Applying Masked Language Models to Search for Suitable Verbs Used in Academic Writing

Chooi-Ling Goh
The University of Kitakyushu
Kitakyushu
Fukuoka, Japan
goh@kitakyu-u.ac.jp

Abstract

Non-native speakers of English tend to use inappropriate expressions when relating to academic writing. This may cause a non-fluent essay and probably unable to convey the precise meaning of the texts. Therefore, it is important to choose the best suitable vocabulary when writing an academic article, especially for international publication. This paper focuses on word infilling on academic writing, where a word embedding model and some pre-trained masked language models are used to predict and fill in some blank verbs in the text like a cloze task. We evaluate the appropriateness of the predictions of the language models by counting how many verbs can be predicted exactly the same as the original texts and also the fluency of the outputs after replacement. We conduct the test on two journals in natural language processing field, one international and one local, in order to compare the fluency of texts written by native and non-native speakers of English. Our experiments show promising results and motivate us to design a learning and writing system that includes both word embedding and masked language model features.

1 Introduction

Recently, many advanced deep learning language models have been trained and proven to improve many natural language processing (NLP) tasks, such as text generation, summarization, machine translation, question answering and etc. General pre-trained language models such as BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) are actively used in these NLP tasks, and achieve excellent performance. In this research, we want to find out whether these language models can help non-native speakers to find the exact verbs used in academic writing. Non-native speakers of English tend to use inappropriate expressions when relating to academic writing. This will cause their essay to be unnatural and non-fluent, and probably unable to convey the precise meaning of the texts. In the worst case, their papers might not be accepted for publication due to the low level of proficiency. Therefore, it is important to choose the best suitable vocabulary in academic writing, especially for international publication. This paper focuses on word infilling on academic writing, where pre-trained language models are used to fill in some blanks, in our research, the verbs in the text.

BERT, a masked language model inspired by the cloze task, provides us a language model to investigate how well it can fill in the blanks in the text in academic writing. Besides, a word embedding model such as Word2vec, can provide words that have close vectors for a specific word, which may have similar meaning or usage. We use these two types of language models to find suitable verbs used in texts for scientific field. BERT-based language models can predict verbs based on the context, whereas Word2vec can only retrieve closer verbs without looking at the context. Experiments are carried out on some English abstracts taken from the journal papers from NLP field. We compare two journals, where one of them is an international journal which is mostly written by native English speakers and the other one is a local journal mostly written by non-native English speakers. In the experiments, some verbs in the texts are masked out, and predicted by the pre-trained language models. We count how many verbs can be predicted same as the original verbs, and also evaluate if the language model can derive better verbs than the original ones. We also compare the fluency of texts before and after the replacement of the predicted words.

The rest of this paper is organized as follows. Section 2 discusses previous work on masked language model, text infilling and word embedding model. Section 3 describes our approach for the verb prediction. Section 4 describes the experiment settings such as pre-trained language models used and datasets, and finally presents the results. Section 5 discusses our findings and shows some concrete outputs from the experiments. Section 6 presents an academic writing aid tool based on the specific language models. Section 7 concludes our discussion and suggests some future potential work.

2 Previous Work

The original masked language model (MLM) BERT is designed to predict randomly masked tokens like in a cloze task, and whether the next sentence is a succeeding sentence (Devlin et al., 2018). BERT is based on the multilayer bidirectional Transformer (Vaswani et al., 2017), which enables representation of left and right contexts for predicting the masked token. BERT is trained by masking 15% of the words on general domain corpora, i.e. Book Corpus and English Wikipedia texts, with 3.3B tokens.

While BERT can only predict single masked token, further research has expanded the model to predict multiple masked tokens, such as in (Joshi et al., 2020). However, in their research, the length of spans must be decided in advance. Later on, variable-length spans are proposed (Donahue et al., 2020; Zhu et al., 2019; Shen et al., 2020). This research is referred to as text infilling by language modeling. These models are able to fill in the blanks with multiple words, and have no limitation of the length of span. Donahue et al. (2020) propose to infill different granularities of text: words, n-grams, sentences, paragraphs, and documents. However, while these text infilling methods can generate fluent text, they have no control on the content of the text generated. In our approach, we do not need to predict multiple masked tokens at the same time. We should provide as much as possible the surrounding context in order to predict the most suitable missing word.

