Automatic Assistance for Academic Word Usage

Dariush Saberi, John Lee, Jonathan Webster
Department of Linguistics and Translation,
The Halliday Centre for Intelligent Applications of Language Studies,
City University of Hong Kong
dsaberi2-c@my.cityu.edu.hk, jsylee@cityu.edu.hk,
ctjjw@cityu.edu.hk

Abstract

This paper describes a writing assistance system that helps students improve their academic writing by incorporating vocabulary that is more typical in the academic setting. Given an input text, the system suggests word substitutions according to an academic word list, and ranks them with a masked language model. Experimental results show that lexical formality analysis can improve the quality of the suggestions, in comparison to a baseline that relies on the masked language model only. Further, a user study demonstrate that students were able to use the system to improve text quality.

1 Introduction

While most research on automatic writing assistance has focused on grammatical error correction (Ng et al., 2014), there has been increasing attention on analyzing academic writing, with respect to the writing style expected in the genre (Bailey, 2011). Examples of recent efforts include identification of argumentation structure (Zhang et al., 2017), automatic assistance for nominalization and sentence restructuring (Lee et al., 2019), and lexical substitution of academic vocabulary (Yimam et al., 2020). This paper focuses on the latter, an example of which is shown in Table 1.

Lexical substitution (LS) is the subfield of natural language processing that aims to replace *target words* in a text without changing its meaning (McCarthy and Navigli, 2009). Our task — academic LS — may be viewed as a special form of LS, with the additional requirement that the substitution be typical for academic discourse. We describe and evaluate a writing assistance system for academic word usage that incorporates lexical formality analysis. Experimental results show that the incorporation of formality analysis helps raise the performance in identifying suitable substitutions.

2 Previous work

The system reported in Yimam et al. (2020), which performs academic lexical substitution (LS) using the typical LS pipeline (Paetzold and Specia, 2016), is closely related to ours. In the first step, the system identifies the non-academic target words with a binary classifier. The second step, Substitution Generation, proposes candidate substitutions for each target word. Finally, the system ranks these candidates to determine the optimal substitution. The candidate substitutions were harvested from PPDB (Pavlick et al., 2015) and WordNet (Fellbaum, 1998), and ranked with the TF-Ranking deep learning model (Pasumarthi et al., 2019).

Automatically mined pairs of informal and formal words (Brooke et al., 2010) have been shown to be useful for natural language generation (Sheikha and Inkpen, 2011). Lexical formality analysis may also be beneficial for academic LS since academic vocabulary tends to be more formal. Although informal-formal text pairs are now publicly available in large volume (Pavlick and Tetreault, 2016), they have not yet been exploited in academic LS.

Proceedings of the 28th International Conference on Computational Linguistics, pages 2163–2168
Barcelona, Spain (Online), December 8-13, 2020

Input:	She	*left*	university training in hydroponics to work for ...
Output:	She	***abandoned***	university training in hyproponics to work for ...

Table 1: Example input and output sentence in academic lexical substitution, with the target word *left* and its substitution *abandon*.

3 Data

There is no existing dataset for academic lexical substitution (LS). We constructed our evaluation data by compiling an academic word list (Section 3.1) and an informal word list (Section 3.2), and transforming an LS dataset with these lists (Section 3.3).

3.1 Academic word list

Our academic word list combines three resources: the *New Academic Word List* (Coxhead, 2016a; Coxhead, 2016b), which consists of 963 headwords selected from an academic corpus of 288 million words; the *Academic Vocabulary List*, with 3,000 lemmas selected from the academic subcorpus of the Corpus of Contemporary American English (Gardner and Davies, 2013); and the 8,625 words labelled as "Technical / Domain Specific" based on their frequency in the same corpus (https://www.academicvocabulary.info). After removing duplicates, the list contains a total of 10,343 words. We will henceforth call a word an *academic word* if it belongs to this list, and a *non-academic word* otherwise.

