Initial Project Overview

SOC10101 Honours Project (40 Credits)

Title of Project:

An Analysis of Neural Machine Translation Approaches for a Low-Resource Language (Scottish Gaelic)

Overview of Project Content and Milestones

To research, implement and analyse the different approaches used in Neural Machine Translation and determine which are best suited for low resource languages

Milestones:

- Introduction complete
- Literature review complete
- NMT training data obtained and cleaned
- NMT models implemented
- Analysis of the NMT models conducted
- Write remaining parts of the dissertation based on the work carried out

The Main Deliverable(s):

- A literature review that covers prior research which identify various approaches in the area of Neural Machine Translation and the challenges that need to be overcome in order to improve translation quality for low-resource language training data.
- Multiple models for machine translation to Gaelic languages
- Visualisations to help demonstrate the accuracy of translations from each model
- A detailed analysis of the different neural models

The Target Audience for the Deliverable(s):

Other researchers and people working in translation or artificial intelligence that have an interest in machine translation or low-resource languages

The Work to be Undertaken:

- General NMT research
- Low resource language translation research
- Literature review on NMT with a focus on low resource languages
- Implement NMT using a high resource language (to demonstrate the baseline performance of a high resource language)
- Collect a large amount of quality training data for the low resource language
- Implement a basic model using the NMT and training data
- Implement complex / alternative models using NMT on the training data
- Benchmark the models and rank them (BLEU score etc.)
- Create visualisations of the results to demonstrate the accuracy of the translation by looking at the attention of the model
- Carry out an analysis of the models based on their individual results
- Write up the dissertation based on the findings of the NMT analysis

Additional Information / Knowledge Required:

Prior to any implementation I need to gain a more thorough understanding of the theory that underpins deep learning and NMT. I will then research and experiment with NMT implementations in Python, extending my current experience with Python development. Another area of knowledge required for the project is the evaluation techniques for determining the effectiveness of the translation models. These techniques will be important for conducting a thorough analysis of the results.

Information Sources that Provide a Context for the Project:

- Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation: https://arxiv.org/pdf/1609.08144.pdf%20(7.pdf
- Corpus Augmentation by Sentence Segmentation for Low-Resource Neural Machine Translation: https://pmt-eu.hosted.exlibrisgroup.com/permalink/f/1aeuh09/TN_proquest2229493935
- Neural machine translation for low-resource languages without parallel corpora: https://pmt-eu.hosted.exlibrisgroup.com/permalink/f/1aeuh09/TN_springer_jour10.1007/s10590-017-9203-5
- Leveraging back-translation to improve machine translation for Gaelic languages http://doras.dcu.ie/23599/1/Backtranslation_Gaelic_languages.pdf
- Machine Translation Evaluation: https://www.cs.cmu.edu/~alavie/papers/GALE-book-Ch5.pdf
- OPUS parallel corpus Scottish Gaelic to English dataset http://opus.nlpl.eu/
- LearnGaelic learning materials dataset https://www.learngaelic.net/

The Importance of the Project:

Neural Machine Translation has greatly improved the quality of translation. However, current methodologies depend heavily on using large quantities of training data. This is a problem for low-resource languages as there is much less training data available and current models that are trained on a low quantity of parallel data often produce low quality translations. As a result, there is a demand in NMT for models that are able to perform well despite having little training data.

By carrying out an analysis of the different approaches available for low-resource language NMT, it will be clear which approach is best suited for the context of the translation. As mentioned earlier, this is important because approaches that work well for high-resource languages do not necessarily work well on low-resource languages.

The Key Challenge(s) to be Overcome:

- Obtaining enough quality training data for the low resource language
- Cleaning any training data that I obtain for the models. Any problems with the data (formatting, spelling mistakes, etc.) will impact the results which would make any analysis inaccurate.
- Finding NMT methods that are different enough to have a variety of BLEU scores based on the limited training data