Word2vec (Mikolov et al., 2013) is a word embedding model which is able to compare the word vectors in order to calculate their similarity using cosine measure. It has been proven that using a Word2vec embedding model trained on specific domain, one can find the most similar word which may be used to replace a word in that domain (Goh and Lepage, 2020).

Based on these previous work, we want to find out if the filling of verbs in the academic articles would be feasible using the MLMs and investigate how well they can predict compared to the original texts. Furthermore, we also compare the results with a word embedding model, Word2vec trained on domain specific scientific articles.

3 Method

Figure 1 shows our approach to find suitable verbs for a particular sentence. In this sentence[1], we need to find two verbs: *[*1] = show* and *[*2] = integrate*. When MLM is used, the context surrounding the verb is provided, then the most probable verbs for this context are returned. In this example, *[*1]* is predicted to be *show, demonstrate, describe, showed, discuss* etc, and *[*2]* is predicted to be *integrate, incorporate, embed, implement, introduce* etc. When Word2vec model is used, no context is provided, but only the verb itself is used to look for verbs that have closer vectors by cosine similarity. Therefore, no matter what the context is, the verb *show* has always closer vectors with *demonstrate, showed, observe, suggest, demonstrating* etc, and *integrate* has always closer vectors with *incorporate, embed, combine, inject, integrating* etc. Finally, the proposed verbs are filled in the sentence in order to form a complete sentence.

4 Experiments

In this section, we describe the pre-trained MLMs and the Word2vec model used for predictions, fol-

[1] A sentence taken from the article V24N04-05 from JNLP.

Input: *We [*1] how to [*2] this method into a standard phrase-based SMT pipeline .*
Answer: *[*1] = show*, *[*2] = integrate*
↓
Predicted by MLM for *[*1]*: *show, demonstrate, describe, showed, discuss, ...*
Predicted by MLM for *[*2]*: *integrate, incorporate, embed, implement, introduce, ...*
Similar by Word2vec for *show*: *demonstrate, showed, observe, suggest, demonstrating, ...*
Similar by Word2vec for *integrate*: *incorporate, embed, combine, inject, integrating, ...*
↓
Possible combination outputs:
We <u>show</u> how to <u>integrate</u> this method into a standard phrase-based SMT pipeline .
We <u>demonstrate</u> how to <u>integrate</u> this method into a standard phrase-based SMT pipeline .
We <u>showed</u> how to <u>incorporate</u> this method into a standard phrase-based SMT pipeline .
etc ...

Figure 1: Method to find suitable verbs using MLM and Word2vec.

lowed by the datasets used for the experiments. Finally, results are shown by calculating the accuracy of the predictions and the fluency of the sentences produced.

4.1 Pre-trained Models

Nowadays, there are plenty of pre-trained language models available. We use some of them for comparisons in our experiments.

For MLM, beside the pioneer BERT, we also compare two variants of BERT-based MLM. DistilBERT is a distilled version of BERT, which is smaller, faster and lighter (Sanh et al., 2019) while retaining 97% of its performance in language understanding. SciBERT is also based on BERT but is trained on scientific texts from Semantic Scholar, with 3.17B tokens (Beltagy et al., 2019). Both BERT and SciBERT has an overlapping of 42% of vocabularies, which shows that general domain and scientific domain have substantial differences on the usage of frequent words. Since MLMs predict words based on the surrounding contexts, they can propose verbs that are conformed to not only semantically, but also functional and morphological similarity, such as word form and tenses. However, we have no control on the selection of semantical meaning of the verbs. We employ the implementation of Hugging Face (Wolf et al., 2019) for using these language models[2].

The Word2vec embedding model is trained on the ACL Anthology Reference Corpus[3] (ACL-ARC). This corpus consists of publications about computational linguistics and natural language processing from selected conferences and journals since 1979 until 2015. It contains 22,878 articles. This embedding model is trained using the gensim implementation[4] of Word2vec. After training, there are 66,453 word vectors in this model, which can be used to propose semantically similar candidates using cosine similarity (Goh and Lepage, 2020). However, this model retrieves words with similar word vectors without any context information. Therefore, sometimes the similar word may be the one with opposite meaning. Moreover, it may not comply with suitable word form and tenses, such as singular form for first person or third person, present tense or past tense, etc. We have chosen this model for comparison because the datasets that we are going to use for experiments are also from NLP field. Therefore, this model should be able to propose words that are most suitable to be used in the texts in NLP field.

