Erroneous data generation for Grammatical Error Correction

Shuyao Xu, Jiehao Zhang, Jin Chen, Long Qin

Singsound Inc.

Introduction

The most effective approaches to Grammatical Error Correction (GEC) task are machine translation based methods. Both Statistical Machine Translation (SMT) approaches and Neural Machine Translation (NMT) methods have achieved promising results in the GEC task.

Utilization of monolingual data for training neural GEC models has been demonstrated to be an effective approach to improve GEC system performance. Following this path, we investigate an approach by systematically generating parallel data for pretraining.

Orig: the primary is open to independent voters .
Gen: the primary is opens to voters independhent .

Orig: the price of alcohol is ramped up at every budget .
Gen: the puice of alchool is ramping up at every budget .

Orig: i do think that some fundamental reforms need to take place .
Gen: i do thought that some fundamentals reforms need to take placement .

Figure 1. Examples of generated data

Pretraining Data Generating Method

Introduce five types of errors to sentences:

- Concatenation: combine two consecutive tokens,
  e.g., hello world → helloworld.
- Deletion: delete the token.
- Transposition: the token exchange position with a consecutive token.
- Misspelling: introduce spelling errors into words,
  - Deletion: delete the character.
  - Insertion: insert a random English letter into the current position.
  - Transposition: exchange position with the consecutive character.
  - Replacement: replace the current character with a random English character.
- Substitution: seven different types of substitutions.
  - Substitution between Prepositions. E.g., in, on, at, through, for, with.
  - Substitution between Articles. E.g., a, an, the.
  - Substitution between Pronouns (Singular). E.g., he, she, his, him, her, hers.
  - Substitution between Pronouns (Plural). E.g., their, them, they, theirs.
  - Substitution between Wh-words. E.g., which, where, what, how, when, who ...
  - Substitution between Modal verbs. E.g., will, shall, can, may, would, …
  - Substitution in a Word Tree (e.g., Figure 3).

Generate pretraining data by stochastically inserting the above error types into correct sentences.

	# Sentences	# Tokens
WMT11+1 Billion Words	145 M	3,100 M
Lang8+FCE+NUCLE+ABCN	1.1 M	14 M

Table 1. Statistics of pretraining data and finetuning data

Word Tree

Introduce Word Tree to make substitutions such as going → gone, useful → usable, administration → administrative possible.

A Word Tree represents a group of words that share the same stem but have different suffixes.

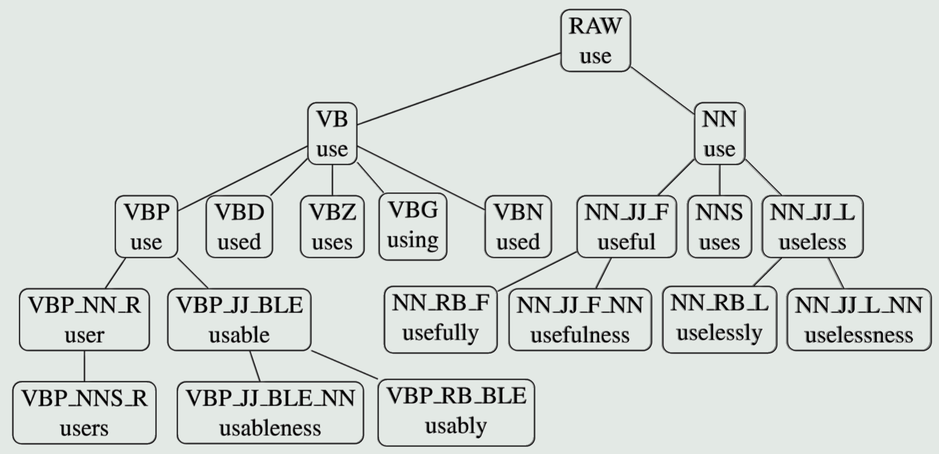

Figure 2. Word Tree: use

Results

The Singsound GEC system is an ensemble of 4 transformer models, which achieve an F0.5 of 66.61 during the GEC shared task. Later, We managed to push the performance up to 69.85 on the ABCN test.

	DEV	TEST		
	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$
Single	53.87	69.26	59.84	67.15
4 Ensemble	55.81	72.99	59.59	69.85

Table 2. Results on the ABCN set

In my opinion its the best film I saw in my life and if you have the oportunited to see , do it .
In my opinion , it's the best film I have seen in my life and if you have the opportunity to see it , do it .

It need us to take long time to separate them .
It takes us a long time to separate them .

He go school yesterday ?
Did he go to school yesterday ?

Figure 3. Examples of system outputs

Conclusion

We present a novel erroneous data generating method for training English GEC models and demonstrate its effectiveness in the GEC task. We also present a novel tool: the Word Tree, and show that one possible application of the Word Tree is generating erroneous text for training GEC models.