Finding Similar Examples for Aiding Academic Writing using Sentence Embeddings

Chooi Ling GOH
The University of Kitakyushu

Yves LEPAGE
Waseda University

{goh@kitakyu-u.ac.jp, yves.lepage@waseda.jp}

1 Introduction

Non-native speakers of English always face problems to compose an article in English. This problem becomes more severe when coming to scientific or academic writing. One way of learning writing scientific articles is to refer to previous published papers. Most of the time, the articles in the same field tend to use similar expressions, vocabularies and writing style. For example, previous published articles in ACL Anthology Reference Corpus (ACL-ARC) [1] could be used as a reference for writing an article in natural language processing (NLP) field, and PubMed articles can be referred in life science and biomedical field. For a non-native speaker, one may produce a simple sentence such as below.

```
Our experiment results are better than
baseline.
```

However, in previous articles, someone has written some similar sentences like:

```
1. Experiments show that our approach
   achieves significantly better results
   than baseline methods.
2. The experimental results show that our
   model still achieves better performance
   than the baseline.
3. Preliminary experiments showed that
   our approach was more effective than
   baseline methods.
```

The writer can refer to these sentences and try to compose a new sentence that is better than the original one like:

```
Our experiment results show that our
model achieves significantly better
performance than the baseline.
```

Recently, sentence embeddings have been proven to improve many NLP downstream tasks. One of the advantages of sentence embeddings is that they can capture the meaning of the sentences and compare the similarity between them. In this paper, we will use sentence embeddings to search for similar sentences in previous articles, and show that they are useful for aiding academic writing.

[1] https://acl-arc.comp.nus.edu.sg

2 Related Work

The recent language representation model BERT [6] has called an attention to the NLP field, where it has been proven that it can improve a lot of NLP downstream tasks, such as question answering and language inference. However, due to computational complexity, it has problem on sentence-pair regression task like semantic textual similarity (STS). This problem has been overcome by Sentence-BERT [7], a sentence based BERT model, applying siamese and triplet networks [8], in order to derive semantically meaningful sentence embeddings.

Stanford Natural Language Inference (SNLI) dataset [1] has been used in many researches. InferSent [5] uses this labeled dataset and shows that it has outperformed the unsupervised methods. The unsupervised learning for Universal Sentence Encoder [3] using various web sources is also further augmented with SNLI dataset. Similar to InferSent [5], training on SNLI improve the results for the transfer tasks in SentEval [4]. All these researches found that SNLI datasets are suitable to train sentence embeddings.

3 Methods

We use the implementation from Sentence-BERT [7] [2]. They propose a sentence transformer which consists of a word embedding model and a pooling model. Each sentence is passed through these two models and transformed into a fixed size sentence vector. They have provided some pre-trained sentence embedding models which is trained on SNLI corpus and semantic textual similarity benchmark (STSb) dataset [2]. Beside SNLI, they have shown that the pre-trained models could be used in other NLP downstream tasks, such as STS shared task, argument facet similarity, Wikipedia sections distinction and seven transfer tasks in SenEval toolkit [4].

We follow the same underlying idea by fine-tuning BERT pre-trained model (bert-base-uncased) to ACL-ARC. First, the BERT model maps the tokens in a sentence into the output embeddings.

[2] https://github.com/UKPLab/sentence-transformers

Then, a pooling model layer is added. We use the mean-pooling here as it has been shown to give better results for SNLI task in previous experiments. Finally, we train the sentence transformer with a triplet loss function.

The sentences in the ACL-ARC are labelled with the year of publication. For example, Y08 represents a paper published in 2008, and Y15 represents a paper in 2015. Totally, we have 37 categories, ranging from year 1979 to 2015. Using these labels, we build weakly labeled sentence triplets, where an anchor sentence is dynamically paired with a positive example which comes from the same category and a negative example which comes from a different category. In total, there are 21,665 articles with 3,992,933 sentences. Figure 1 shows the number of articles and sentences over the years. There are increasingly more papers published recently which shows that NLP field is advancing rapidly.

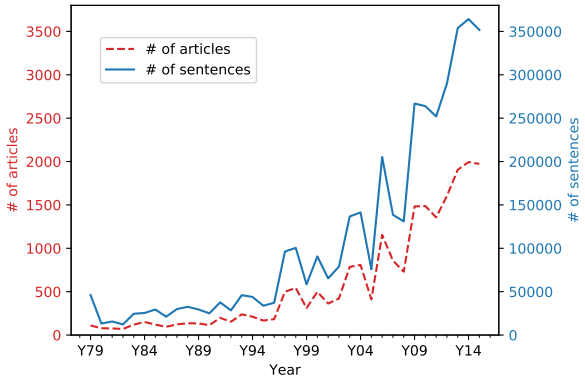

Figure 1: Number of articles and number of sentences in ACL-ARC each year from 1979 until 2015.

