An Assessment of Substitute Words in the Context of Academic Writing Proposed by Pre-trained and Specific Word Embedding Models

Chooi Ling Goh[1] and Yves Lepage[2]

[1] The University of Kitakyushu, Kitakyushu, Fukuoka, Japan
goh@kitakyu-u.ac.jp
[2] Waseda University, Kitakyushu, Fukuoka, Japan
yves.lepage@waseda.jp

Abstract. Researchers who are non-native speakers of English always face some problems when composing scientific articles in this language. Most of the time, it is due to lack of vocabulary or knowledge of alternate ways of expression. In this paper, we suggest to use word embeddings to look for substitute words used for academic writing in a specific domain. Word embeddings may not only contain semantically similar words but also other words with similar word vectors, that could be better expressions. A word embedding model trained on a collection of academic articles in a specific domain might suggest similar expressions that comply to that writing style and are suited to that domain. Our experiment results show that a word embedding model trained on the NLP domain is able to propose possible substitutes that could be used to replace the target words in a certain context.

Keywords: word embedding · word similarity · dictionary lookup · synonym · academic writing.

1 Introduction

Many researchers face problems in composing scientific articles. For non-native speakers of English, the problem becomes more severe. They can use machine translation systems to translate from their mother tongues to English, but most of the time, the translation output quality is not satisfactory, or does not comply with the academic writing style. A bilingual dictionary or a thesaurus may be used to search for suitable expressions, when only simple words come across the mind. However, not all expressions suggested comply to the academic writing style. Moreover, it is difficult to know whether the alternate ways of expression are in style and in lexicon.

In this paper, we suggest to use word embeddings to search for substitute words in the context of academic writing. Word embeddings have been applied to many natural language processing tasks such as information retrieval, sentiment analysis, question answering and document classification. As opposed to a thesaurus, which usually provides only semantically similar words or expressions, word embeddings may not only show semantically similar words but also other words with similar word vectors, that could be even better. Furthermore, if we train an embedding model on only a collection of academic articles, then, possibly, similar expressions which comply to that writing style and which are suited to the domain might be suggested. This can be very helpful for non-native speakers of English to guide them to write articles in style and in lexicon. As an example, a less proficient person may know the easy word "*but*", but word vectors may propose "*however*" or "*although*" as alternative words. Similarly, a target word like "*show*" may be replaced with more sophisticated words like "*reveal*" or "*depict*".

One may suggest to use the English lexical database WordNet [11] for finding the synonyms while writing an article. Certainly, it is a good idea to consult a well defined semantic network for this purpose, but we are wondering whether word embeddings would propose some different expressions. It has been reported that the number of human-judged synonyms extracted from word embeddings is about twice the number given by WordNet in a survey [6] for extracting synonyms to be used in the machine translation evaluation metric METEOR [3]. Therefore, there is a high possibility that word embeddings could give better suggestion of word proposal in academic writing.

The goal of this paper is to compare the word similarity results for a word embedding model built on a specific domain with some other general large pre-trained models. We use the ACL Anthology Reference Corpus[3] (ACL-ARC hereafter) in the natural language processing (NLP) domain as our specific domain. 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. We would like to know whether a specific model trained on specific domain could give equivalent or better word similarity than large pre-trained models.

2 Specific Word Embedding Model Trained on ACL-ARC

As mentioned above, we used ACL-ARC to build a specific word embedding model. ACL-ARC is a subset of ACL Anthology[4]. The corpus consists of the publications about computational linguistics and natural language processing from selected conferences and journals since 1979 until 2015. It consists of 22,878 articles.

We used the gensim[5] implementation of Word2Vec to build our model. We trained our model with the continuous bag-of-words (CBoW) model as this model has been shown to be the optimal choice for building better models for English [9, 6]. The parameter settings are as follows.

[3] https://acl-arc.comp.nus.edu.sg/
[4] https://aclanthology.coli.uni-saarland.de/
[5] https://radimrehurek.com/gensim/

– 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 extracted 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. We also excluded the section references in the papers. However, there still exists some noise or uncleaned texts. Most of the noise is coming from conference names, mathematical equations, figures and tables. All the texts are lowercased, 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 66k word vectors.

Table 1. Statistics on the word embedding model built on ACL-ARC

# of articles used	21,636
# of tokens	88,006,598
# of distinct word	578,960
# of word vectors (include references)	77,311
# of word vectors (exclude references)	66,453

3 Large Pre-trained Models

There exist three standard models for word embeddings at the moment [6] : Word2vec [9], GloVe [13] and fastText [2]. These models allow us to compute the semantic similarity between two words, so as 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 [2] by capturing subword information. While Word2vec [9] is limited to a vector space locally, GloVe [13] also considers word co-occurrence globally.

