Word Embeddings in place of Dictionary Lookup in the context of Academic Writing

Chooi Ling GOH
The University of Kitakyushu

Yves LEPAGE
Waseda University

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

1 Introduction

Researchers who are non-native speakers of English always face some problems to compose a scientific article in this language. Most of the time, it is due to the lack of vocabulary or knowledge of alternate ways of expression. They can use machine translation systems to translate from their mother tongues to English when writing English articles. However, sometimes the translation output is not correct, or does not comply with the academic writing style. Some researchers may also use a bilingual dictionary or a thesaurus to search for suitable words.

Word embeddings have been applied to many natural language processing tasks such as information retrieval, sentiment analysis, question answering and document classification. In this paper, we propose to use word embedding models as an alternative to dictionary lexicon or term bank look-up. As opposed to a thesaurus, which usually provides only semantically similar words or expressions, a word embedding model may not only show semantically similar words but also other words that have similar word vectors. Hopefully, a word embedding model trained on a collection of academic articles should comply with the academic writing style and contain terms which are similar in the domain considered. Such dictionary lookup can be of help to non-native speakers of English to search for vocabularies that are suitable for writing an article in style and in lexicon. For example, a lower level proficiency person may know the easy word "but", but word vectors may propose "however" or "although" as alternative expressions.

Following a trend in natural language processing (NLP), we first apply our methods to the NLP research field, i.e., we apply NLP methods on NLP data. This has been called NLP4NLP [2]. We use the ACL Anthology Reference Corpus[1] (ACL-ARC hereafter) as our NLP domain corpus. ACL Anthology is a digital archive of research papers in the premium conferences in NLP and the English language quality of the papers is reputed. The goal of this paper is to run a preliminary experiment. We will use ACL-ARC to build a word embedding model, and compare the word similarity results with some other large pre-trained models.

2 Large Pre-trained Models

There exist three standard models for word embeddings at the moment[2]: Word2vec [3], GloVe [5] and fastText [1].

The above word embedding models allow us to compute the semantic similarity between two words, or to find the most similar words given a target word. The ability to obtain word vectors for out-of-vocabulary words is featured in fastText [1] by capturing the subword information. While Word2vec [3] is limited to a vector space locally, GloVe [5] considers also word co-occurrence globally.

We will use large pre-trained models available from the three methods above to compare with our word embedding model trained on ACL-ARC.

- Word2vec[3]: trained on GoogleNews, GoogleNews-vectors-negative300.bin.gz, 3 billion tokens, 3 million word vectors.
- GloVe[4]: trained on Wikipedia 2014 + Gigaword 5, glove.6B.zip, 6 billion tokens, 400K word vectors.
- fastText[5]: trained on Wikipedia 2017 + UMBC webbase corpus + statmt.org news dataset, wiki-news-300d-1M.vec.zip [4], 16 billion tokens, 1 million word vectors.

GoogleNews model contains compound words (e.g. "ANTARA_News_PRNewswire_AsiaNet" and "eerily_similar"), whereas in other models, no compound word is to be found. Besides, GoogleNews and fastText models have more uncleaned texts, such as erroneous spelling, than other models (e.g. "baed", "similiar", "infomation", see Table 4).

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

[2] We leave aside the more recent ELMo [6] which is based on deep context.

[3] https://code.google.com/archive/p/word2vec/

[4] https://nlp.stanford.edu/projects/glove/

[5] https://fasttext.cc/docs/en/english-vectors.html

3 Specific Word Embedding Model Trained on ACL-ARC

We use ACL-ARC to build our word embedding model. ACL-ARC is a subset of ACL Anthology[6]. The corpus consists of publications about computational linguistics and natural language processing from selected conferences and journals since 1979 until 2015. It consists of 22,878 articles.

We use gensim[7] implementation of Word2Vec to build our model. The parameter settings are as follows.

- Dimensionality of the word vectors (size=300)
- Distance between the current word with the predicted word (window=5)
- Minimum count of word occurrence (min_count=5)

As pre-processing, we extract the texts from the XML output generated by the commercial optical character recognition (OCR) software, Nuance Omnipage. The front pages from the conferences are excluded. There exists some noise or uncleaned text. We did not care about it, and just used them as it is. Most of the noise is coming from conference names, mathematical equations and references. All the text is lower-cased, and words containing numbers, symbols or punctuations are removed. Table 1 shows some statistics on the corpus used for building our word embedding model. From 88 million tokens, we built a model containing 77K word vectors.