4 Experiment and Results

First of all, we train a sentence embedding model using only BERT and ACL-ARC. During the training, the development set is the STSb devset and we evaluate the models using STSb testset. Next, we train a model by fine-tuning the pre-trained Sentence-BERT models: SBERT-NLI-base and SBERT-NLI-STSb-base. We compare the results with the previous work stated in Sentence-BERT [7]. Table 1 shows the evaluation results[3]. Training with ACL-ARC only does not give good performance as expected since it does not suit to the STSb task. However, after fine-tuning with the pre-trained models, they have shown a slight improvement. This experiment shows that fine-tuned models trained on ACL-ARC encode reasonable sentence embeddings.

[3]The results of SBERT-NLI-base and SBERT-NLI-STSb-based are taken from github, which are slightly different from their paper [7].

Model	Spearman
SBERT-NLI-base	77.12
SBERT-NLI-STSb-base	85.14
SBERT-ACLARC-base	48.23
SBERT-NLI-ACLARC-base	78.52
SBERT-NLI-STSb-ACLARC-base	**85.19**

Table 1: Evaluation on the STSb testset. SBERT-ACLARC is trained only on ACL-ARC, SBERT-NLI-ACLARC is using SBERT-NLI pre-trained model and fine-tuned with ACL-ARC, and SBERT-NLI-STSb-ACLARC is using SBERT-NLI-STSb pre-trained model and fine-tuned with ACL-ARC.

5 Search in Abstract

Using the sentence embedding model trained in previous section, we search for similar sentences in ACL-ARC for some input sentences. For the first attempt, we use only the texts from abstracts as the search database. Out of 21,665 articles, we can only extract 18,786 abstracts, with 105,637 (102,683 unique) sentences. Figure 2 shows the distribution of sentence length in the abstracts. Due to some OCR errors, we have about 1% of abstracts that are noises, which could not be split into proper sentences and are very long (more than 60 words).

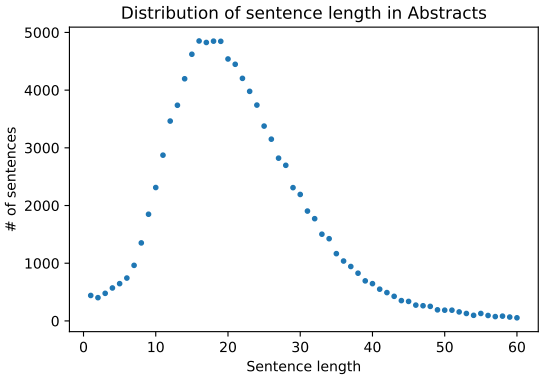

Figure 2: Distribution of sentence length in abstracts. This graph shows only sentences with less than or equal to 60 words. Only 887 sentences have more than 60 words which are almost noises.

For a second language learner, it is usually very difficult to comprehend long sentences. Therefore, we try to keep the sentences short. Since some of the sentences are very long, we further cut the sentences into N-gram phrases. From the distribution, we have fixed the N to 20, and slide through the sentence for every 10 words. That means, if a sentence has 38

words, it will be cut into 1˜20-word, 11˜30-word and 21˜38-word N-gram phrases. At the end, we produce 177,340 (172,998 unique) N-gram phrases in total. Combining the two sets, we have 282,977 (232,874 unique) sentences/N-gram phrases for search.

Table 2 shows an example of the search results, by descending order of cosine similarity. We can use some simple, incomplete, or even with grammatical error sentence as the query. We are able to obtain some reasonable search results. Not only the same words in the query could be found in the search results, but similar words with similar meaning could also be found. For example, "work better" could be replaced by "perform better", instead of saying "We want to show", it is better to use "Our experiment show" and etc. Furthermore, we also obtain some extra vocabularies such as "proposed system", "significantly" and "remarkably", which could be used to improve the proficiency level of the writing. However, whether a writer can make use of the search results efficiently, will be another issue.