We will use large pre-trained models available from the three methods above to compare with our specific word embedding model trained on ACL-ARC. All the models, including the specific model, are trained with 300 dimensions which has been proven to deliver optimal performance [9, 7].

– Word2vec[7]: trained on GoogleNews, GoogleNews-vectors-negative300.bin.gz, 3 billion tokens, 3 million word vectors.

[6] We leave aside the more recent ELMo [14] that is based on deep context, and BERT [4] that uses masked language model.
[7] https://code.google.com/archive/p/word2vec/

– GloVe[8]: trained on Wikipedia 2014 + Gigaword 5, glove.6B.zip (300d), 6 billion tokens, 400 thousand word vectors.
– fastText[9]: trained on Wikipedia 2017 + UMBC webbase corpus + statmt.org news dataset, wiki-news-300d-1M.vec.zip [10], 16 billion tokens, 1 million word vectors.

The GoogleNews model[10] contains compound words (e.g. "*ANTARA_News_PRNewswire_AsiaNet*" and "*eerily_similar*"), whereas in other models, no compound word is found. Besides, GoogleNews and fastText models have more uncleaned items, like erroneous spelling, than other models (e.g. "*baed*", "*similiar*", "*infomation*"). Furthermore, these models are case-sensitive, e.g. "*show*" and "*Show*", both exist in the models. A preliminary experiment has shown that these noise words appeared in higher ranking of word similarity, it is therefore better to remove them from the models.

Hence, in order to have a fair comparison, we further filter the large pre-trained models, so that they only contain words that are found in the ACL-ARC word vectors. A large number of word vectors are removed by this filtering. The number of word vectors left is shown in Table 2. After filtering, all the words in all models are in lowercase, and no compound words or erroneous words left.

Table 2. Statistics on the word embedding models after filtering

	ACL-ARC	GoogleNews	GloVe	fastText
Training size	88M	3B	6B	16B
Before filtering	66,453	3M	400k	1M
After filtering	66,453	24,912	33,231	32,569

4 Experiments

We chose 12 highly frequent words from ACL-ARC which look like producing more choice of substitute words to be the target words. We extracted similar words using the four models presented above. The 12 target words used for evaluation are shown below. Table 3 shows the synonyms taken from the WordNet lemmas.

with, by, each, using, results, some, however, methods, see, very, thus, shows

[8] https://nlp.stanford.edu/projects/glove/
[9] https://fasttext.cc/docs/en/english-vectors.html
[10] For simplicity, the four models are referred as ACL-ARC, GoogleNews, GloVe and fastText hereafter.

Table 3. 12 frequent words selected from ACL-ARC used for evaluation. Right column shows the synonyms taken from the WordNet lemmas.

target word	Synonyms from WordNet
with	[NOT FOUND]
by	aside, away, past
each	apiece, for_each_one, from_each_one, to_each_one
using	apply, employ, expend, exploitation, habituate, practice, use, utilise, utilize, victimisation, victimization
results	answer, consequence, effect, ensue, event, final_result, issue, lead, leave, outcome, resolution, result, resultant, resultant_role, solution, solvent, termination, upshot
some	about, approximately, around, close_to, just_about, more_or_less, or_so, roughly
however	all_the_same, even_so, nevertheless, nonetheless, notwithstanding, still, withal, yet
methods	method, method_acting
see	ascertain, assure, attend, catch, check, come_across, consider, construe, control, date, determine, discover, encounter, ensure, envision, escort, examine, experience, fancy, figure, find, find_out, get_a_line, get_wind, get_word, go_out, go_steady, go_through, hear, image, insure, interpret, learn, look, meet, pick_up, picture, project, realise, realize, reckon, regard, run_across, run_into, see_to_it, take_care, take_in, understand, view, visit, visualise, visualize, watch, witness
very	identical, rattling, real, really, selfsame
thus	frankincense, gum_olibanum, hence, olibanum, so, thence, therefore, thusly
shows	appearance, bear_witness, demo, demonstrate, depict, designate, display, establish, evidence, evince, exhibit, express, indicate, picture, point, present, prove, read, record, register, render, shew, show, show_up, testify, usher

For each word, we extracted the 10 nearest neighbor words based on cosine similarity from each word embedding model (see Sections 2 and 3). According to [1], word embedding performance is affected by various factors such as corpus size, length of individual texts or existence of specific content. [1] suggest to measure the stability of the performance by testing on multiple bootstrap samples. On the other hand, word vectors are concatenated in [8] in order to combine two vectors so as to obtain better performance in some extrinsic tasks. In our approach, we did not combine word embeddings, but only combined similar words extracted from the four models (i.e. JointModel) using the heuristic below.

