Automatic detection of unexpected/erroneous collocations in learner corpus

Jen-Yu Li and Thomas Gaillat
Linguistique Ingénierie et Didactique
des Langues (LIDILE),
Université Rennes 2
Place du recteur Henri Le Moal,
CS 24307 - 35043 Rennes cedex,
France
jenyuli@gmail.com

Abstract

This research investigates the collocational errors made by English learners in a learner corpus. It focuses on the extraction of unexpected collocations. A system was proposed and implemented with open source toolkit. Firstly, the collocation extraction module was evaluated by a corpus with manually annotated collocations. Secondly, a standard collocation list was collected from a corpus of native speaker. Thirdly, a list of unexpected collocations was generated by extracting candidates from a learner corpus and discarding the standard collocations on the list. The overall performance was evaluated, and possible sources of error were pointed out for future improvement.

1 Introduction

Multiword expressions (MWEs) are word combinations which present lexical, syntactic, semantic, pragmatic or statistical idiosyncrasies. The boundary between MWEs and collocations is subtle. In Ramisch et al. (2018), they defined collocations as combinations of words whose idiosyncrasy is purely statistical and show no substantial semantic idiosyncrasy. In this way they oppose MWEs to collocations. Some researchers (Sag et al., 2002) regard collocations as any statistically significant cooccurrences, which include all kinds of MWEs. Some other researchers (Garcia et al., 2019; Baldwin and Kim, 2010) consider collocations as a subset of MWEs. For Tutin (2013), collocation is a category of semantic phraseme. As defined by Mel'čuk (1998), a phraseme is a set of phrase which is not free (without freedom of selection of its signified and without freedom of combination of its components). In this sense, the meaning of phraseme is quite similar to MWE. In this research, we considered collocation as a subset of semantic phraseme and a subset of MWEs as well. To constrain the set of collocation candidates, we focus on the Verb-Noun (VN) construction.

Second language learners usually have problems with collocations. Some researchers have reported that the errors are related to the learners' L1 (Nesselhauf, 2003; Hong et al., 2011). The correction of wrong collocations[1], such as *to *create [construct] a taller and safer building*, in written essays can help learners increase their competence and thus their proficiency in English writing (Meunier and Granger, 2008). Therefore, the automatic detection and correction of erroneous collocations would be helpful for

[1] In this research, the terms *wrong collocations, erroneous collocations, unexpected collocations*, and *collocational errors* are interchangeable.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 101–106
Barcelona, Spain (Online), December 13, 2020.

learners. Designing such a system would support specific feedback messages that could be employed to guide learners in their meta-cognitive learning processes (Shute 2008).

Such a system may be based on two kinds of corpora: a learner corpus which is used to extract known collocational errors, and a reference corpus to extract standard English collocations (Shei and Pain, 2000). Chang et al. (2008) proposed a method of bilingual collocation extraction from a parallel corpus to provide phrasal translation memory. Their system performance was exceptionally good (precision=0.98, recall=0.91). However, this approach required a bilingual dictionary, a parallel corpus for a specific L1 and English, as well as word-alignment matching of translations.

This paper presents a preliminary research on a learner corpus. In the following sections, we will briefly explain the method, present the results, and give some discussions.

2 Method

We propose a system to extract unexpected collocations in three stages: (a) implementation and evaluation of a collocation extraction module; (b) collection of standard collocations from a native corpus; (c) extraction of wrong collocations from a learner corpus. The main principle is, firstly, to extract all possible collocations in the learner corpus, and then identify standard collocations by the reference (collocations extracted from native corpus); the remainder of the items are considered as wrong collocations. Three evaluation points were made, aiming at the collocation extraction module, the reference of standard collocations, and the extraction of wrong collocations, respectively. The system diagram and the three stages are shown in Figure 1.