3.2 Informal word list

Not every non-academic word could or should be revised. Many technical terms, for example "hydroponics" (Table 1), are too domain-specific for inclusion in the academic word list. The list also excludes many words, such as function words, that are essential building blocks of sentences. If the system attempts to replace every non-academic word with an academic counterpart, many substitutions may be superfluous. One solution is to train a classifier to determine if a non-academic word requires revision (Yimam et al., 2020).

While target word selection is an important task, it is not the focus of this study. We will assume as target words all non-academic words that are considered "informal" (e.g., the word "left" in Table 1), since informal words are likely to require revision in the academic context. For this purpose, we compiled a list of informal words, which include the 396 informal words from Brooke et al. (2010), and 4,198 words with human formality judgment (Pavlick and Nenkova, 2015). We will henceforth use the term *informal word* to refer to a word in this list.

3.3 Evaluation data

We derived our evaluation data from the 2,474 sentences in the Concepts in Context (CoInCo) corpus (Kremer et al., 2014). As an "all-words" lexical substitution dataset, every word in the text that can be replaced with another (i.e., a "target word") is manually annotated with the possible substitutions.

We retained only those target words that are both non-academic and informal (cf. Section 3.2). For example, neither the word "university" nor "hydroponics" in Table 1 is considered a target word: the former is academic; the latter is non-academic but is not informal, and therefore does not require revision. Further, a target word must have at least one gold substitution that is academic, though it may be formal or informal. This procedure left us with a total of 1,545 instances satisfying these conditions.

4 Approach

Our approach makes use of BERT (Devlin et al., 2019), a masked language model that has been shown to attain good performance in lexical substitution (Zhou et al., 2019), among many other tasks in natural language processing. We masked each target word in the input sentence, and then retrieved the top-N ranked list of candidates for the masked position from BERT. We will refer to this list as the "BERT list". We performed evaluation on values of N ranging from 1 to 50.

Approach	Precision	Recall	$F_{0.5}$
Baseline	20.91%	**17.94%**	0.2024
PPDB	19.43%	11.38%	0.1702
Gensim	28.81%	11.38%	0.2206
Lexical Formality	**34.22%**	15.52%	**0.2758**

Table 2: Experimental results on different filter approaches during substitution candidate generation

Words in the BERT list may not be academically appropriate or semantically similar to the target word. We removed non-academic words from the list, and then further filtered it with the following methods:

Baseline Keep all remaining words without further filtering.

PPDB Include only those words that are paraphrases of the target word in PPDB (Pavlick et al., 2015).

Gensim Include only those words that are among the 50 words most related to the target word, as estimated by the Google News pre-trained Gensim model (Mikolov et al., 2013).

Lexical Formality Include only those words that are considered formal equivalents of the target word. The formal equivalents are taken from the informal-formal word pairs from Brooke et al. (2010) and the Style Lexicon from Pavlick and Nenkova (2015), including both the manually crafted (4,196 pairs) and automatically generated (654,385 pairs) ones.

If no word survives the filter, the system does not attempt substitution. Otherwise, it predicts the top-ranked candidate as the substitution.

5 Experiments

We first used the annotated corpus (Section 3.3) to automatically evaluate the proposed methods. Then, we applied the optimal method on a user study.

