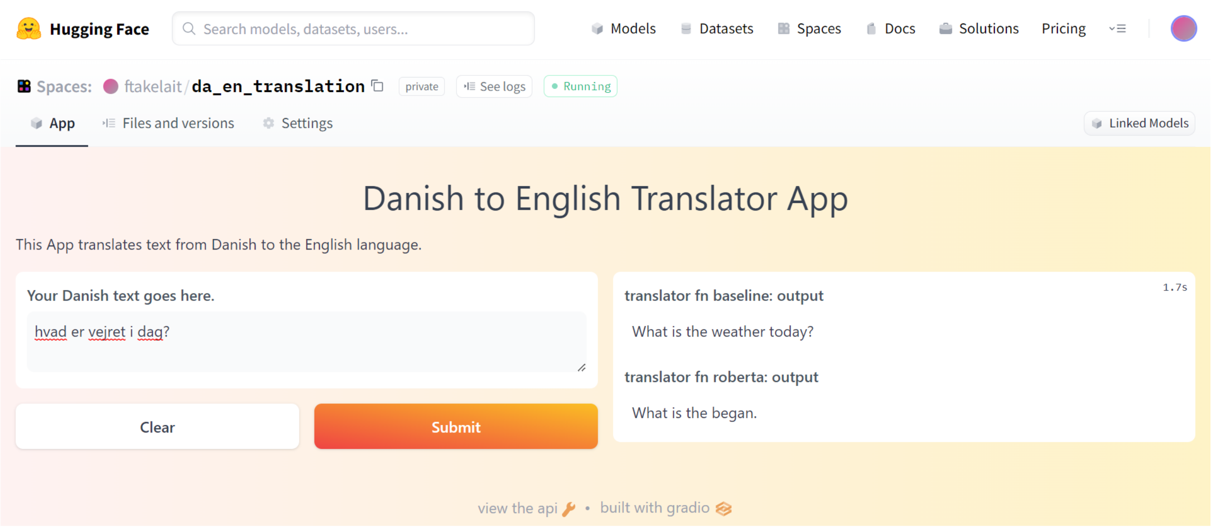
Table 2: Total number of sentences, L1 Words, and English Words for the proceedings of the European Parliament data of Danish-English parallel language translation.

Model name in Hugging Face	**Language**	**Data Size**
flax-community/roberta-base-danish	da-en	<100M

Table 3: Model used in the experiments and the size of its corresponding training data. 'M' is short for million

8 Danish to English Translator App Demo

You can try our danish-english translator app here [4].



Danish to English Translator App

This App translates text from Danish to the English language.

Your Danish text goes here.

hvad er vejret i dag?

1.7s

translator fn baseline: output

What is the weather today?

translator fn roberta: output

What is the began.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

[4] huggingface.co/spaces/ftakelait/da_en_translation.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.