Cloze Test for Verbs in Academic Writing by Masked Language Models

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

1 Introduction

Recently, many advanced 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 [1] and GPT-2 [2] have been used in many NLP tasks, and achieve excellent performance. This paper focuses on text infilling on academic writing, where pre-trained language models are used to fill in some blanks, especially verbs, in the text.

We use BERT, a masked language model inspired by the Cloze task, and investigate how well can it fill in the blanks in the text in academic writing. 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 and 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 words can be predicted same as the original words, and see if the language model can derive better words than the original ones. We also compare the fluency of texts before and after the replacement of the predicted words.

2 Text Infilling

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 [1]. BERT is based on the multilayer bidirectional Transformer [3], which enables representation of left and right contexts for predicting the masked token. BERT is trained by masking 15% of the words. It is trained 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 [4]. However, in this research, the length of spans must be decided in advance. Later on, variable-length spans are proposed [5, 6, 7]. 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. [5] propose to infill different granularities of text: words, n-grams, sentences, paragraphs, and documents. However, these text infilling methods can generate fluent text, but has no control on the meaning of generated text.

In this paper, we want to know if the filling of verbs in the academic articles would be possible using the MLMs and investigate how well can they predict compared to the original texts. Furthermore, we also compare the results with a word embedding model, Word2vec [8], trained on domain specific scientific articles.

3 Experiments

3.1 Pre-trained Models

We compare four language models in our experiments: Word2vec, BERT, DistilBERT and SciBERT.

Word2vec [8] 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 words which can be used to replace the words in academic writing. A Word2vec[1] model trained on the ACL Anthology Reference Corpus[2] (ACL-ARC) can propose semantically similar candidates using cosine similarity [9]. There are 66,453 word vectors in this model. However, this model can only compare the

[1] The Word2vec embedding model is trained using the gensim implementation. https://radimrehurek.com/gensim/

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

word vectors without any context information. Therefore, sometimes the similar word may be the one with opposite meaning.

Beside the original BERT, we also compare two MLMs which are based on BERT. DistilBERT is a distilled version of BERT, which is smaller, faster and lighter [10] 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 [11]. Both BERT and SciBERT has an overlapping of 42% of vocabularies, which shows that general domain and scientific domain have substantial difference on use of frequent words.

Below shows precisely the models we used for comparison.

- BERT
  `bert-base-uncased`
- DistilBERT
  `distilbert-base-uncased`
- SciBERT
  `allenai/scibert_scivocab_uncased`

We employ the implementation of Hugging Face [12] for using these language models.

3.2 Datasets

We collected 631 English abstracts from the Journal of Natural Language Processing (JNLP) published by The Association for Natural Language Processing (ANLP), Japan[3]. These articles are mostly written by non-native speakers of English[4]. These abstracts contributes to 4,564 sentences, and 108,322 tokens. We mask out all the verbs[5], and try to fill in these verbs with the MLMs. There are 11,224 masked words, which covers 10.36% of the tokens. For comparison, we collected 662 abstract from the Computational Linguistics Journal (CLJ) published by The Association for Computational Linguistics (ACL), USA[6]. On the contrary, these articles are mostly written by native English speakers. 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 the datasets.

Table 1 Statistics on the datasets for JNLP and CLJ.

	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%

3.3 Results

This section explains the prediction results. First, we count how many words suggested by the MLM matched with the original words. The word may be found in the first position, top 5 suggestions or top 10 suggestions. Table 2 shows the accuracy rates. Apparently, SciBERT's suggestions 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 suggestions, SciBERT achieves an accuracy of about three-quarter of the verbs.

Table 2 Accuracy from each model 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%

Second, we evaluate the performance of the models using perplexity (PPL). Lower value of perplexity reflects better fluency of texts. The perplexity (PPL) is calculated based on the GPT-2 language model [2][7]. 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 in any domain. Table 3 shows the perplexity obtained. The Word2vec model fills in the masked words with the most similar words using cosine similarity. In other words, none of the 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

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

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

[5] Except auxiliary verbs.

[6] https://www.mitpressjournals.org/loi/coli

[7] https://huggingface.co/transformers/perplexity.html

similarity. Although 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 is higher level than CLJ. 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 filing the verbs in academic text.

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

	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

4 Discussion

Table 4 shows some examples of the prediction outputs. 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.

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

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 MLM 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 MLM are useful in this cloze test.

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 has the similar meaning as "*method*".

5 Conclusion

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. 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. 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.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP18K11446 .

Table 4 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}.

	Examples from Journal of Natural Language Processing
S1	We $\left\{\begin{array}{l}\textbf{[show]}\\\text{demonstrate}\\\text{know}\\\text{know}\\\text{show}\end{array}\right\}$ how to $\left\{\begin{array}{l}\textbf{[integrate]}\\\text{incorporate}\\\text{integrate}\\\text{incorporate}\\\text{integrate}\end{array}\right\}$ this method into a standard phrase-based SMT pipeline .
S2	However , when we $\left\{\begin{array}{l}\textbf{[generate]}\\\text{create}\\\text{write}\\\text{write}\\\text{have}\end{array}\right\}$ a summary , we $\left\{\begin{array}{l}\textbf{[use]}\\\text{employ}\\\text{have}\\\text{have}\\\text{have}\end{array}\right\}$ much knowledge and experience in our mind . Therefore , it is difficult to $\left\{\begin{array}{l}\textbf{[compute]}\\\text{calculate}\\\text{determine}\\\text{understand}\\\text{determine}\end{array}\right\}$ the importance which $\left\{\begin{array}{l}\textbf{[correlates]}\\\text{correlate}\\\text{varies}\\\text{comes}\\\text{is}\end{array}\right\}$ with human sense .
	Examples from Computational Linguistics Journal
S3	The core of our approach is a new model that $\left\{\begin{array}{l}\textbf{[combines]}\\\text{integrates}\\\text{combines}\\\text{combines}\\\text{combines}\end{array}\right\}$ phrases and dependency syntax , $\left\{\begin{array}{l}\textbf{[integrating]}\\\text{incorporating}\\\text{demonstrating}\\\text{with}\\\text{combining}\end{array}\right\}$ the advantages of phrase-based and syntax-based translation .
S4	We $\left\{\begin{array}{l}\textbf{[employ]}\\\text{utilize}\\\text{utilize}\\\text{use}\\\text{use}\end{array}\right\}$ empirical corpus studies and machine learning experiments to $\left\{\begin{array}{l}\textbf{[understand]}\\\text{comprehend}\\\text{determine}\\\text{understand}\\\text{investigate}\end{array}\right\}$ the mechanisms that people $\left\{\begin{array}{l}\textbf{[use]}\\\text{employ}\\\text{engage}\\\text{use}\\\text{engage}\end{array}\right\}$ in $\left\{\begin{array}{l}\textbf{[managing]}\\\text{multimedia}\\\text{solving}\\\text{managing}\\\text{handling}\end{array}\right\}$ these complex interactions .

References

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

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

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 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*, pp. 5998–6008. Curran Associates, Inc., 2017.

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

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

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

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

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

[9] Chooi Ling Goh and Yves Lepage. 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*, pp. 414–427, Singapore, 2020. Springer Singapore.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 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*, 2019.

[11] Iz Beltagy, Kyle Lo, and Arman Cohan. 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)*, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] 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. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, Vol. abs/1910.03771, , 2019.