5.1 Automatic Evaluation

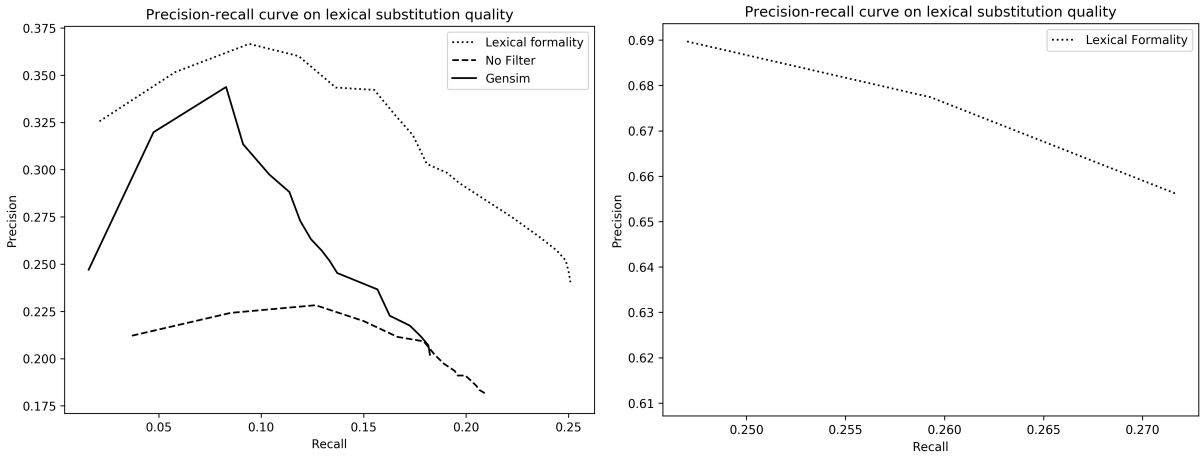
Precision is more important than recall since inappropriate suggestions can mislead students, who are expected to be the main users of the system. Following the shared task in grammatical error correction (Ng et al., 2014), we adopted the $F_{0.5}$ metric, placing twice as much emphasis on precision than recall.

We evaluated the proposed filtering methods (Section 4) on BERT lists whose length ranged from $N = 1$ to $N = 50$. As expected, a smaller N produced a conservative system that made fewer substitution attempts, and generally attained higher precision but lower recall. As shown in Figure 1a, at all N values, the Lexical Formality method outperformed the other methods in both precision and recall. It achieved the highest $F_{0.5}$ at $N = 20$.

Table 2 shows system performance on our evaluation dataset (Section 3.3) with the BERT list length fixed at $N = 20$. The baseline achieved 20.91% precision and 17.94% recall. Using PPDB lowered both the precision (19.43%) and recall (11.38%). Compared to the baseline, Gensim yielded higher precision (28.81%) at the cost of recall (11.38%). Incorporating Lexical Formality led to the best performance, in terms of precision (34.22%), recall (15.52%) and $F_{0.5}$ (0.2758). These results are unfortunately not directly comparable to Yimam et al. (2020), who reported only ranking correlation.

5.2 Manual Evaluation: Experts

To enable our subjects to consider a substantial number of lexical substitution within a reasonable time, we searched for short paragraphs of no more than 100 words with relatively large proportions of informal words. Within these constraints, we randomly selected 16 paragraphs taken from the British Academic Written English (BAWE) corpus (Nesi, 2008) and 4 paragraphs with academic topics from the Quora Question Answering Corpus (Sharma et al., 2019). These 20 paragraphs had an average length of 66 words. The system suggested a total of 81 substitutions for 38 target words.



(a) (b)

Figure 1: (a) Precision-recall curve for the three best performing filtering methods (Section 4) in the automatic evaluation (Section 5.1), using BERT lists whose length ranged from $N = 1$ to $N = 50$; (b) Precision-recall curve for the Lexical Formality method on the 20 paragraphs used in the manual evaluation (Section 5.2), over the same range of BERT list lengths

Two native English speakers, both PhD candidates in Linguistics, judged the quality of each suggestion on a three-point scale of "better", "same" and "worse", in comparison to the target word in the original text. The raters achieved a pairwise Cohen's Kappa of 0.65, which is considered "substantial agreement" (Landis and Koch, 1977). Overall, 28.39% (23) of the suggestions were rated "better", 19.75% (16) were rated "same", and 51.58% (42) were rated "worse". Figure 1b shows the precision-recall curve.

