Progress Report

Project Details

Name: Jiaxin Wang
Email: jw2117@cam.ac.uk
Project Title: A Recursive Recurrent Neural Network Decoder for Grammatical Error Correction
Name of Supervisors: Zheng Yuan (primary), Christopher Bryant
Name of Director of Studies: Thomas Sauerwald
Name of Overseers: Alan Blackwell, Srinivasan Keshav

Work Accomplished

Moses SMT *Due 5th Nov / Completed 15th Oct*

[Original Plan: Data preprocessing]

Obtained standardised data (FCE corpus) in m2 format from BEA19[1]. Converted it into sentence-aligned format which is required for moses training.

[Original Plan: Language model training]

Built a **5-gram language model** using lmplz (which is a KenLM language model creation program that comes with moses installation). Trained with One Billion Words[2] data. Binarised the model for faster loading.

[Original Plan: Translation model training]

Obtained **word alignment** using GIZA++. Trained a translation model using preprocessed FCE training data. A **phrase table** was produced.

[Original Plan: Tuning with development data]

Tuned the translation model with preprocessed FCE development data.

[Original Plan: Evaluate with testing data]

Evaluated the performance of moses SMT using ERRANT[3].

========== Span-Based Correction ==========

TP	FP	FN	Prec	Rec	F0.5
479	632	4070	0.4311	0.1053	0.2663

==========

R^2NN

[Original Plan: Phrase pair embedding for sparse features] *Due 19th Nov / Completed 17th Dec*

Data preprocessing: obtained one-hot encoding for each sentence in the training file using information from the **phrase table**.

Implemented a **one-hot-layer neural network**. Trained the network with sentences in the training file.

[Original Plan: Phrase pair embedding using rnn] *Due 3rd Dec / Completed 23rd Jan*

---

[1] https://www.cl.cam.ac.uk/research/nl/bea2019st/#data

[2] https://opensource.google/projects/lm-benchmark

[3] https://github.com/chrisjbryant/errant

Data preprocessing: obtained word embedding using fastText[4].
Implemented a **recurrent neural network**. Trained the network with word embeddings for each word in the training file.

Unexpected Difficulties

[Original work schedule]
Problem:
The translation confidence based phrase pair embedding (TCBPPE) in the R^2NN paper[5] is more complicated than I originally expected (more details below). Hence, even though I finished building a moses SMT three weeks ahead, my overall progress is four weeks behind.
Actions:
My original plan gives two weeks for model evaluation and four weeks for possible extensions. I will devise a new work plan to catch up with the progress.

[Phrase pair embedding for sparse features]
Problem:
In the R^2NN paper, the authors used forced decoding to get positive samples. I couldn't get moses work with forced decoding, and there seems to be no documentation about it online that could help with my confusion.
Actions:
Instead of using forced decoding, I will store the state of my neural network after each training epoch. When it comes to testing, I will use the neural network state that has been trained for 10 epochs. I can experiment with other stored neural network states as an extension task.

New Work Plan

Preprocess testing data to be used by the **one-hidden-layer neural network** and the **recurrent neural network.**	*22nd Jan - 4th Feb 2022*
Build a recursive recurrent neural network (R^2NN).	*5th Feb - 18th Feb 2022*
Train the parameters of R^2NN decoder.	*19th Feb - 4th Mar 2022*
Evaluation of the performance of R^2NN decoder.	*5th Mar - 18th Mar 2022*
Write the dissertation. Start with possible extension.	*19th Mar - 22nd Apr 2022 [Easter Break]*
Finish writing dissertation and send in the draft for review.	*23rd Apr - 6th May 2022*
Dissertation Deadline	Friday 13th May 2022 (12 noon)

[4] https://fasttext.cc/
[5] https://www.microsoft.com/en-us/research/wp-content/uploads/2014/06/P14-1140-2.pdf