Semi-automatic Extraction of Academic Transdisciplinary Phraseology: Expanding a Corpus of Categorized Academic Phraseology Using BERT Machine Learning Models (Short Paper)

Micaela Aguiar [1,], José Monteiro [1] and Sílvia Araújo [1]

[1] *University of Minho, Rua da Universidade, Braga, 4710-057 , Braga*

Abstract

The lack of access to academic literacy skills and tools is a serious problem, as it furthers inequality among students. In this paper, we propose a methodology to semi-automatically extend a corpus of academic phraseology, previously manually extracted and categorized, using BERT machine learning models. We begin by describing the constitution of the manually extracted and categorized corpus. Next, we briefly discuss how the BERT machine learning model works. Then, we explore the methodology for the semi-automatic extension of the initial corpus: from the constitution of a new corpus, to the preparation of the documents to be processed by the model, to the manual evaluation of the retrieved sentences. We ran two tests, the results of which we report in this paper.

Keywords

academic phraseology, transdisciplinary scientific phraseology, natural language processing, BERT, semi-automatic extraction

1. Introduction

Many are the reasons given for the lack of academic literacy among higher education students. Some point the finger at earlier levels of education, and, in particular, at language subjects, whose curriculum is often focused mainly on literary texts [1]. Others point out that teachers themselves sometimes "have insufficient meta-language skills and knowledge to discuss writing issues with students and to explain their expectations with respect to student assignments" [2]. Institutions try to address the issue by offering academic writing courses. However, not all institutions have the same offer and a lot of the time these types of courses are subject to fees. This promotes inequality in the access to academic literacy skills.

As part of the research project PortLinguE, a Portuguese project financed by European funds, we are developing a tool that will assist Portuguese and non-native students in academic writing tasks. This tool will be freely available to all students and one of its goals is to bridge the gap in the access to academic literacy skills. This tool centers around the creation of a phrase bank of European Portuguese academic phraseology, that was manually extracted and categorized.

This paper describes a methodology to expand the initial corpus of manually extracted and categorized academic phrases using BERT machine learning models. We will begin by addressing the framework of academic phraseology, which will be followed by a brief description of the constitution of the initial corpus and the description of the methodology for the corpus expansion, ending with results of the tests we ran.

2nd International Conference on "Multilingual Digital Terminology Today, Design, Representation Formats and Management Systems", (MDTT) 2023, June 29-30, 2023, Lisbon, Portugal

EMAIL: maguiar60@gmail.com (M. Aguiar); jdiogoxmonteiro@gmail.com (J. Monteiro); saraujo@elach.uminho.pt (S. Araújo)
ORCID: 0000-0002-5923-9257 (M. Aguiar); 0000-0002-2904-3501 (J. Monteiro); 0000-0003-4321-4511 (S. Araújo)

2. Framework

Formulaic language is favored by native speakers in their communication [3], however, using academic formulas is not a “linguistic universal skill” [4], in fact failing to use formulaic language may even be considered a signal of “lack of mastery of a novice writer in a specific disciplinary community” [5]. The concept of “transdisciplinary scientific lexicon” accounts for the non terminological lexicon common to the scientific community and widely shared by the disciplines [6]. Transdisciplinary scientific lexicon may refer to single words, multiword referential sequences (like collocations and fixed expressions), multiword discursive sequences (recurring expressions used to structure the speech), multiword interpersonal sequences (expression used to convey pragmatic or modal functions) and semantic-rhetorical routines [7]. According to Tutin, semantic-rhetorical routines are typical utterances of scientific writing corresponding to a specific rhetorical function: they correspond to complete statements, built around a predicate.

An academic phrase bank is a list of collocations, discourse markers, hedging devices, but mostly of semantic-rhetorical routines, that perform various functions, such as referring to sources, describing the results of an experiment or stating the conclusions of a study. There are some academic phrase banks available online, such as the Ref-N-Write Academic Phrasebank for English (https://www.ref-n-write.com/academic-phrasebank/) or the Dictionnaire des expressions from Base ARTES for multiple languages (https://artes.app.univ-paris-diderot.fr/artes-symfony/web/app.php/fr). In Portuguese, Bab.la (https://en.bab.la/phrases/academic/opening/english-portuguese) offers a multilingual Portuguese phrase bank. Morley [8] developed the most popular English phrase bank, the Manchester University's Academic Phrasebank (https://www.phrasebank.manchester.ac.uk/). The Academic Phrasebank corpus originally consisted of 100 postgraduate dissertations from the University of Manchester, and has since incorporated academic material from a variety of sources. Morley drew on Swale's concept of move to manually extract and categorize sections of text serving a particular communicative purpose. There aren’t many academic phrase banks exclusively for European Portuguese, so we are developing one as a freely available tool for academic literacy.