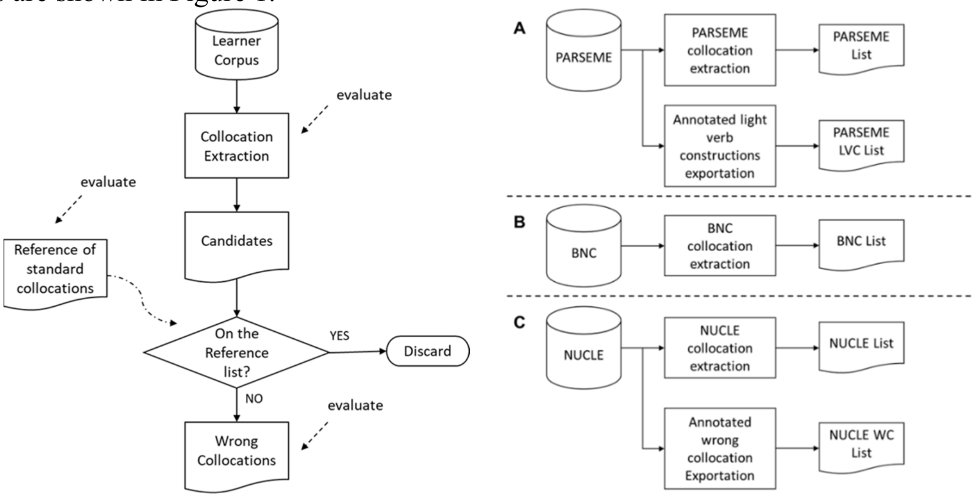


Figure 1. The system diagram and the three stages.

Stage A. Implementation and evaluation of the collocation extraction module: collocations were extracted from the PARSing and Multi-word Expressions (PARSEME[2]) corpus (Savary et al., 2015) with the implemented module. The results were saved as the PARSEME List. According to Garcia et al. (2019), light verb constructions (LVCs) can be regarded as collocations in VN form. The manually annotated LVCs were therefore retrieved and saved as the PARSEME LVC List. It is the gold standard (i.e. the ground truth) to evaluate the extraction module and to fine tune the parameters in the scripts.

Stage B. Collection of standard collocations: to have a large list of standard collocations, we used the implemented module to extract collocations from the British National Corpus (BNC[3]) (BNC Consortium, 2007) to form a list of standard collocations (the BNC List). The reference of standard collocations was built by merging the BNC List and the PARSEME LVC List. It was evaluated by manual verification. The errors in the reference list would degrade the credibility of our gold standard and thus might have a negative influence on the overall performance.

Stage C. Extraction of wrong collocations: we used the implemented module to extract candidate collocations (named as the NUCLE List) from the National University of Singapore Corpus of Learner

[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842
[3] https://ota.bodleian.ox.ac.uk/repository/xmlui/

English (NUCLE[4]) (Dahlmeier et al., 2013). The sentences manually annotated with erroneous collocations (*Wci* tag) were also exported, and the VN terms in these sentences were detected and saved in the NUCLE WC List. It was used to evaluate the overall performance of our system.

The scripts[5] were written in Python with Natural Language Toolkit (NLTK)[6] (Bird and Loper, 2004). Five lexical association measures were used in collocation extraction tasks, namely the raw frequency counting, t-test, chi-square test, log likelihood ratio, and pointwise mutual information. The formulas as well as an evaluation of 84 measures can be found in Pecina (2010).

3 Results

3.1 Evaluation of the collocation extraction module

To evaluate the module, we extracted the collocations from PARSEME and compared them with the PARSEME LVC List. The precision, recall, F_1 and $F_{0.5}$ scores were used as the accuracy metric. The best precision rate is 0.11 for the bigram detection with minimal frequency of 2, using raw frequency measure, and with the top 200 collocations. Meanwhile, the best recall rate is 0.11 when both bigram and trigram detection are used, and with minimal frequency equals 2 for top 300 collocations, with the log likelihood ratio or with the raw frequency measure. the best F_1 and $F_{0.5}$ are both 0.08 for the bigram detection using raw frequency measure with a minimal frequency of 2 and with top 300 collocations. Pointwise mutual information and chi-square methods cannot give good results even without applying filters. The results obtained by t-test methods are similar to raw frequency method. The window size was set to four. Shorter or longer window lengths were tried but did not have good results, which means the words of a collocation tends to co-occur in the span of four words.

3.2 Evaluation of the BNC list

For manual verification, 200 candidates were randomly sampled from the BNC list and given to an experienced English teacher. He validated firstly obvious collocations like *take place*. For the candidates that he was not sure about, he consulted the Corpus of Contemporary American English (COCA) collocate search tool[7]. If he found the candidate in the COCA corpus, it was validated; if not, the candidate was discarded. The final precision rate is 0.57.