Finally, we have four language models to be used in our experiments: BERT, DistilBERT, SciBERT and Word2vec. Below shows precisely the models we used for comparison.

- BERT
  `bert-base-uncased`

[2] `https://huggingface.co/models`

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

[4] `https://radimrehurek.com/gensim/`

- DistilBERT
  `distilbert-base-uncased`

- SciBERT
  `allenai/scibert_scivocab_uncased`

- Word2vec (trained on ACL-ARC)

4.2 Datasets

We conducted our experiments using the abstracts taken from the journals in NLP field. In order to compare the differences between native and non-native speaker writings, we have chosen a local journal and an international journal for our experiments. For the local journal, we collected 631 English abstracts[5] from the *Journal of Natural Language Processing*[6] (hereafter JNLP) published by The Association for Natural Language Processing (ANLP), Japan. Most of the articles in this journal are written in Japanese language, but they have provided an abstract in English. These articles are mostly written by non-native speakers of English[7]. These abstracts contributes to 4,564 sentences, and 108,322 tokens. We mask out all the verbs[8], and try to fill in these verbs with the language models. There are 11,224 masked words, which covers 10.36% of the tokens. For the international journal, we collected 662 abstract[9] from the *Computational Linguistics Journal*[10] (hereafter CLJ) published by The Association for Computational Linguistics (ACL), USA. On the contrary, these articles are mostly written by native English speakers with high level of language proficiency. There are 4,409 sentences, and 116,644 tokens, with 11,995 masked verbs, which is 10.28% of the tokens. Table 1 shows the summary of these datasets.

	JNLP	**CLJ**
# of abstracts	631	662
# of sentences	4,564	4,409
# of tokens	108,322	116,644
# of masked words	11,224	11,995
Masked rate	10.36%	10.28%

Table 1: Statistics on the datasets for JNLP and CLJ.

	First	**Top5**	**Top10**
JNLP			
DistilBERT	22.95%	43.53%	52.88%
BERT	22.92%	45.43%	55.36%
SciBERT	37.79%	65.40%	74.96%
CLJ			
DistilBERT	22.19%	43.94%	53.55%
BERT	24.08%	46.25%	56.02%
SciBERT	39.27%	67.43%	76.11%

Table 2: Accuracy from each MLM model for predictions in top first position, top 5 or top 10 positions.

4.3 Results

This section presents the prediction results. First, for evaluation on accuracy, we count how many words suggested by the MLM matched with the original words. The word may be found in the top first position, top 5 suggestions or top 10 suggestions. Table 2 shows the accuracy rates. Apparently, SciBERT's predictions are far better than the other two models. This proves that domain specific MLM is useful in suggesting correct vocabularies for that domain. When restricted to top 10 predictions, SciBERT achieves an accuracy of about three-quarter of the verbs. Since Word2vec proposes words that have closer vectors with the original word, but never proposes the word itself, therefore there is no accuracy result for it.

Second, we evaluate the performance of the models by fluency using perplexity (PPL). Lower value of perplexity reflects better fluency of texts. The perplexity is calculated based on the GPT-2 language model (Radford et al., 2019)[11]. This model has been successful to improve many NLP tasks with zero-shot task transfer. We believe that this model can provide fair results for evaluating texts

[5] Downloaded from `https://www.anlp.jp/resource/journal_latex/index.html` in LaTeX form.

[6] `https://www.anlp.jp/guide/index.html`

[7] The authors may have asked for proofreading service to correct their English.

[8] Except auxiliary verbs.

[9] Mostly taken from ACL-ARC and partly downloaded from `https://aclanthology.org/venues/cl/` for articles after 2015.