In order to show that the similar sentences found by sentence embeddings are useful for real essay writing, we ask 2 post-graduate students who are non-native speaker of English to write 5 sentences each without any aid by machine translation system or dictionary. Then, for each sentence, we show them 15 similar sentences/phrases in the corpus using the fine-tuned SBERT-NLI-ACLARC-base model. Finally, we ask them to correct the original sentence by referring only to the search results. Table 3 show the sentences before and after correction. Words in *italic* form in the corrections are words that do not exist in the original sentences but found in the similar sentences. Out of the 10 sentences, only one sentence could not be corrected, as he/she could not find any suitable words/phrases to use. Based on this result, we believe that the similar sentences found by sentence embeddings are helpful for aiding academic writing.

6 Conclusion

We showed in this paper that sentence embeddings are useful to find similar sentences in order to aid sentence-level revisions in academic writing. However, the problem is that the precision of the search results is still quite low, therefore some of the sentences that are not related to the original sentences will also be shown. These will cause some frustration to the writers because it is wasting their time to read through all the fifteen sentences. Therefore, for a real system, it is important to increase the precision and show less examples. Apart from the abstracts, the search will also cover other sections of articles in the future.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K11446.

References

[1] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Sept. 2015.

[2] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Aug. 2017.

[3] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. Universal Sentence Encoder. *CoRR*, abs/1803.11175, 2018.

[4] A. Conneau and D. Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, 2018.

[5] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Sept. 2017.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[7] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Nov. 2019.

[8] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

Model	Sentence	Sim
Query	We want to show our models work better than previous work.	
ACLARC	With this work we want to draw attention to this fact.	0.8287
	We focus on the tools used for the creation of our corpus and present some results refuting the idea that	0.8259
	the ultimate goal being to use what we learn to improve computer dialogue systems.	0.8225
	Our goal is to describe the problems and to show the solutions proposed.	0.8205
	In this work, we want to verify how this hybrid approach would improve with better classifiers.	0.8144
NLI-ACLARC	with other method, it is proved that our model is more effective by showing the better results.	0.8635
	In case of the same number of parameters are used with other method, it is proved that our model is more effective by showing the better results.	0.8421
	Our experiments show that our method performs better than standard CRF training.	0.8339
	We evaluate our approach on data sets used in prior studies, and demonstrate that our proposed methods perform better than	0.8288
	studies, and demonstrate that our proposed methods perform better than the state-of-the-art systems.	0.8268
NLI-STSb-ACLARC	with other method, it is proved that our model is more effective by showing the better results.	0.7141
	In case of the same number of parameters are used with other method, it is proved that our model is more effective by showing the better results.	0.6740
	produced by our system performs better than the best previous results.	0.6704
	to better ranking performance and speeds up the model training remarkably.	0.6548
	Our proposed system attains significantly better performance than previous approaches for both image caption generalization and In addition, our work	0.6542

Table 2: Examples of similar sentences sorted by descending order of cosine similarity.

Original	After correction
This paper proposes a method which can generate formal sentences under some constraints.	This paper proposes a method which can generate formal sentences under some constraints.
A good sentence structure can make article more readable and persuadable.	The *logical organization* of sentence can *help improve* the *readability* and persuasibility of a *text*.
People usually use some tools to help them writing in the case of lacking the experience of academic writing.	The academic *aid* tools *cater for* people who are *with low skills* of academic writing.
The purpose of academic writing is that using objective words to illustrate the finished scientific work.	The purpose of academic writing is that using objective words to illustrate the *current research*.
Article which translate by other people may not precisely express the meaning from author.	Article which translate by other people may not be *directly adapted to* the meaning from author.
One of advantages of machine translation is that it can promote academic research.	Machine translation has *significant* advantages to promote the *development* of academic research.
Machine translation can help researchers to read papers which are written in different languages.	Machine translation *contributes* to the *study* of papers, which is *vital* to *parsing* the *actual meaning* of papers *presented* in different languages.
There are some drawbacks of neural machine translation system.	Neural machine translation system *suffers* from some *critical problematic issues*.
In order to improve accuracy without huge data, we use analogy method.	*On the basis of theoretical considerations*, *it is suggested that* analogy method *could be used* in order to improve accuracy and *eliminate the need for* huge data.
State-of-the-art research is limited on translating short sentence by analogy.	*A general problem of* state-of-the-art *approaches* is that they are limited to the translation of short sentence.

Table 3: Before and after correction by showing the similar sentences to the post-graduate students. Words in *italic* form are new words added to the original sentences which are found in the similar sentences.