3. Manually Extracted Corpus

To create the European Portuguese academic phrase bank, we started with a corpus of 40 scientific papers taken from RepositoriUM, the repository of the University of Minho, and Repositório Aberto, the repository of the University of Porto. The papers were divided into four scientific areas: Life and Health Sciences, Exact and Engineering Sciences, Natural and Environmental Sciences, and Social Sciences and Humanities. Papers were only included in the corpus if they were written in European Portuguese and were available in open access.

Taking into account the pedagogical nature of the phrase bank, we elected to create a corpus of scientific papers, because the scientific paper as a genre is ubiquitous in the work of any researcher and because it can be used as a model for young researchers and undergraduates. As Tulin and Jacques [6] point out, the lexicon shared by scientific productions is a lexicon of genre. That is why, with regard to the semantic-rhetorical categories, we used an adapted version of the typology put forward by Morley.

Five main categories were considered: introduction, literature review, methodology, results, discussion and conclusions. These categories were informed by the concept of included genders (*genre inclu* in the French original) [9]. Included genres refers to sections of text such as the introduction or the conclusion that can be, for example, found in distinct genres, such as the scientific paper, the doctoral thesis or the conference paper. The extraction and categorization was carried out using the qualitative data analysis and mixed methods software, MAXQDA. Afterwards, the extracted phraseological units were simplified: any particular content was removed or substituted by more general terms. This highlights the phraseological element, makes it easier for students to use it in their writing and discourages plagiarism. Fifty sub-categories were identified and nearly a thousand phraseological units were extracted.

4. Extending the Corpus

Tulin [7] points out that, unlike other types of lexical sequences, the extraction and categorization of the semantic-rhetorical routines is hard to automatize, due to its large lexical, syntactic and semantic variety. To try to automatically extract this type of more complex phraseological units, we will utilize the manually extracted and categorized corpus as a starting point to find similar elements in a larger corpus, using the natural language processing model BERT [10].

4.1. Bert Model

BERT (Bidirectional Encoder Representations from Transformers) is a Natural Language Processing model that analyzes text corpus in terms of similarities at word, collocation and sentence level and distributes the processed data based on semantic similarity, thus generating semantic vectors. Transformers [11] are a deep learning model, first introduced in the 2017 paper "Attention is all you need" [12], that uses attention (a concept that arose in NLP and Machine Translation) to weigh the relevance of each part of the input data. Similarly to humans, previous recurrent neural nets (RNNs) were not able to process long sentences, because RNN's process them sequentially (left to right or right to left) and tended to "to forget information from timesteps that are far behind" [13]. The concept of attention addresses this problem by proposing to "look at all the different words at the same time and learn to 'pay attention' to the correct ones depending on the task at hand" [13]. Transformers are capable of reading a sentence in both directions at the same time (hence the concept of bidirectionality). Since Transformers can process data in any order, it enabled the models to train in large amounts of data and create pre-trained models [14].

BERT is the most popular and used pre-trained model in Natural Language Processing. BERT is usually trained for two main purposes: masked language modeling (which entails predicting a randomly masked word) and Next Sentence Prediction (which involves predicting if two sentences are consecutive or not [15]. This way, BERT is able to process context since "words are defined by their surroundings, not by a pre-fixed identity" [14]. This feature enables BERT to perform semantic searches. This type of search is unique given that it seeks to "determine the intent and contextual meaning of the words a person is using for a search". We will train BERT models to perform semantic searches and find similar phrases and structures in a large text corpus, using the initial manually extracted and categorized corpus.

4.2. Methodology

In this section, we will describe the steps in the methodology we propose to semi-automatically extend a corpus of previously categorized academic phraseology.

The first step in this methodology is to compile a new corpus. The new corpus will be composed of 40 PhD dissertations and 40 Master's thesis, drawn from the four disciplinary areas mentioned above. The new corpus will meet the same inclusion criteria as the original corpus: being available in open access and written in European Portuguese. This is the only step where human involvement is necessary, and our system only necessitates that the user creates a file with the repository link of the files he wishes to extract.