1. For each proposed word, sum up the cosine similarity values from all models.
2. Order the list by number of occurrences and total cosine similarity in descending order.
3. Consider only the 10 highest ranking words in that order.

In order to evaluate the performance of each model, we conducted two experiments:

– evaluation on an extrinsic task using the machine translation output, and
– an intrinsic evaluation by human judgement.

4.1 Evaluation using Machine Translation

For each target word, we collected 10 sentence pairs from English-French translation pairs in Linguee[11]. We tried not to collect sentences that are too long or too short. Too long sentences may not produce satisfactory machine translation results and too short sentences may not provide enough context. In average, the length is about 17 words per sentence. We chose the English-French language pair as the translation pair because it exhibits better machine translation results currently. For 12 words, we collected 120 sentence pairs in total. We then used the deepL[12] translator to translate from English to French. Since deepL is trained on top of Linguee, we also translated the sentences using Google Translate[13] for comparison. We evaluate the translation results using the BLEU metric [12]. Higher BLEU scores are obtained by translations that are closer to the target reference translations, which imply better translations. The translation performance for deepL and Google Translate are shown at the top part of Table 4. It shows that deepL delivers better translation results than Google Translate. In the following experiments, we consequently used only deepL for translation.

For each target word in each model, we replaced the corresponding 10 sentences with the 10 candidate words, i.e., each candidate word is replaced in 10 sentences, therefore, we generated 100 sentences per target word. In total for 12 target words, we have 1,200 sentences per model. We translated these sentences using deepL and calculated the BLEU scores. The bottom part of Table 4 shows the translation results. As for individual models, GoogleNews proposes the best candidates and fastText is the worst. By combining all the models into a JointModel, better candidates are proposed.

[11] https://www.linguee.com/
[12] https://www.deepl.com/translator
[13] https://translate.google.com/

Table 4. Machine translation results using BLEU scores

	BLEU score
deepL	45.09
Google Translate	39.48
Word embedding model	
ACL-ARC	39.05
GoogleNews	39.61
GloVe	38.69
fastText	38.65
JointModel	40.02

4.2 Human Judgement

We also evaluate candidate word proposals obtained by human judgement. For each target word, if a proposed candidate word can 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). The substitute word must also conform to morphological features, i.e., it should exhibit the correct word form according to the tense, number, etc and it should be semantically similar. Basically, different word forms of the same lemma are not necessarily substitutable, e.g. "*using*" and "*used*". We enquire how many possible substitutes are proposed by each model.

Table 5. Results by human judgement

Model	Total	Avg/person	Avg/word
Before inconsistency correction			
ACL-ARC	227	32.43	2.70
GoogleNews	223	31.86	2.65
GloVe	125	17.86	1.49
fastText	244	34.86	2.90
JointModel	251	35.86	2.99
After inconsistency correction			
ACL-ARC	237	33.86	2.82
GoogleNews	240	34.29	2.86
GloVe	131	18.71	1.56
fastText	259	37.00	3.08
JointModel	**263**	**37.57**	**3.13**

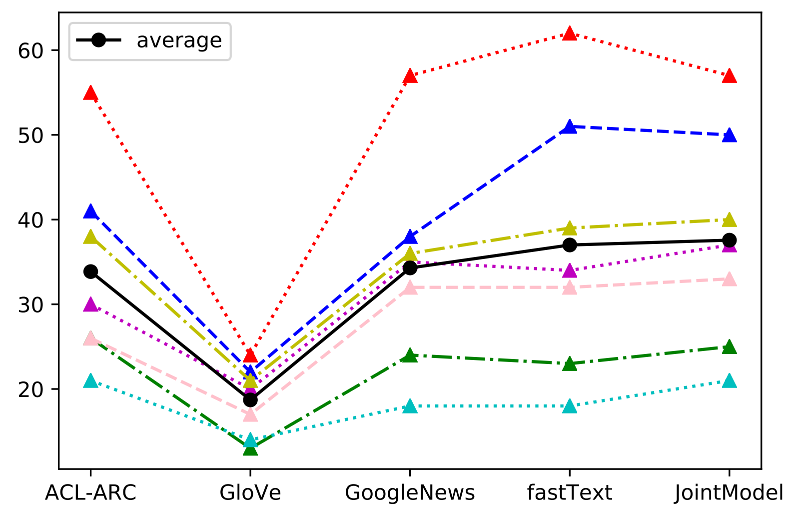

Fig. 1. Comparison of results of evaluators. Each dotted line shows a different evaluator. The black solid line shows the average. The lines between the points are just for better identification of evaluators.

