Initial project idea

Who is on the team?

- Mikhail Tokarev
- Nuttaree Busarapongpanich
- Rashmi Jadhav
- Yu-Wen Chen

A tentative title: Save your words

It can be interpreted in two ways. 1) An article is constructed by a bunch of words. Therefore, it literally "saves" your article by paraphrasing and extending the idea. 2) It also has the meaning of saying "Save your word. I've reached the word count limit you set." to your teacher.
Our intention is the first one.

Brief description of project goals and proposed methods

Languages, being a communication medium, play an important role in human lives. Many tools out there in the world try to make sentences more concise, aiming for the sentences to be easier for readers to understand. However, we propose an idea of expanding sentences and paragraphs whilst also preserving the original meaning or intent of the text. There are many challenges in this project. Some of them are preserving the grammar and sentence structure, preserving the meaning and context(the measure of which is dependent on humans) and, avoiding plagiarism in the generation of new sentences.

Goal:

Our goal is to represent ideas in a linguistically more elegant way. One example could be that exams like TOEFL/IELTS require us to use fancy words or sentence forms which we do not use in daily lives while communicating. Even research papers in a particular field have a language that people usually follow. That being said, each field would have some writing standards or language in their literature. Such tasks thus require insertion of words in the basic form of text and/or substitution of some words with another. Here, we want to build a model which can improve the level of English of the text provided or rather make the language conform to the literature. Additionally, we want to expand the sentences to make them longer and keep the semantic sense of the sentence. Whilst doing this, we need to also make sure that we do not end up plagiarizing from the corpus itself.

There are 2 approaches that we could have:

1. We will replace some words with the synonym and add more words (adjective or adverb) to the sentence. For example, let's take the first sentence from the goal. "Our goal is to represent ideas in a linguistically more elegant way." we can change to "Our objective is to represent ideas in an etymologically more beautiful way." (This is one of the first method)

2. We can change the whole sentence by changing its whole structure. For instance, we can change the phrase from “The home-made food” to “The food I am making at home” (Transform the phrase), or change from “I have done the research” to “the research has been done by me” (Transform the sentence, change of active-passive voice)

Proposed methods:

We will apply the Natural Language Processing technique with Deep learning to this work. We will be using a dataset called - summarization dataset. This dataset includes 1.3M articles and summaries for them written manually by writers and reporters.

1. word2vec, End-to-End memory:
   We want to use the article dataset as a training set and implement learned GloVe or PPMI word to vec [1] dependencies in order to capture semantic of words and insert words or find synonyms where they look like a good fit. If this seems too easy we can use semantic implementation of sentence meanings in order to make sure that we keep the same idea of the text/sentence. But it is mostly for fun since DNN which checks itself with another DNN does not look very reliable. For this method, we could also use the End-to-End memory approach [3] where we have the whole sentence and we can get the meaning from the whole sentence. We can also use this approach to find the part-of-speech tagging to see where we should add words. This will yield the first outcome where we replace the word by its synonym and add more words in between the sentence.

2. BERT:
   We can also group the sentences together using BERT [9] to extract the meaning of sentences and pick the longest sentence that has the most similar meaning.

3. LSTM (Bidirectional LSTM)
   We can use LSTM [2] to achieve the text generation by creating the stories from the given input stories.

4. Text transformer:
   The text transformer would be another good choice that can be used to achieve this work. We can use a text transformer approach to transform new sentences based on the main idea we have. [6] is a demo on how the text transformation works.

5. Skip-Thought Vector (RNN encoder with GRU):
   Using Skip-Thought Vector [4]. Building an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage. Instead of using a single word to predict its surrounding context, we can encode a whole sentence to predict the sentence around it. By this, we can just generate whole new sentences by passing a sentence as an input.

Reference to at least 3 related works

Papers:

Word2vec & GloVe

[1] https://academic.oup.com/jamiaopen/article/2/2/246/5423083

LSTM (Bidirectional LSTM):

[2] https://pdfs.semanticscholar.org/c51d/13034b2df47dae8f33bd0efad996de99ed4c.pdf

End-to-End memory

[3] https://www.aclweb.org/anthology/W17-7305.pdf

Skip-Thought Vectors (RNN encoder with GRU):

[4] https://arxiv.org/pdf/1506.06726.pdf

Articles:

LSTM:

[5] https://towardsdatascience.com/sentence-classification-using-bi-lstm-b74151ffa565

Text transformer:

[6] https://app.inferkit.com/demo
[7] https://medium.com/phrasee/neural-text-generation-generating-text-using-conditional-language-models-a37b69c7cd4b
[8] https://huggingface.co/blog/how-to-generate

BERT

[9] https://towardsdatascience.com/cutting-edge-semantic-search-and-sentence-similarity-53380328c655

Dataset:

http://lil.nlp.cornell.edu/newsroom/explore/index.html#