English/Spanish Translation with Language Identification - Project Milestone

Alex Shah

11/12/17

0.1 Abstract

This project's goal is to combine different architectures of neural networks to form a functional English/Spanish translator with the ability to detect the input language and translate to the corresponding opposite language. The resulting product combines a network to classify the input language with a sequence to sequence based recurrent neural networks for creating contextually aware translations.

0.2 Introduction

Software translation is a massively useful tool to bridge the gap between cultures. Your cell phone does not need a desktop PC's compute power to leverage neural networks improving the quality of software translation. However training a neural network does require a large amount of compute power and time.

Training a model is more efficient if we can train both directions simultaneously. By using encoder/decoder as a basis, there is an interim step (or steps) by which languages and context are broken down into mathematical representations the network is able to use to translate any language to any other. A neural network does not specifically need to train translation from language A to language B and vice versa, but can learn language A to any language based on that language A's features.

In addition, a translation is only as good as it's accuracy. Context is key when comparing the accuracy of various models. By utilizing language agnostic training, we can improve the speed at which we train. While beneficial to training time, this method is also beneficial to accuracy. Due to the necessary level of abstraction where the model learns feature sets, we can use attention mechanisms to fine tune the level of context to improve accuracy results.

0.3 Background

0.3.1 Language Classification

Classifying a given input is handled by a Convolutional Neural Network (CNN). It is crucial to quickly determine the input language as identifying the language is an intermediary step before translation. However, incorrectly determining the input language will create incorrect translations. Therefore it is important that the classification network is able to quickly but accurately detect the input language.

0.3.2 Contextually Aware Translation

Strides have been made using deep learning and deep network models for Machine Translation (MT) accuracy, making MT ever more accurate in contextually aware translation. Sequence to Sequence networks (Seq2Seq) are built like an encoder/decoder. The encoded input text is examined to determine the decoded output in the target language. An attention mechanism is used to limit the amount of backward and forward information sharing. Too much or too little information while determining context can have adverse effects on translation.

Neural Machine Translation (NMT) is a specifically derivated method of Seq2Seq focusing on accuracy and efficiency of training translation models.

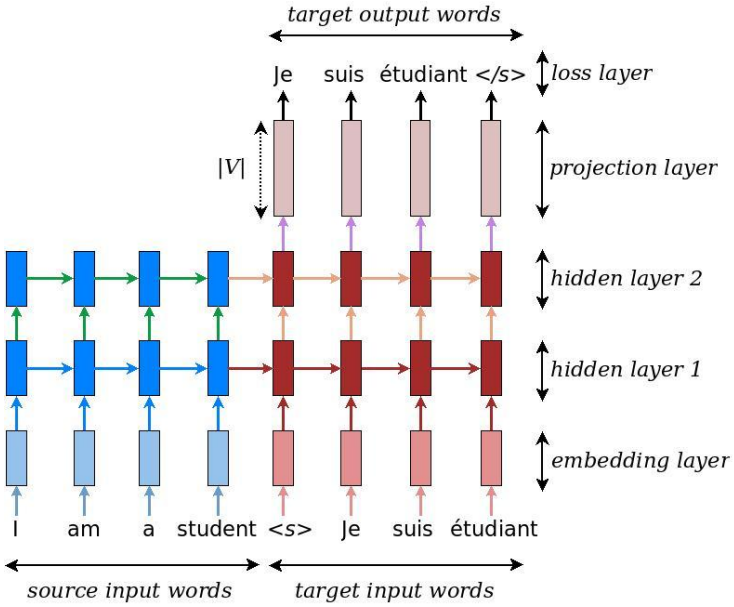

Figure 1: Sequence to Sequence Model (Luong, 2017)

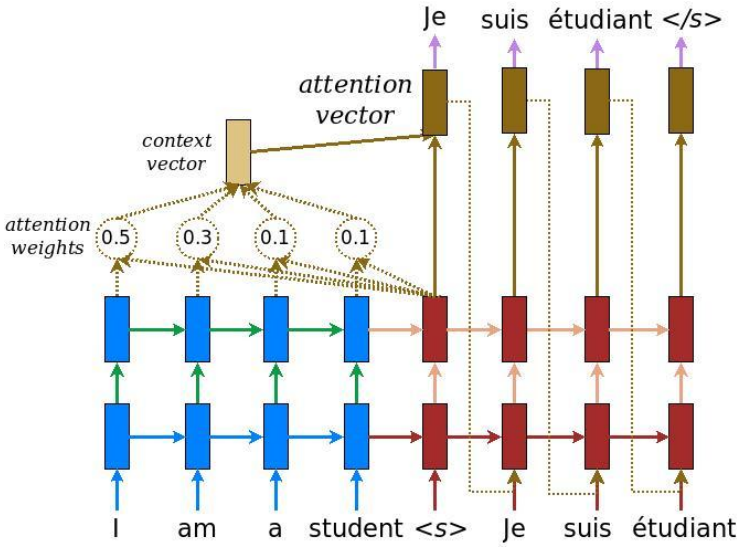

Figure 2: Sequence to Sequence Model with Attention Mechanism (Luong, 2017)

0.4 Methodology

In this project there are two networks. A Convolutional Neural Network is used as a language classifier, to detect whether a given input is English or Spanish. Next a Sequence to Sequence network, specifically using Google's Nural Machine Translation framework translates the given input to the opposite language. For example a given input is identified as English and translated to Spanish. And vice versa. The structure of the program is divided into the two main components. Classify will determine the input language, Translate will translate the input. The main program will use these two functions. Training will be done separately from the application's client facing usage. In this case, the end product will remain a command line interface but can be implemented in an API or web interface easily.

0.5 Experiments

Datasets in Spanish were difficult to find, but there are many datasets in other languages. In order to train translation, the model needs to study the target language and human corrected translations. In order to translate from one language to another, it is not necesary to train the model withthe input language and human corrected translations of the target language, but rather any human corrected language. In this way a trained model can train on multiple languages without need for datasets to overlap languages, or even add a new language at a later date.

The data used in this project comes from hand translated European Parliament records dating from 1996 to 2011. The dataset contains nearly 2 million sentences hand translated between English and Spanish.

The classifier will be trained on the same dataset which the Spanish and English corpuses reach almost 54 million words each.

0.6 References

0.6.1 Further Reading

https://github.com/tensorflow/nmt
https://github.com/google/seq2seq
http://www.nltk.org/

0.6.2 Articles

https://nlp.stanford.edu/projects/nmt/
https://research.googleblog.com/2017/07/building-your-own-neural-machine.html
https://sites.google.com/site/acl16nmt/
https://github.com/lmthang/thesis
https://google.github.io/seq2seq/data/

0.6.3 Papers

https://arxiv.org/pdf/1611.04558v1.pdf
https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
http://aclweb.org/anthology/D/D14/D14-1179.pdf
https://arxiv.org/pdf/1409.0473.pdf
https://arxiv.org/pdf/1508.04025.pdf
https://arxiv.org/abs/1609.08144
https://arxiv.org/abs/1703.01619

0.6.4 Data

http://www.statmt.org/europarl/
https://conferences.unite.un.org/UNCorpus
http://www.statmt.org/wmt17/translation-task.html