Comp8715 Project Plan

Project title: Deep check: Deep Learning based Grammar Checker
Supervisor: Dr. Lizhen Qu
Student: Kiarie Ndegwa (u4742829)

Problem definition:
There are millions of people around the world trying to learn the English language effectively. However, crossing over into the native speaker threshold continues to remain a challenge due to nuanced language idiosyncrasies; which fall out of the demarcation of semantic or syntactic formalism. These idiosyncrasies are therefore mostly learnt the time consuming traditional way i.e. by repeatedly making errors and getting corrected.

More often than not these lapses of "incorrect" English can be attributed to collocations errors. These include multiword expressions(MWEs) such as "to kick the bucket" as opposed to "kick the buckets", or "telephone booth/box" as opposed to "telephone cabinet". There are currently no effective analytical approaches to the problem as it is estimated that there exist as many collocation relationships as there are words in the English language.

This project wishes to "take the bull by the horns" by implementing a sequence to sequence deep network that will serve as a state of the art grammar error checker. The projects primary focus will be on developing a grammar corrector based on an architecture based on a sequence to sequence stacked recurrent neural networks(RNNs).

It is intended that Long Short Term Memory(LSTM) or Gated recurrent unit (GRU) will serve as the primary computational unit in this deep network architecture. As they have the critical advantage of being able to model contextual information contained within sentences of arbitrary length. Common to MWEs as shown above. Thus acting as a one-size fits all solution to the different flavors of MWE’s and their respective unique challenges.

Tasks:
1. Understand language models and how they are generated using neural networks
2. Understand the how recurrent neural networks function both theoretically and when implemented on GPUs.
3. Understand the fundamental formal challenges involved in natural language processing.
4. Learn to critically analyze literature from current Natural Language Processing journals.
5. Learn to use the Torch framework effectively and GPU clusters.
6. Implement and exhaustively test a deep network algorithm wholly designed throughout the duration of this project. Compare results generated during this project with the current state of the art neural network grammar correctors.

Schedule:
Week 1 -3
- Research on various architectures.
- Be familiarized with the components of neural networks.
- Learn computational graphs and their implementation on Torch and Tensor flow.
Week 4 -6

- Implement simple stock standard algorithms on Torch and run them on NICTAs computational facilities.
- Begin research onto the more advanced components of the RNNs.
- Be familiarized with the data sets and the state of the art that will be used to train and benchmark.
- Unsupervised training: Get model to learn language from untagged corpus such as the billion word Wikipedia corpus

Week 7 -8

- Supervised training: Get model to correct error prone sentences.
- Test various architectures, i.e. modify attention and highway network designs
- Start writing thesis – finish literature review.

Week 9-10:

- Get model to replicate common collocation errors.
- Get model to generate corrected sentence alternatives.
- Optimize model to train using less computational resources.

Week 11-12

- Continue evaluation and comparison with other approaches
- Finish writing thesis – methodology, analysis, conclusion.
- Prepare final presentation.