

# 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<sup>1</sup>. Converted it into sentence-aligned format for Moses training.

*[Original Plan: Language model training]*

Built a **5-gram language model** using Implz (which is a KenLM language model creation program that comes with Moses installation). Trained with One Billion Words<sup>2</sup> 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<sup>3</sup>.

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

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

=====

### **R<sup>2</sup>NN**

*[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 the **phrase table**.

Implemented a **one-hidden-layer neural network** which gives a new feature for each phrase pair. Trained the network with sentences in the training file.

Evaluated the performance of the sparse feature using ERRANT. It gives a lower F0.5 score than the baseline (Moses), but it is still a work in progress and not a full model.

---

<sup>1</sup> <https://www.cl.cam.ac.uk/research/nl/bea2019st/#data>

<sup>2</sup> <https://opensource.google/projects/lm-benchmark>

<sup>3</sup> <https://github.com/chrisjbryant/errant>

```

===== Span-Based Correction (5 epochs) =====
TP    FP    FN    Prec  Rec  F0.5
543   2366   4006   0.1867 0.1194 0.1677
===== Span-Based Correction (10 epochs) =====
TP    FP    FN    Prec  Rec  F0.5
545   2363   4004   0.1874 0.1198 0.1684
=====

```

*[Original Plan: Phrase pair embedding using rnn]*

*Due 3rd Dec / Completed 23rd Jan*

Data preprocessing: obtained word embedding using fastText<sup>4</sup>.

Implemented a **recurrent neural network**. Trained the network with word embeddings for each word in the training file.

## Unexpected Difficulties

*[Original plan: Build  $R^2NN$ ]*

Problem: The translation confidence based phrase pair embedding (TCBPPE) in the  $R^2NN$  paper<sup>5</sup> took longer to implement than I originally expected. Hence, even though I finished building a Moses SMT three weeks ahead, my overall progress is four weeks behind.

Actions: I will adjust my work plan to catch up with the progress.

*[Original plan: Use forced decoding during model training]*

Problem: 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: I will use the neural network that has been trained for 10 epochs instead of using forced decoding and revisit this as an extension if there is time.

## 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)

---

<sup>4</sup> <https://fasttext.cc/>

<sup>5</sup> <https://www.microsoft.com/en-us/research/wp-content/uploads/2014/06/P14-1140-2.pdf>