Effective Grammatical Error Correction with Neural Machine Translation Techniques

Shubha Guha, s1473099
Supervised by Kenneth Heafield
Progress Report

14 July 2017

Goals

Grammatical error correction (GEC) is the task of automatically transforming text with potential grammatical errors into grammatically correct text. Although both input and output text are in the same language, this transformation is similar to the task of translating text from one language to another. As a result, recent work in GEC has predominantly used techniques that have been refined for the task of machine translation (MT). At the same time, the tasks are sufficiently different that standard MT techniques must be altered to tackle the particularities of GEC. The fact that for GEC both input and output are in the same language means that in most cases (at a word level) they are even the same text, which can result in a model learning to copy the input text as output and failing to learn to correct grammatical errors.

The goal of this project is to apply neural machine translation (NMT) methods to GEC with appropriate modifications to make learning more effective. Since the main difference between MT and GEC is that there is a much higher occurrence of input words being identical to output words, the main idea behind this thesis project is to apply a higher weighting to the training loss from words that are not identical between input and output sequences.

Methods

The early stages of the project timeline were spent in familiarization with tools, packages, and cloud services. Initially on Google Cloud, then Microsoft Azure, various baseline models were trained with Nematus (ultimately only the git version at commit 73037e9). All baselines were trained using cross-entropy minimization as the training objective, while details around data pre- and postprocessing and model architecture were altered from one baseline to the next.

The final baseline model was trained on the NUCLE and Lang-8 corpora (2.2 million parallel sentences), validated on the CoNLL-2013 test set (1381 sentences), and finally evaluated on the CoNLL-2014 test set (1312 sentences). These were all preprocessed with a trained Moses truecaser, followed by the application of byte-pair encoding using the subword-nmt package, and finally the replacement of all pipe symbols (|) with a special token (<pipe>).

The final baseline model's architecture used layer normalization and dropout (0.1 on the source and target layers, 0.2 on embedding and hidden layers), in addition to Nematus default settings such as an embedding size of 512, hidden layer size of 1000 units, and an encoder and decoder depth of one layer each. Training hyperparameters used were a batch size was 60 and Nematus defaults including an early stopping patience of 10 and the Adam optimizer.

Results

The final baseline model finished training after 590,000 minibatches. With the MaxMatch scorer from the CoNLL-2013 shared task, it achieved a precision of 0.3319 and recall of 0.1413 on the test set. This recall confirms the expected high incidence of false negatives, i.e. the models fails to correct many grammatical errors. The samples below in Table 1 confirm that the system has learned often to copy input sequences without correcting grammatical errors (false negatives), and in the case of Sample 3 even to make edits that are not necessary (false positives):

Source 1	Mizu@@ shima seaside industrial area is especially well known as one of the largest industrial area in Japan .
Truth 1	Mizu@@ shima 's seaside industrial area is especially well known as one of the largest industrial areas in Japan .
Sample 1	Mizu@@ shima seaside industrial area is especially well known as one of the largest industrial areas in Japan .
Source 2	it is very difficult for me to use " listen to " and " hear " properly .
Truth 2	it is very difficult for me to understand the difference between " listen to " and " hear . "
Sample 2	it is very difficult for me to use " listen to " and " hear " properly .
Source 3	I learnt English earlier than learning Japanese , but the latter is more skilled than the former .
Truth 3	I learnt English earlier than learning Japanese , but the latter I 'm more skilled at than the former .
Sample 3	I have learnt English earlier than learning the Japanese , but the latter is more skilled than the former .

Table 1: Samples generated by fully trained baseline model.

To Do

At the time of this writing, a new objective function is being added to Nematus that will multiply the existing cross-entropy cost vector for each batch of words by a weight vector that identifies and emphasize which words of the batch must be edited from source to target sequence. Once implemented, another model will be trained using the same model architecture, training hyperparameters (except for the training objective), and data preprocessing that were used to train the baseline model. Once evaluated on the test set, we expect to find a higher recall.

If time remains after these results are computed and the final writing is being completed, we hope to train a second baseline model using minimum risk training (MRT) with MaxMatch $F_{0.5}$ score as the scoring function, and a final model using MRT with a different word-level weighted scoring function.