We asked seven postgraduate students who are conducting research in the NLP domain to evaluate the word candidates. These students are non-native speakers of English: two are French, one is Thai, and four others are Chinese. These students have experience on writing at least one conference paper or their thesis in English.

As usually observed with human evaluators, there exists some inconsistency, where evaluators have chosen a word in a model, but have not chosen the same word in another model. For example, from the target word "*very*", one of the evaluators had chosen the word "*remarkably*" in the JointModel, but did not choose it in the specific ACL-ARC model and the fastText model. This phenomenon happened for almost all the evaluators. We corrected this kind of mistakes made by the evaluators by double checking the selections, and adding the corresponding selections to the possible substitute word lists for all models.

Table 5 shows the results for human judgement. The top part shows the original results made by the evaluators without correction and the bottom part is the version corrected for inconsistency. In average, each word has about two to three possible substitute proposals. Although the specific ACL-ACR model is much smaller than the large pre-trained models, it gives comparable results for finding similar words. Figure 1 shows the annotations for each evaluator[14]. In

[14] The line graph is used just to better identify the evaluators.

general, the JointModel gives the best word proposals and the GloVe model has the least suitable candidates. All the evaluators show similar tendency to all the models. The human judgement has a fixed-marginal Fleiss's kappa [5] value of 0.49, which is considered as moderate agreement among the evaluators.

4.3 Discussion

Table 6 and Table 7 show some examples of proposed similar words. Double underline shows words selected by at least four evaluators and single underline by at least one evaluator. Words with gray background shows mutual agreement by all evaluators. There is not much mutual agreement among the evaluators. This could be caused by the different English proficiency levels among them. However, in general, many of the proposed words look good, and conform to the lexicon used in scientific articles. Moreover, word embeddings also propose words that are not in their WordNet synsets (referred to Table 3). For example, the target word "*very*" has the lemmas "*identical, rattling, real, really, selfsame*" in its WordNet synset, but our models suggested that "*quite, fairly, extremely, pretty, remarkably, highly*" etc. are possible substitutions. We conclude that it is possible to use word embedding models to find appropriate substituitions in the context of academic writing.

Based on the evaluation in Table 5, the specific ACL-ARC model provides a slightly lower number of possible substitutes compared to other models (except the GloVe model). But, it exhibits a larger variety of proposals which conform to the academic writing style. The human evaluation is very much dependent on the English proficiency level of the evaluators. Some of the proposed words seem to be too difficult for them to judge, especially in the absence of any context. 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 shall provide writers with usage samples of the possible substitutes.

Table 8 and Table 9 show two examples of translations after substituting the proposed words to the target word. Table 8 shows the target word "*using*" with the substitute words proposed by the specific ACL-ARC model; Table 9 shows the target word "*however*" with the substitute words proposed by the Joint-Model model. The substitute sources with double underline are words selected by at least four evaluators and single underline by at least one evaluator. The translation outputs show that even we use the words with different form for the same lemma, as "*uses*" and "*used*", to replace the target word "*using*", it will be translated into the same French word "*utiliser*". However, the right contexts are different. For the target word "*however*", some substituted words are omitted in the translations, as in the case of "*although*" and "*though*". For this kind of conjunction word, it is difficult to replace the target word directly, but one needs to rewrite the whole sentence in order to keep it semantically similar. Hence, it is also difficult to judge by translation output.