3.3 Intersections between lists

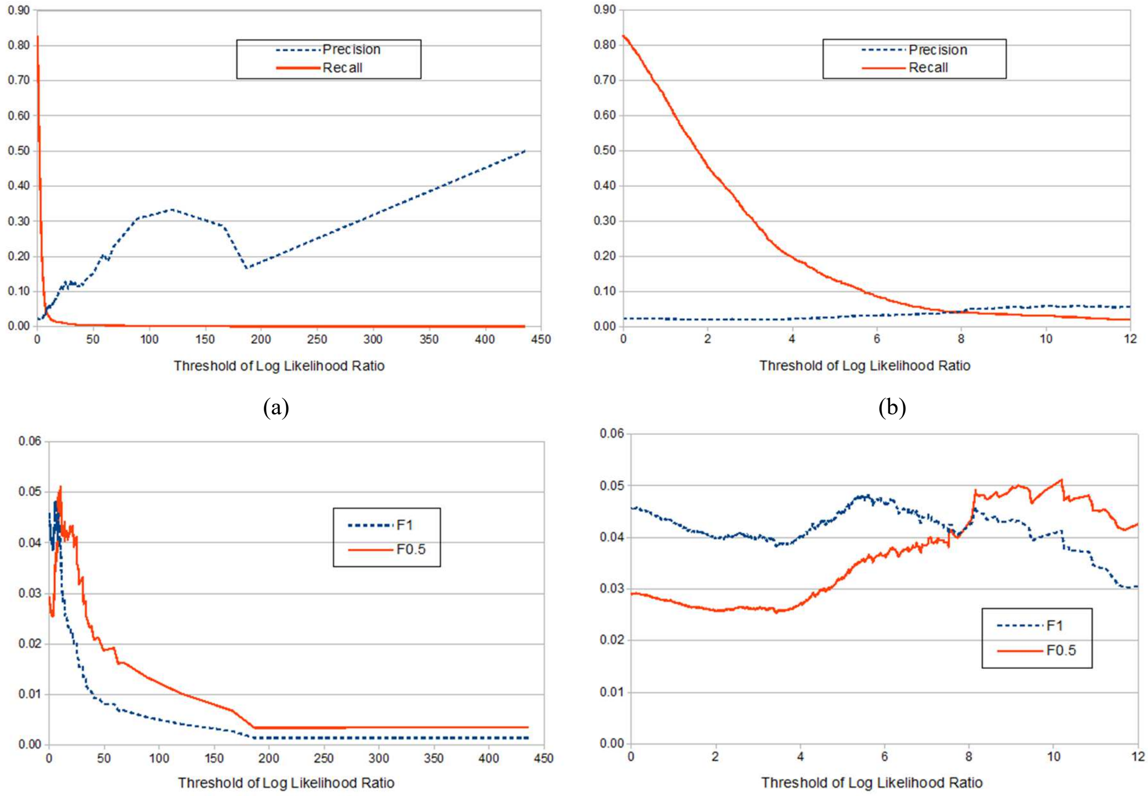
Ideally the union of the BNC List and the PARSEME LVC List (noted as **BNC ∪ PARSEME LVC**) gives us the standard collocations, and NUCLE WC List gives the wrong collocations. Ideally there should be no overlapping in standard and wrong collocations. However, we found that there are intersections between the NUCLE WC List and the PARSEME LVC (11 collocations), between the NUCLE WC List and the BNC List (20 collocations), and between all three lists (4 collocations). The amount of this overlapping is therefore 27 (20+11-4=27), noted as **NUCLE WC ∩ (BNC ∪ PARSEME LVC)**; it is about 1.8% of the NUCLE WC List.

3.4 Optimization by selecting a threshold of Log Likelihood Ratio

Candidates were extracted from NUCLE and compared with the gold standard, i.e. the NUCLE WC List (1,471 erroneous VN collocations). Various thresholds of log likelihood ratio were tested for optimization. Figure 2(a) shows the global view of precision and recall versus different thresholds, and Figure 2(b) gives a zoom-in of threshold from zero to twelve. The highest precision is 0.5 when the threshold value is set to 430, where only two candidates are extracted. The precision and recall meet at the same level about 0.04 when the threshold is set to eight, and 1,408 candidates are extracted. The maximal

[4] NUCLE is a collection of 1,414 essays (in a total of 1.2 million words) written by students who are non-native English speakers. It is available by submitting a license agreement via https://www.comp.nus.edu.sg/~nlp/corpora.html
[5] Source codes are available online: https://github.com/jenyuli/wrong_collocation_extraction
[6] https://www.nltk.org/
[7] https://www.english-corpora.org/coca/

recall (0.83) is obtained by extracting all possible candidates (54,471), and the precision becomes extremely low (0.02).



(a) (b)

(c) (d)

Figure 2. Precision, Recall, F_1 and $F_{0.5}$ scores versus log likelihood ratio.

Figure 2(c) and 2(d) demonstrate the global view and zoom-in of the F_1 and $F_{0.5}$ trends. We can see that the $F_{0.5}$ reaches its peaks (0.05) when the threshold is set to eight or ten; while the F_1 fluctuates around 0.04 to 0.05 when threshold is set lower than eight. Considering all four indices, the optimal value of the threshold can be set about eight.