5.3 Manual Evaluation: Students

We conducted a user study with the 4 best-performing paragraphs from Section 5.2, with an average length of 54 words. The user study involved 35 subjects, all first-year undergraduate students who were non-native speakers of English taking a course on academic English at City University of Hong Kong. The subjects revised the paragraphs by selecting substitutions from the drop-down lists for the target words on a webpage. Of the 26 revision suggestions offered by the system for the 14 target words, 65.41% would improve the text. The breakdown includes 42.31% which were rated "better", 23.10% "same", and 34.62% rated "worse". Those rated as "same" also constitute an improvement, since they would replace the non-academic target word with an appropriate academic word.

The subjects selected a total of 396 substitutions, left the target word unchanged 84 times, and also supplied their own revision 10 times. Of the substitutions selected by the subjects, 80.30% were judged to be an improvement (with 62.88% rated "better", 17.42% "same", and 19.70% "worse"), substantially higher than the original 65.41%. These results suggest that students were able to avail themselves of the appropriate suggestions to improve the texts.

6 Conclusions

We have described a writing assistance system that automatically provides lexical substitution suggestions to incorporate more academic vocabulary. A significant novelty of the system is the use of lexical formality in selecting substitution candidates. Experimental results show that the formality analysis led to improved performance over a number of competitive baselines. Further, a user study demonstrates that students were able to use the system to improve text quality.

Acknowledgements

This project was supported by an HKSAR UGC Teaching & Learning Grant (Meeting the Challenge of Teaching and Learning Language in the University: Enhancing Linguistic Competence and Performance

in English and Chinese, 2016-19 Triennium), and by an Applied Research Grant (#9667151) from City University of Hong Kong.

References

Stephen Bailey. 2011. *Academic Writing: A Handbook for International Students*. Routledge, New York, NY.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic Acquisition of Lexical Formality. In *Proc. 23rd International Conference on Computational Linguistics (COLING)*.

Averil Coxhead. 2016a. A New Academic Word List. *TESOL Quarterly*, 34(2):213–238.

Averil Coxhead. 2016b. Reflecting on Coxhead (2000), "A New Academic Word List". *TESOL Quarterly*, 50(1):181–185.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

D. Gardner and M. Davies. 2013. A New Academic Vocabulary List. *Applied Linguistics*, 35(3):305–327.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What Substitutes Tell Us – Analysis of an "All-Words" Lexical Substitution Corpus. In *Proc. EACL*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.

John Lee, Eric L. M Cheung, Dariush Saberi, and Jonathan Webster. 2019. Expanding Students' Registerial Repertoire with a Writing Assistance Tool. *Journal of English for Academic Purposes*.

Diana McCarthy and Roberto Navigli. 2009. The English Lexical Substitution Task. *Language Resources and Evaluation*, 43:139–159.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. International Conference on Learning Representations (ICLR)*.

Hilary Nesi. 2008. BAWE: an introduction to a new resource. In A. Frankenberg Garcia, T. Rkibi, M. Braga da Cruz, R. Carvalho, C. Direito, and D. Santos-Rosa, editors, *Proc. Eighth Teaching and Language Corpora Conference*, page 239–46, Lisbon, Portugal. ISLA.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proc. 18th Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Gustavo H. Paetzold and Lucia Specia. 2016. Benchmarking Lexical Simplification Systems. In *Proc. LREC*.

Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Michael Bendersky, Marc A. Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. 2019. TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery (KDD)*, pages 2970–2978.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado, May–June. Association for Computational Linguistics.

Ellie Pavlick and Joel Tetreault. 2016. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better Paraphrase Ranking, Fine-grained Entailment Relations, Word Embeddings, and Style Classification. In *Proc. ACL*.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of Formal and Informal Sentences. In *Proc. 13th European Workshop on Natural Language Generation (ENLG)*, pages 187–193.

Seid Muhie Yimam, Gopalakrishnan Venkatesh, John Lee, and Chris Biemann. 2020. Automatic Compilation of Resources for Academic Writing and Evaluating with Informal Word Identification and Paraphrasing System. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A Corpus of Annotated Revisions for Studying Argumentative Writing. In *Proc. ACL*.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based Lexical Substitution. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 3368–3373.