Table 6. Examples of proposed similar words sorted by descending order of cosine similarity. Double underline (word) shows words selected by at least four evaluators and single underline (word) by at least one evaluator. Words with gray background shows mutual agreement by all evaluators.

target word	ACL-ARC	GoogleNews	GloVe	fastText	JointModel
using	employing, utilizing, applying, exploiting, uses, via, used, relying, employs, utilizes	use, utilizing, used, uses, incorporating, applying, employing, utilize, utilized, utilizes	used, use, uses, method, methods, instead, techniques, technique, example, can	utilizing, employing, utilising, applying, incorporating, use, constructing, combining, creating, substituting	use, utilizing, used, uses, employing, applying, incorporating, utilizes, utilising, constructing
however	moreover, furthermore, although, but, nevertheless, unfortunately, therefore, because, though, nonetheless	though, although, nevertheless, nonetheless, that, meanwhile, also, but, only, not	although, though, but, that, not, nevertheless, because, only, fact, also	although, though, nevertheless, but, therefore, nonetheless, indeed, unfortunately, consequently, yet	although, though, but, nevertheless, nonetheless, that, not, only, also, therefore
methods	techniques, approaches, algorithms, methodologies, strategies, method, metrics, mechanisms, technique, schemes	techniques, method, methodologies, technique, tactics, methodology, strategies, approaches, mechanisms, procedures	techniques, method, technique, using, methodologies, methodology, use, tactics, procedures, tools	techniques, method, methodologies, strategies, approaches, methodology, procedures, technique, ways, tactics	techniques, method, methodologies, technique, approaches, methodology, strategies, tactics, procedures, mechanisms

continue on Table 7

Table 7. Continue from Table 6.

target word	ACL-ARC	GoogleNews	GloVe	fastText	JointModel
some	many, numerous, certain, several, borderline, lots, conflicting, themselves, exceptional, all	many, few, lot, several, plenty, those, lots, these, all, little	many, few, those, other, have, more, several, others, these, even	many, several, these, certain, all, most, various, other, those, few	many, several, few, these, those, all, other, certain, lots, have
very	quite, fairly, extremely, relatively, remarkably, too, comparatively, overly, reasonably, sufficiently	extremely, quite, pretty, fairly, particularly, really, so, obviously, too, especially	extremely, quite, so, pretty, too, really, well, always, especially, but	extremely, pretty, quite, remarkably, highly, relatively, fairly, too, most, somewhat	extremely, quite, too, pretty, fairly, so, really, remarkably, relatively, especially
shows	demonstrates, illustrates, showing, reveals, depicts, summarizes, displays, indicates, suggests, show	show, shown, showed, showing, indicates, reveals, demonstrates, suggests, illustrates, confirms	show, shown, showing, showed, seen, tv, television, featured, appears, recent	show, showing, indicates, shown, demonstrates, showed, illustrates, reveals, displays, depicts	show, showing, shown, demonstrates, showed, illustrates, reveals, indicates, depicts, displays

Table 8. An example for translation of "*using*" with the substitute words proposed by the ACL-ARC model.

Source	Reference
Sometimes you may want to create a window running a program directly, without **using** a shell first.	Parfois vous voulez créer une fenêtre executant directement un programme, sans **passer** par l'invite de commande.
Translation by deepL	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** un shell au préalable.
Translation by Google Translate	Parfois, vous souhaiterez peut-tre crer une fen-tre excutant un programme directement, sans **utiliser** d'abord un shell.
Substitution source	Translation by deepL
Sometimes you may want to create a window running a program directly, without **employing** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** un shell au préalable.
Sometimes you may want to create a window running a program directly, without **utilizing** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** d'abord un shell.
Sometimes you may want to create a window running a program directly, without **applying** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **appliquer** d'interpréteur de commandes au préalable.
Sometimes you may want to create a window running a program directly, without **exploiting** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **exploiter** d'abord un shell.
Sometimes you may want to create a window running a program directly, without **uses** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** d'interpréteur de commandes au préalable.
Sometimes you may want to create a window running a program directly, without **via** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **passer** par un shell d'abord.
Sometimes you may want to create a window running a program directly, without **used** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** d'interpréteur de commandes au préalable.
Sometimes you may want to create a window running a program directly, without **relying** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **avoir besoin de faire appel à** un shell au préalable.
Sometimes you may want to create a window running a program directly, without **employs** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** un shell au préalable.
Sometimes you may want to create a window running a program directly, without **utilizes** a shell first.	Parfois, vous pouvez vouloir créer une fenêtre exécutant un programme directement, sans **utiliser** d'abord un shell.

Table 9. An example for translation of "*however*" with the substitute words proposed by the JointModel model.