The second step is to prepare the collected documents to be processed by the model. Generally speaking, academic texts are deposited in repositories in pdf format. These documents must be downloaded and then converted into text format. To perform this task, our *python* pipeline uses the requests module to download the documents (using the URLs provided by the user). Then, the *pdfplumber* [16] package allows for extracting the text from each document.

After the documents have been converted into text format, the third step is to extract individual sentences from each document. To parse each sentence, we opted to use the NLTK (natural language tool-kit) [17] platform for *python*, specifically the *portuguese.pickles tokenizer*.

The fourth step is to use the BERT model to convert the individual extracted sentences into semantic vectors, and then store them in an efficient text search database FAISS. Instead of manually producing the embeddings and managing the database, we opted to use Haystack [18], a end-to-end framework that enables the construction of powerful and production-ready pipelines for searching text. Our Haystack pipeline was set up to use an efficient FAISS database and a pre-trained BERT model called BERTimbau [19]. BERTimbau has the limitation of being a model trained with the BrWaC corpus (Brazilian Web as Corpus). However, there are no large BERT models trained for European Portuguese yet.

With our Haystack framework working and fully indexed with sentences, the fifth step consists in finding, for each phrase of the original corpus, the top 10 most similar sentences from the new corpus. The framework facilitates this process by having methods for querying the database. These methods have two purposes: first, they convert the original phrase into a sentence vector (using the same pre-trained model); second, they perform a dot product search between the new query vector and the vectors stored in the database. The results from performing the search are 10 similar phrases from the new corpus being recommended. Given the semantic nature of the process, the phrases should be similar in both content and context.

The sixth step is the manual evaluation of the results.

5. Results & Discussion

In the application of this methodology, we first tested the rufimelo/Legal-BERTimbau-sts-base-ma-v2 [20] model and evaluated the results manually. This first trial served for both exploratory and baseline purposes. We found that the results of this first test were not particularly positive. However, they still showed promising results for the application of our methodology. For this reason we decided to run a new test, this time using the larger rufimelo/Legal-BERTimbau-sts-large-ma-v3 [21] model. We also carry out a manual evaluation of the results of the second model, whose results we will describe next.

For the manual evaluation, we defined 8 categories: 1 - it corresponds to a variation of the template; 2 - it is a repetition of the template; 3 - it deviates from the original meaning, by focusing on words in the context; 4 - it is not an expression (titles, numbers, punctuation, etc...) or it is not a complete expression; 5 - it is an expression that is not related to the original template; 6 - it is an ambiguous expression, that only the context can clarify; 7 - it is an expression that corresponds to another function; 8 - it is an expression in a language other than Portuguese. The manual evaluation worked as follows: for each of the 50 subcategories in the phrase bank, 3 templates were categorized (two template-proposed sentences per template).

The initial corpus contains variables like x, y, and z that replace expressions with particular content, so in evaluating the first model we evaluated the results with variables and without variables (we determined that the variables were a source of noise for the model). That is why for the first model we evaluated 600 sentences and for the second we evaluated 300. The results of the two models are described in Table 1.

Table 1
Model Comparison

Categories	Model 1	Model 2
Total	600	300
1 - Match	121 - 20,1%	134 - 45,5%
2 - Repetition	0 - 0%	0 - 0%
3 - Contextual Focus	187 - 31,1%	25 - 4,1%
4 - No Expression	125 - 20,8%	8 - 2,6%
5 - Random	97 - 16,1%	38 -12,6%
6 - Ambiguous	58 - 9,6%	11 - 3,6%
7 - Mismatch	10 - 1,6%	63 - 21%
8 - Other Language	4 - 0,6%	21 -7%

As can be seen, there was a considerable increase in the amount of matches from the first model (20.1%) to the second (45.5%). The first model presented many instances where the results were bound to a word from the context of the template (31,1%), kept for the sake of readability; the second (4,1%) model presented considerably fewer occurrences of this problem. Another result that should be highlighted is the increase in mismatch cases from the first model (1,6%) to the second (21%). We categorized as mismatches occurrences in which the expression corresponds to another function, usually very close to the one presented by the model. This is because the phrase bank categories often present nuances that are difficult to resolve, for example, between the functions of "reporting unexpected results" , "commenting on the results" or "summarizing the results". In these cases the expressions will later be incorporated into the most appropriate functions. A limitation of this method is that it is not always possible to tell which plane of the text (introduction, methodology, conclusions, etc.) the expression is taken from, which explains the percentage of ambiguous results. Of the 50 categories, the first model showed positive results for 31 categories, the second for 41 categories.

