**MACHINE TRANSLATION USING SEQUENCE TO SEQUENCE MODELS**

12.02.2018

Rachana Movva

Manasa Bhimaraya Singhekar

Final Project: INFO 7374 Cognitive Computing and Deep Leaning

# Overview:

Machine translation uses software to translate text or speech from one language to another. Machine translation engine performs simple substitution of words in one language for words in another, but that alone doesn’t usually produce the highest quality translation of a text. For more accurate translation, recognition of whole phrases and their closest counterparts in the target language is needed.

Although big players like Google Translate and Microsoft Translator offer near-accurate, real-time translations, some “domains” or industries call for highly-specific training data related to the particular domain in order to improve accuracy and relevancy.

# Goals:

* Unlike the announced Google Neural Machine Translation system (GNMT) to achieve the largest improvements to date for machine translation quality using DNNs, we aim to show a straightforward application of the Long Short-Term Memory (LSTM) architecture. This solves general sequence to sequence problems whose dimensionality of the inputs and outputs is of variable lengths, thereby improving accuracy and relevancy of making translations from English to French.
* The idea is to use one LSTM to read the input sequence, one timestep at a time, to obtain large fixed dimensional vector representation, and then to use another LSTM to extract the output sequence from that vector. The second LSTM (handles long term dependencies) is essentially a recurrent neural network language model except that it is conditioned on the input sequence.
* Each input and output sequence must be encoded to integers and padded to the maximum phrase length. This is because we will use a word embedding for the input sequences and one hot encode the output sequences.
* Our model reads an input sentence “ABC” in English and produces “WXYZ” in French as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short-term dependencies in the data that make the optimization problem much easier.
* Different architectures of LSTM’s in Sequence to sequence models and recently released attentions which has shown high accuracy in machine translation problems would be tried.

# Use Cases:

To make translations from English to French with high accuracy and relevancy.

# Data:

**Europarl Machine Translation Dataset**

Europarl is a standard dataset used for statistical machine translation, and more recently, neural machine translation.

Link: http://www.statmt.org/europarl/

# Process Outline:

1. Data Preprocessing

* Data Cleaning-Removing punctuations, stop words etc.
* Tokenizing
* Converting English Text to Sequences
* Converting French text to One-hot encodings

2. Study of different architectures of LSTM’s and train different models

3.Evaluate the trained models and select the best model for translation

# Evaluation:

The evaluation for the model will be based on an inference model.

The inference model will leverage all the network parameters learnt during training but we define them separately because the inputs and outputs during inference are different from what they were during training the network.

# Milestones:

|  |  |
| --- | --- |
| **Timeframe** | **Delivery** |
| Day 1-4 | Data Cleaning |
| Day 5-9 | Training and Choosing the best model for Translation |
| Day 10-11 | Documentation |

## 

# Deliverables: Model that efficiently converts from English to French.

## **Reference and Sources:**

# Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation - https://arxiv.org/abs/1406.1078

# Sequence to Sequence Learning with Neural Networks - <https://arxiv.org/abs/1409.3215>

# Attention Is All You Need - https://arxiv.org/abs/1706.03762