Source	Reference
It is not **however** the sole tool, nor in the end is it the most important one.	Ce n'est **cependant** pas le seul, ni le plus important au bout du compte.
Translation by deepL	**Mais** ce n'est pas le seul outil, ni en fin de compte le plus important.
Translation by Google Translate	Ce n'est **cependant** pas le seul outil, ni à la fin le plus important.
Substitution source	Translation by deepL
It is not **although** the sole tool, nor in the end is it the most important one.	Ce n'est pas le seul outil, ni en fin de compte le plus important.
It is not **though** the sole tool, nor in the end is it the most important one.	Ce n'est pas le seul outil, ni en fin de compte le plus important.
It is not **but** the sole tool, nor in the end is it the most important one.	Ce n'est pas **seulement** le seul outil, ni en fin de compte le plus important.
It is not **nevertheless** the sole tool, nor in the end is it the most important one.	Ce n'est **cependant** pas le seul outil, ni en fin de compte le plus important.
It is not **nonetheless** the sole tool, nor in the end is it the most important one.	**Il n'en reste pas moins que** ce n'est pas le seul outil, ni le plus important en fin de compte.
It is not **that** the sole tool, nor in the end is it the most important one.	Ce n'est pas **que** le seul outil, ni en fin de compte le plus important.
It is not **not** the sole tool, nor in the end is it the most important one.	Ce n'est pas le seul outil, ni en fin de compte le plus important.
It is not **only** the sole tool, nor in the end is it the most important one.	Ce n'est pas **seulement** le seul outil, ni en fin de compte le plus important.
It is not **also** the sole tool, nor in the end is it the most important one.	Ce n'est pas **non plus** le seul outil, ni en fin de compte le plus important.
It is not **therefore** the sole tool, nor in the end is it the most important one.	Ce n'est **donc** pas le seul outil, ni en fin de compte le plus important.

5 Conclusion

The purpose of this paper was to inspect the use of various word embedding models, in order to look for substitute words for a certain target word in the context of academic writing in place of dictionary lookup. Our experiment focused on proposing words for articles in the natural language processing domain, using the ACL-ARC as a corpus for training a specific word embedding model. We limited the word vectors of the large pre-trained models to the vocabulary found in the specific ACL-ARC model for a fair comparison. Compared to large pre-trained models, the specific model proposed more words conform to the academic writing style. By combining the proposed words from all the models into a JointModel model, we further improved the word proposals.

We conclude that word embeddings are useful for suggesting substitute words for writing academic articles. They can help a non-native speaker of English to transform a low level proficiency text into proper academic style writing.

In the future, we will explore into suggesting different expressions, not only at the word level, but also at the phrase, sentence or even paragraph level. We also want to enforce functional similarity using substitute vectors [15, 7], so that proposed words are conform not only by semantic similarity, but also by morphological similarity. Finally, some words may lead to more lexical choice than other words, which points at varying the number of proposed substitutes. We may apply relative cosine similarity as suggested by [6], and decide on a threshold so as to suggest a relevant variable number of word proposals.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K11446 .

References

1. Antoniak, M., Mimno, D.: Evaluating the Stability of Embedding-based Word Similarities. Transactions of the Association for Computational Linguistics **6**, 107–119 (2018)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics **5**(1), 135–146 (2017)
3. Denkowski, M., Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation. pp. 376–380 (2014)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (Jun 2019)
5. Fleiss, J.L.: Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin **76**(5), 378–382 (Nov 1971)
6. Leeuwenberg, A., Vela, M., Dehdari, J., Genabith, J.: A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. The Prague Bulletin of Mathematical Linguistics **105**, 111–142 (Apr 2016)

7. Melamud, O., Dagan, I., Goldberger, J.: Modeling Word Meaning in Context with Substitute Vectors. In: Proceedings of the NAACL. pp. 472–482 (2015)
8. Melamud, O., McClosky, D., Patwardhan, S., Bansal, M.: The Role of Context Types and Dimensionality in Learning Word Embeddings. In: Proceedings of the NAACL-HLT. pp. 1030–1040 (Jun 2016)
9. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR (2013)
10. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in Pre-Training Distributed Word Representations. In: Proceedings of LREC (2018)
11. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM **38**(11), 39–41 (1995)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of ACL. pp. 311–318 (Jul 2002)
13. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: Proceedings of EMNLP. pp. 1532–1543 (2014)
14. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep Contextualized Word Representations. In: Proceedings of NAACL-HLT. pp. 2227–2237 (Jun 2018)
15. Yatbaz, M.A., Sert, E., Yuret, D.: Learning Syntactic Categories Using Paradigmatic Representations of Word Context. In: Proceedings of the EMNLP-CoNLL. pp. 940–951 (Jul 2012)