In the future, we will try to eliminate other causes of noise that we have identified for the models, such as numerals, dates and place-holder names, like Smith and Jones, to see if we can achieve better results. We will also run a new test with the second model using a larger corpus of thesis and dissertations.

The final step in our work will be to annotate all occurrences of the model with the best results; recategorize the sentences that have been mismatched, and finally, transform the sentences into simplified templates in order to be incorporated into the phrase bank. Phrases containing expressions that already exist in the phrase bank will be used as examples of the template in context.

6. Conclusions

In this paper, we proposed a methodology to semi-automatically extend a corpus of categorized academic phraseology using machine learning models, BERT. The aim was to enrich the European Portuguese academic phrase bank, which is being developed within the PortLinguE project and will be made available as a tool to support academic literacy.

In the future, this extended phrase bank will be available for free online and in pdf format on the digital platform created by the PortLinguE project, called Lang2Science. Furthermore, we intend to embed this phrase bank in a search engine, using technology already developed within the PortLinguEproject. The search engine also uses BERT models, so this integration will allow users to search for expressions, similar to what they do when using Google, and obtain similar phraseology or phraseology with similar functions as a result. This offers users a more dynamic way to interact with the phrase bank.

7. Acknowledgements

This work was carried out within the scope of the "PortLinguE" project (PTDC / LLT-LIG / 31113/2017) financed by FEDER under Portugal 2020 and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT,I.P.).

8. References

[1] J. A. Brandão, Literacia Académica: Da Escola Básica Ao Ensino Superior – Uma Visão Integradora, Letras & Letras (2013) 17.

[2] K.M. Jonsmoen, and M. Greek, 2017, 'Lecturers' text competencies and guidance towards academic literacy', Educational Action Research, 25(3) (2017) 354–69.

[3] A. Wray, Formulaic language and the lexicon. Cambridge University Press, Cambridge, 2002.

[4] C. Pérez-Llantada, Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage, Journal of English for Academic Purposes, 14 (2014) 84–94.

[5] J. Li, and N. Schmitt, The acquisition of lexical phrases in academic writing: a longitudinal case study, Journal of Second Language Writing, 18(2) (2009) 85–10.

[6] A. Tutin, and M.-P. Jacques, Le lexique scientifique transdisciplinaire : une introduction, in: M.-P. Jacques and A. Tutin (Eds.), Lexique transversal et formules discursives des sciences humaines, ISTE Editions, 2018, pp.1-26.

[7] A. Tutin, La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico-rhétoriques, in: A. Tutin and F. Grossmann (Eds.) L'écrit scientifique : du lexique au discours. Autour de Scientext, Presses Universitaires de Rennes, 2014, pp. 27-44.

[8] J. Morley, A compendium of commonly used phrasal elements in academic English in PDF format, The University of Manchester, 2014.

[9] F. Rastier, Arts et sciences du texte, PUF, Paris, 2001.

[10] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL (2019).

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in: 31st Conference on Neural Information Processing Systems Long Beach, CA, USA, 2017

[13] N. Adaloglou, How Attention works in Deep Learning: understanding the attention mechanism in sequence models, 2020, URL: https://theaisummer.com/attention/.

[14] B. Lutkevich, BERT language model, 2020, URL: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

[15] H. Tayyar Madabushi, L. Romain, D. Divjak, and P. Milin, CxGBERT: BERT meets Construction Grammar, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4020–4032. https://doi.org/10.18653/v1/2020.coling-main.355

[16] pdfplumber. URL: https://github.com/jsvine/pdfplumber

[17] NLTK. URL: https://www.nltk.org/

[18] haystack. URL: https://github.com/deepset-ai/haystack

[19] F. Souza, R. Nogueira, R. Lotufo, BERTimbau: Pretrained BERT Models for Brazilian Portuguese, in: R. Cerri, R. C. Prati (Eds.), Intelligent Systems, BRACIS 2020, Springer, Cham, 2020, https://doi.org/10.1007/978-3-030-61377-8_28.

[20] rufimelo/Legal-BERTimbau-sts-base-ma-v2. URL: https://huggingface.co/rufimelo/Legal-BERTimbau-sts-base-ma-v2

[21] rufimelo/Legal-BERTimbau-sts-large-ma-v3. URL: https://huggingface.co/rufimelo/Legal-BERTimbau-sts-large-ma-v3