# of articles used	21,636
# of tokens	88,006,598
# of distinct word	578,960
# of word vectors	77,311

Table 1: Some statistics on the word embedding model built on ACL-ARC.

4 Experiments

We choose 60 highly frequent words from ACL-ARC, and extract similar words using the four models presented above. From the highest 200 frequent words, we omit some functional words like "the", "of", "and", single character words like "a", "x", "e", and words that are too specific for the NLP field like "semantic", "dependency", "discourse". We further filter words that do not look like having more choices of alternative expressions. Figure 1 shows the 60 base words left finally. For each word, we extract the 10 nearest neighbor words based on cosine similarity from each word embedding model (see Sections 2 and 3).

with, by, which, model, each, data, system, using, used, information, features, results, also, corpus, text, different, some, based, approach, work, given, english, algorithm, evaluation, most, method, performance, new, machine, parsing, however, structure, methods, paper, knowledge, research, processing, possible, while, following, phrase, because, problem, since, experiments, annotation, many, accuracy, form, well, see, very, similar, classification, human, thus, process, best, score, shows

Figure 1: 60 highly frequent words selected from ACL-ARC used for evaluation.

4.1 Evaluation Guidelines

For each base word, if a proposed word could be used to replace the original word, by any form of rephrasing, then it is considered as a possible substitute (1 point), or else it is not (0 point)[8]. We count how many possible substitutes have been proposed by each model.

4.2 Human Judgement

We asked three master second year students (S1, S2, S3) to evaluate the results. These students are non-native speakers of English, but they have some experience in writing at least one international English research article. Their English proficiency levels as of TOEIC (Test of English for International Communication) are shown in Table 2.

	S1	S2	S3
Education level	M2	M2	M2
English education (yrs)	11	11	14
TOEIC level	625	450	430
# of scientific English papers published	1	2	1

Table 2: Information about the evaluators. (M2 stands for master's student 2nd year)

[6] https://aclanthology.coli.uni-saarland.de/

[7] https://radimrehurek.com/gensim/

[8] For GoogleNews and fastText pre-trained models, the same possible substitute may be proposed several times with only differences in case (lower/upper), e.g. "show" and "Show". In this case, we count 0.5 point. The Word2vec model trained on ACL-ARC and the large pre-trained GloVe model lowercase the text before training.

5 Results

Table 3 shows the result of the manual evaluation by the three evaluators. In average, each word has about two possible substitute proposals. Although the ACL-ACR model is much smaller than the large pre-trained models, it gives comparative results for finding similar words.

Model	S1	S2	S3	Avg
Word2vec (ACL-ARC)	2.07	1.53	1.92	1.84
Word2vec (GoogleNews)	1.58 (1.88)	1.72 (2.01)	2.22 (2.52)	1.84 (2.14)
GloVe (wiki+Gigaword)	1.63	2.38	2.55	2.19
fastText (wiki-news)	1.87 (2.14)	1.97 (2.16)	2.23 (2.48)	2.02 (2.26)

Table 3: Evaluation results by three evaluators. Brackets show the results where a same word with different case is counted as 0.5 point.

Table 4 shows some examples of proposed similar words. Double underline shows mutual agreements among the three evaluators. Single underline shows words selected by at least one judge. There is not much mutual agreements among the evaluators.

6 Discussion

Based on the evaluation in Table 3, ACL-ARC provides a slightly lower number of possible substitutes compared to other models. However, it exhibits a larger variety of proposals which conform to the academic writing style. From the authors' point of view, the proposals for "using" and "used" are almost perfect for the ACL-ARC model if the writers are able to rephrase the sentence properly. However, due to the low proficiency level of English of the evaluators in this experiment, they do not have any idea on how to use the words. Since our purpose is to help the writers to compose an article, in the case where the writers do not have the ability to use the proposed substitutes, we need to look for other ways to help them. This experiment was useful to help us in the design of a writing aid tool: we understood that just proposing a list of possible substitutes is not enough, we also need to provide writers with usage samples of the possible substitutes.

As usually observed with human evaluators, there exists some inconsistency, where evaluators have chosen a word in a model, but did not choose it in an other model. For example, from the base word "shows", the word "show" should be mutually agreed upon, but one of the evaluators did not choose it in the ACL-ARC model. The same goes for "similar", where "identical" should be chosen from the GoogleNews model as well.