[10] `https://direct.mit.edu/coli`

[11] `https://huggingface.co/transformers/perplexity.html`

	JNLP	CLJ
Tokenized	32.68	34.15
Word2vec	45.13	47.96
DistilBERT	33.15	35.46
BERT	31.65	33.87
SciBERT	30.13	32.04

Table 3: Perplexity for original tokenized text and output from each model.

in any domain. Table 3 shows the perplexity obtained. The tokenized text is the original one. Despite the JNLP's articles are mostly written by non-native English speakers, the fluency is slightly better than CLJ based on perplexity. However, since we do not assess on the proficiency level, it is hard to say that JNLP has higher level than CLJ. The Word2vec model fills in the masked words with the most similar words using cosine similarity. In other words, none of the proposed words are the same as the original words. Therefore, the perplexity is higher, implying lower fluency, as Word2vec does not take contexts into account. Many word proposals by Word2vec do not conform to neither functional nor morphological similarity. For MLMs, only the first suggestion is used for evaluation. From Table 2, we noticed that only 22%–39% of the first suggestions are the same as the original words. However, these do not deteriorate much on the perplexity, or rather better than the original text, especially for SciBERT. This implies that in-domain MLM could offer good suggestions for filling the verbs in academic text.

5 Discussion

In this section, some examples of the predictions are shown and discussed. Figure 2 shows some examples of the prediction outputs[12]. The words in bold face with square brackets are masked words used for prediction. The outputs of each model are in the order as below.

$$\left\{\begin{array}{l}\textbf{[Masked]}\\ \text{Word2vec}\\ \text{DistilBERT}\\ \text{BERT}\\ \text{SciBERT}\end{array}\right\}$$

[12]These sentences were taken from the articles as follows: S1:V24N04-05, S2:V15N03-06, S3:J14-2005, S4:J11-1004.

Some of the words although are not the same as the original words, they make sense to be replaced. For example, it is certainly reasonable to use "*demonstrate*" to replace "*show*" in sentence S1, and "*combining*" to replace "*integrating*" in the sentence S3. Since Word2vec does not take contexts into account, it may introduce some grammatically or functionality erroneous words. For example, in S2, "*correlates*" is replaced by "*correlate*", and in S4, "*managing*" has become "*multimedia*". On the other hand, the problem with MLMs is that although they can predict suitable words based on the contexts, which make the sentence become fluent, sometimes they do not convey the same meaning as the original word. For example, it is fine to replace "*understand*" with "*comprehend*" or "*investigate*" in sentence S4, but certainly "*determine*" is running out from the meaning of the sentence. However, in general, both Word2vec and MLMs are useful in this cloze task which enable the writers to have more choices in selections of proper words.

This experiment results are promising to motivate us in the design of a writing system: we can either use Word2vec to only look for similar words, or masked language model to fill in the blanks. For example, in the input sentence below,

We [*] how to integrate this [method] into a standard phrase-based SMT pipeline .

where *[*]* is used to look for suitable words, and *[method]* is used to look for alternative words that have the similar meaning as "*method*".

In fact, many researchers face problems in composing scientific research articles. For non-native speakers of English, the problem becomes more severe. A machine translation system may help them to translate from their mother tongue language to English but sometimes the translation does not comply with the academic writing style. A language model that trained on specific domain can help them to search for words or expressions that are more suitable to be used when composing an article in that domain, such as in our example, the scientific texts. Therefore, a system that could propose alternative suggestion of vocabularies is very helpful when composing a scientific article, especially for non-native speakers.

6 An Academic Writing Aid Tool

Based on our findings above, we designed a simple writing aid tool (AwTool). In AwTool, users are allowed to provide two types of placeholders in order to search for suitable vocabularies. Here, a set of square bracket is used as a placeholder. In this placeholder, we can either put an asterisk as an unknown word for filling in the blank, where MLM can be used for this purpose. Alternately, if we have a word inside this bracket, it means that we can used Word2vec to look for similar words, and also MLM to look for alternative words that can fill in the blank based on the context. Figure 3 shows an input sentence where *[*1]* is a blank word to be filled, and *[tend]* is a word to be searched for its similar words and also as a place to be filled. The writing process can be a repeated process. Especially for MLM, when the context is changed, the possible suggestions may change too. Therefore, when the user has chosen a certain word and feels comfortable with it, she/he can remove the placeholder and fill in with the word of choice and re-run the search. Figure 4 shows the same example after the placeholders have been removed for *[tend]*, *[coming]* and *[convey]*. As we can see, the word suggestions for *[*1]* and *[correct]* have become slightly different.

