The University of Texas at Dallas
CS 6320
Natural Language Processing
Spring 2019

Class Project

Machine Translation: Danish-English Translator

Aadish Joshi, Sushant Singh Dahiya.

1. The Problem

The idea was to implement a Machine translator encoder decoder architecture for translating one human language to another. For this project we would like to translate Danish to English.

2. Methodology

An encoder will be trained on Danish language to generate a "Thought Vector" which will be an aggregate of the trained encoder data. This thought vector along with the English encoded Vector will be used to train the decoder unit. Both encoder and decoder modules will consist of Recurrent neural networks for training purposes.

3. Data Annotations

We used European Parliament Proceedings Parallel Corpus 1996-2011 containing 21 European languages. This corpus was created by European Union with the goal to generate sentence aligned text for statistical machine translation systems. http://www.statmt.org/europarl/

4. Implementation

Encoder will be trained on source language (Danish). Danish input sentences will be preprocessed to add padding for the beginning and the end of the sentences. Encoder will consist of Tokenizer unit to tokenize Danish sentences. These tokens will then be mapped to vector embedding i.e. a float point valued vector which are then fed in to the Gated Recurrent Units. Gated Recurrent units are programmed to generate a single valued output.

Decoder will be trained on destination language (English) in combination of a thought vector generated from encoder. Similar to the encoder, decoder will consist of English tokenizer and then English vector embeddings. Gated Recurrent Units at the decoder will generate English text as an output.

The predicted flowchart will be like this.

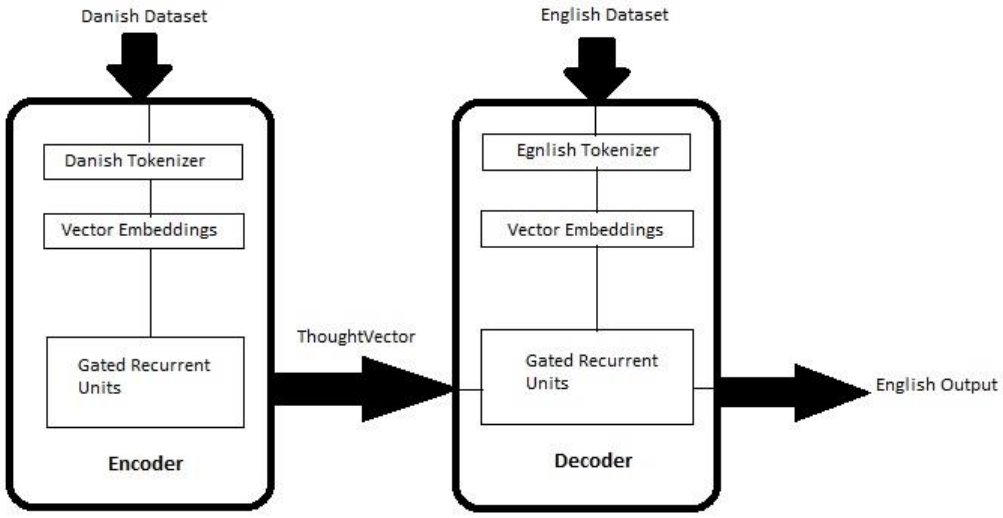

Fig. Encoder-Decoder Architecture
Machine Translation using Recurrent Neural Networks

5. Experiments Projected

We think LSTM i.e. Long Short Term Memory units can be used instead of Gated Recurrent Units for a slight better accuracy. GRU's are preferred than LSTM as they are similar performance wise and much easier to integrate.

We think with this architecture the goal of creating Danish to English Machine Translator can be satisfied.