Text Corruption Detection and Generation

Hao Zou
Department of Computer Science and Engineering
University of Minnesota, Twin Cites
Minneapolis, MN 55414
zou00080@umn.edu

1 Part 1

Nowadays Transformer achieves State-Of-The-Art performance for many NLP tasks, hence I just treated this problem as a simple binary classification task which might be the simplest way. Pre-trained BERT was used to fine-tune the corpus for classification issue (logistic regression). AdamW from HuggingFace was used as optimizer to fix weight decay. 4 epochs were trained with 2-hour average for each epoch. The model with smallest valid loss 0.34 was chosen for testing. The valid accuracy was 0.86. To advanced the performance, the feature of paired-data can be taken into consideration, like making the object of BERT be the score of English sentence minus the score of Corruption sentence, then make the inference based on that after training.

Other methods for rule-based error type checking: *ERRor ANnotation Toolkit*, which extract and classify grammatical errors in parallel original and corrected sentences.

2 Part 2

It is natural to approach the error correction task as a machine translation problem from incorrect into correct English. Hence I spent two day working on OpenNMT training the provided dataset to fulfill the research purpose. However, I soon realized that the additional training corpus was not provided and I was using the same corpus for both training and generating part. Based on that, this task cannot be treated as a Seq2Seq generation problem.

I then turned to rule-based ill-formed sentences generation and found a toolkit GenERRate, a good tool for automatically introducing syntactic errors into sentences. It was introduced by Prof. Jennifer Foster. Unfortunately, this toolkit has not been maintained for a long time and is not easy to use. Hence I just generate corruptions from scratch based on their paper and related work. This might mot be the most efficient way to automatically introduce syntactic errors into sentences. But since we are not allowed extra data or derivatives, this rule-based method might be the most direct way.

Four different kinds of grammatical and ill-formed errors were introduced into train English sentences. [2] reported one ordering of the substitution, deletion and insertion correction operators in a study of native speaker spoken language slips, which is:

$$substitute(48\%) > insert(24\%) > delete(17\%) > combination(11\%)$$

The same ordering was used for the errors generation. The dataset was shuffled for random purpose. Here are some examples of those error types:

1. **Missing Word Errors:**
   There is no European code of ethics, however, and there are **no** agreements at national level .
   There is no European code of ethics, however, and there are agreements at national level .

2. **Extra Word Errors:**
   The wand given to Alora transforms into a spear .

Preprint. Under review.

The wand **wand** given to Alora transforms into a spear .

3. **Real-Word Spelling Errors:**

Vrenna drew her saber and San'doro drew his knives .
Vrenna drew her saber and San'doro drew his **knifes** .

4. **Agreement Errors:**

Returning home from working the graveyard shift , Jill is shocked that Molly is missing .
Returning home from working the graveyard shift , Jill is **shocking** that Molly is missing .

Missing word errors were classified based on the part-of-speech (POS) of the missing word. The frequency distribution was picked from [1]'s work:

$$det(28\%) > verb(23\%) > prep(21\%) > pro(10\%) > noun(7\%) > "to"(7\%) > conj(2\%)$$

For each sentence, all words with the above POS tags are noted. One of these is selected and deleted. For those sentences that do not contained POS tags above, other tags were selected randomly and deleted. Real-Word Spelling Errors were classified as replacing the erroneous word with another word with a Levenshtein distance of one from the erroneous word, and Agreement Errors were classified as replacing a singular determiner, noun or verb with its plural counterpart [1]. Here for simplicity sake, these two errors were categorized into one substitute error. Characters were randomly placing between a random choice of words for corresponding sentences. Rule-based methods can hardly be comprehensive or comparably efficient, large room of possibilities can to be extended to handle a wider class of errors for this case.

References

[1] Jennifer Foster. Good reasons for noting bad grammar : empirical investigations into the parsing of ungrammatical written english. 2005.

[2] J. Stemberger. Syntactic errors in speech. *Journal of Psycholinguistic Research*, 11:313–345, 1982.