7 Conclusion

An academic writing system is indispensable to non-native speakers of English to publish their research work in English with professional standard. The purpose of this research was to investigate the use of masked language models in aiding academic writing. By providing the MLMs the left-right contexts of a sentence, they are able to predict some useful words to fill in the blanks. Using models trained on specific domain, such as SciBERT and Word2vec trained on ACL-ARC, we can control the selections of vocabularies used in scientific articles, and improve the proficiency of academic writing style in that domain. Our experiments were carried out on the abstracts taken from the NLP journal articles written by both native and non-native English speakers. The results were promising and encouraging us to design a writing system that includes both word embedding and language model features. In the future, we would also like to add in AwTool the suggestions to use lexical bundles, which are also indispensable in writing a fluent and native-like essay (Mizumoto et al., 2017; Goh and Lepage, 2019).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP18K11446.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online, July. Association for Computational Linguistics.

Chooi Ling Goh and Yves Lepage. 2019. Extraction of lexical bundles used in natural language processing articles. In *2019 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 223–228.

Chooi Ling Goh and Yves Lepage. 2020. An assessment of substitute words in the context of academic writing proposed by pre-trained and specific word embedding models. In Le-Minh Nguyen, Xuan-Hieu Phan, Kôiti Hasida, and Satoshi Tojo, editors, *Computational Linguistics. PACLING 2019. Communications in Computer and Information Science, vol 1215*, pages 414–427, Singapore. Springer Singapore.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*.

Atsushi Mizumoto, Sawako Hamatani, and Yasuhiro Imao. 2017. Applying the bundle–move connection approach to the development of an online writing support tool for research articles. *Language Learning*, 67(4):885–921, December.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. *arXiv preprint arXiv:2002.03079*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wanrong Zhu, Zhiting Hu, and Eric P. Xing. 2019. Text infilling. *CoRR*, abs/1901.00158.

Examples from Journal of Natural Language Processing	
S1	We {**[show]** demonstrate know know show} how to {**[integrate]** incorporate integrate incorporate integrate} this method into a standard phrase-based SMT pipeline .
S2	However , when we {**[generate]** create write write have} a summary , we {**[use]** employ have have have} much knowledge and experience in our mind . Therefore , it is difficult to {**[compute]** calculate determine understand determine} the importance which {**[correlates]** correlate varies comes is} with human sense .

Examples from Computational Linguistics Journal	
S3	The core of our approach is a new model that {**[combines]** integrates combines combines combines} phrases and dependency syntax , {**[integrating]** incorporating demonstrating with combining} the advantages of phrase-based and syntax-based translation .
S4	We {**[employ]** utilize utilize use use} empirical corpus studies and machine learning experiments to {**[understand]** comprehend determine understand investigate} the mechanisms that people {**[use]** employ engage use engage} in {**[managing]** multimedia solving managing handling} these complex interactions .

Figure 2: Some output examples from each model for JNLP and CLJ. Words in bold face with square brackets are masked words used for predictions. The outputs of each model are in the order of {**[Masked]**, Word2vec, DistilBERT, BERT, SciBERT}.