We also spot another problem: the different forms of spelling, especially between American and British English. For example, evaluators have chosen "utilized" but not "utilised".

As shown in Table 4, GoogleNews and fastText models exhibit a lot of erroneous words. In reality, we should not propose this kind of substitutes to the writers as they may not know that these words are incorrect and they may misuse them. For example, one of the evaluators has mistakenly chosen "infromation" although it is a mis-spelled word.

7 Conclusion

The purpose of this paper was to survey on various word embeddings models, in order to look for alternative expressions for a certain word in place of dictionary lookup in the context of academic writing. As a preliminary experiment, we focused here on proposing words for writing articles in the natural language processing field, by using the ACL-ARC as our search corpus. Compared to large pre-trained models, a specific model proposed more words which conform with the academic writing style.

We conclude that word embeddings are useful for suggesting substitute words for writing academic articles. They can help to transform a low level proficiency writing into academic style writing, especially for non-native speakers of English. In the future, we will explore into suggesting different expressions, not only at word level, but also at phrase, sentence or even paragraph level.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K11446.

References

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics – Volume 5, Issue 1*, pages 135–146, 2017.

[2] G. Francopoulo, J.-J. Mariani, and P. Paroubek. NLP4NLP: Applying NLP to scientific corpora about written and spoken language processing. In *Proceedings of the First Workshop on Mining Scientific Papers: CLBib@ISSI*, pages 5–11, 2015.

base word	Word2vec (ACL-ARC)	Word2vec (GoogleNews)	GloVe (wiki+Gigaword)	fastText (wiki-news)
using	employing, utilizing, applying, uses, via, combining, use, employs, used, utilizes	use, utilizing, Using, used, uses, incorporating, applying, Use, employing, Utilize	used, use, uses, instead, or, types, allows, can, directly, intended	utilizing, employing, utilising, applying, incorporating, use, constructing, combining, creating, substituting
used	utilized, employed, exploited, applied, adopted, leveraged, designed, utilised, reused, useful	utilized, using, use, uses, intended, Used, Using, misused, designed, beused	using, use, uses, or, types, instead, example, such, as, similar	utilized, employed, utilised, uses, designed, applied, relied, referred, devised, allowed
based	relies, relying, basis, rely, relied, depends, depending, focuses, depend, focusing	Based, headquartered, baed, headquarted, basing, basedin, bsed, Basing, basd, ANTARA_News_PR-Newswire_AsiaNet	company, group, new, its, which, firm, business, also, part, research	basing, Based, predicated, relying, relies, derived, centered, rely, focusing, premised
however	although, moreover, furthermore, indeed, nevertheless, unfortunately, but, though, because, nonetheless	though, although, nevertheless, nonetheless, that, meanwhile, also, but, only, not	although, though, as, both, this, but, be, also, latter, .	although, though, nevertheless, but, therefore, nonetheless, indeed, unfortunately, consequently, yet
similar	analogous, identical, dissimilar, close, competitive, comparable, related, divergent, promising, complementary	similiar, strikingly_similar, Similar, identical, dissimilar, eerily_similar, virtually_identical, different, simliar, same	example, instance, such, same, this, unusual, which, although, particular, common	similiar, comparable, identical, analogous, dissimilar, same, different, related, akin, simlar
shows	illustrates, reveals, demonstrates, showing, depicts, indicates, displays, suggests, show, summarizes	show, shown, showed, showing, indicates, reveals, demonstrates, suggests, illustrates, Shows	show, shown, picture, feature, showing, seen, appearing, featured, appeared, this	show, showing, indicates, shown, demonstrates, showed, illustrates, reveals, displays, Shows
information	clues, meta-information, knowledge, meta-data, shifters, metadata, worry, cues, generalisations, exclusivity	info, infomation, infor_mation, informaiton, informa_tion, informationon, informationabout, Information, informaion, details	source, data, sources, provided, documents, search, web, provide, knowledge, intelligence	infomation, informaton, info, informtion, informatin, infromation, information-, informaion, inforamtion, infomration

Table 4: Examples of proposed similar words sorted by descending order of cosine similarity. Double underline shows mutual agreements among the three evaluators. Single underline shows words selected by at least one judge.

[3] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[4] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of LREC*, 2018.

[5] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, June 2018.