4 Discussions and conclusion

As our experiment configuration is capable to extract wrong collocations from the leaner corpus, the overall performance is not satisfactory. Hence, we reviewed the results and point out some possible sources of errors for future studies.

First, regarding the PARSEME corpus, the gold standard was built based on the *LVC* tag, so it may be that the verbs of the collocations were biased. In fact, 44 out of 85 collocations on the list were constructed only by five verbs, namely *do, get, give, have,* and *take*. Therefore, the evaluation of the module was also biased. Regarding the BNC List, we have reached a precision of 0.57 due to the large size of corpus (100 million words) and a strict selection (top 10 for each sub-directory of the BNC). However, comparing with a previous study (Jian et al., 2004) which extracted 631,638 VN collocations from the BNC, we found that our standard collocation reference list (BNC ∪ PARSEME LVC) was much smaller (n=942) and may have a negative influence on the performance. Regarding the NUCLE, because the Part-Of-Speech (POS) and the lemma are not available, we used a POS tagger and a Lemmatizer. Yet, their performances were not evaluated, so the gold standard NUCLE WC List was not perfectly accurate. As for the whole system, it may be helpful to incorporate a word dependency parser module to identify the object noun which received the action of the verb.

Our approach has shown a method to detect erroneous collocations in learner English. As it relies on the accurate extraction of a reference list, our next step will consist in exploring larger corpora for extraction. Such an extraction module would be of great benefit as part of a Computer Aided Language Learning System dedicated to the analysis of phraseology in learner texts.

Reference

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. Chapman and Hall/CRC, Boca Raton, FL, USA, Second edition.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

BNC Consortium, 2007, *The British National Corpus*. Distributed by Bodleian Libraries, University of Oxford.

Yu-Chia Chang, Jason S. Chang, Hao-Jan Chen, and Hsien-Chin Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299, July.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Dana Gablasova, Vaclav Brezina, and Tony McEnery. 2017. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, 67(S1):155–179.

Marcos Garcia, Marcos García Salido, Susana Sotelo, Estela Mosqueira, and Margarita Alonso-Ramos. 2019. Pay Attention when you Pay the Bills. A Multilingual Corpus with Dependency-based and Semantic Annotation of Collocations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4012–4019, Florence, Italy, July. Association for Computational Linguistics.

Ang Leng Hong, Hajar Abdul Rahim, Tan Kim Hua, and Khazriyati Salehuddin. 2011. Collocations in Malaysian English learners' writing: A corpus-based error analysis. *3L: The Southeast Asian Journal of English Language Studies*, 17(Special Issue):31–44.

Jia-Yan Jian, Yu-Chia Chang, and Jason S. Chang. 2004. Collocational Translation Memory Extraction Based on Statistical and Linguistic Information. In *Proceedings of the 16th Conference on Computational Linguistics and Speech Processing*, pages 257–264, Taipei, Taiwan, September. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Batia Laufer and Tina Waldman. 2011. Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning*, 61(2):647–672.

Claudia Leacock. 2010. Collocation Errors. In *Automated grammatical error detection for language learners*, pages 63–71. Morgan & Claypool Publishers, California.

Igor Mel'čuk. 1998. Collocations and Lexical Functions. In Anthony P. Cowie, editor, *Phraseology: theory, analysis, and applications*, Oxford linguistics, pages 23–53. Oxford Univ. Press, Oxford.

Fanny Meunier and Sylviane Granger, editors. 2008. *Phraseology in foreign language learning and teaching*. John Benjamins Pub. Co, Amsterdam ; Philadelphia.

Nadja Nesselhauf. 2003. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, 24(2):223–242, June.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1/2):137–158.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, et al. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, United States, August. Association for Computational Linguistics.

Ute Römer. 2005. Section 4.5.4 Verbs and objects [BNC/BoE]. In *Progressives, patterns. pedagogy: a corpus-driven approach to English progressive forms, functions, contexts, and didactics*, pages 130–135. J. Benjamins Pub. Co, Amsterdam ; Philadelphia.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, number 2276, pages 1–15. Springer.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November.

Chi-Chiang Shei and Helen Pain. 2000. An ESL Writer's Collocational Aid. *Computer Assisted Language Learning*, 13(2):167–182, April.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2019. Verbal Multiword Expressions: Idiomaticity and flexibility. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions: Current trends*, pages 35–68. Language Science Press, Berlin.

Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78(1):153–89, March.

Agnès Tutin. 2013. Les collocations lexicales : une relation essentiellement binaire définie par la relation prédicat-argument. *Langages*, n° 189(1):47–63, April.