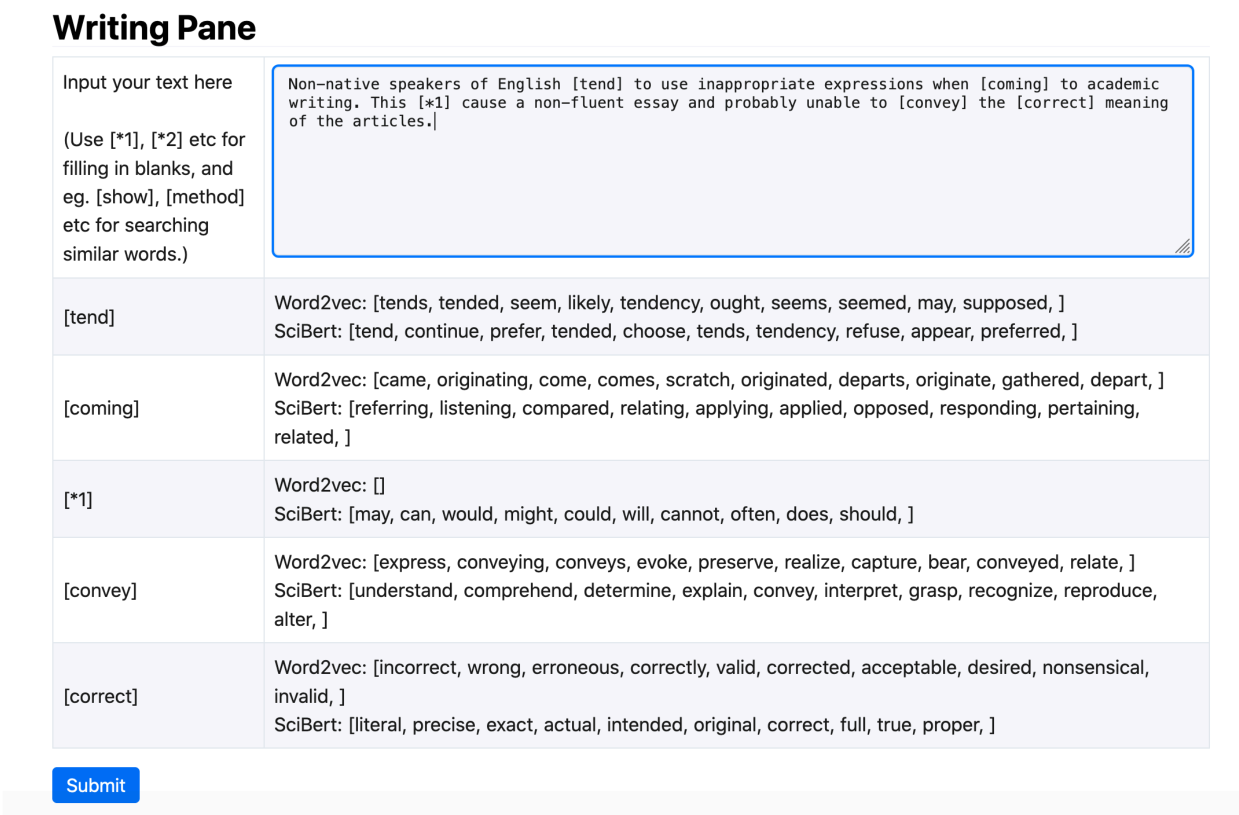

Figure 3: A screenshot for AwTool. *[*1]* is a blank word to be filled, and *[tend]* is a word to be searched for its similar words and also as a place to be filled. Below shows the word suggestions by Word2vec and SciBERT for each placeholder.

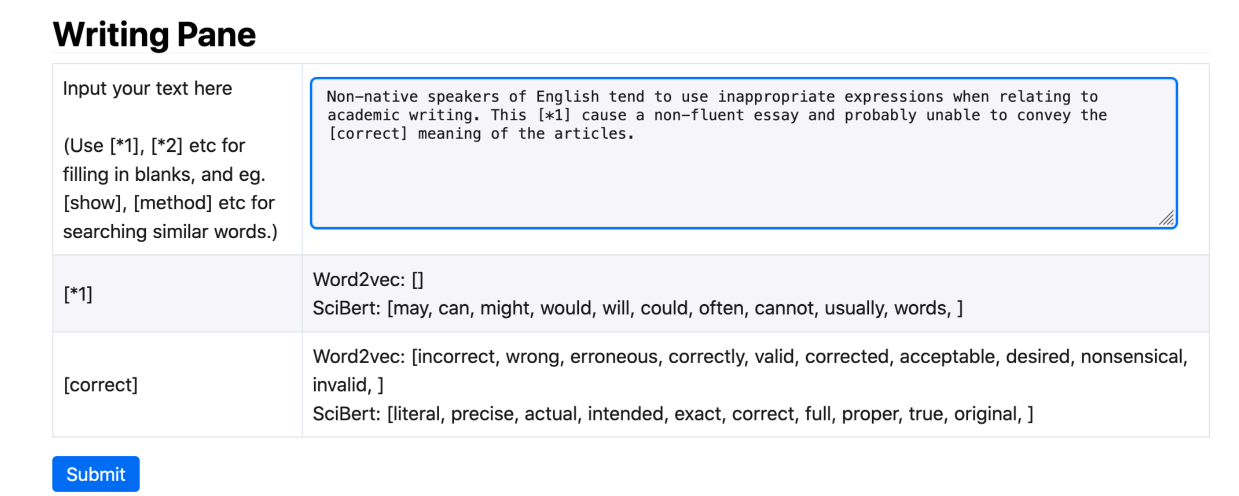

Figure 4: A screenshot for AwTool after a few placeholders have been removed. The word suggestions will be different as the context has changed.