Typicality of Lexical Bundles in Different Sections of Scientific Articles

Haotong WANG
Waseda University
Kitakyushu, Fukuoka, Japan
+81-93-692-5287
wanghaotong0925@toki.waseda.jp

Yves LEPAGE
Waseda University
Kitakyushu, Fukuoka, Japan
+81-93-692-5287
yves.lepage@waseda.jp

Chooi Ling GOH
The University of Kitakyushu
Kitakyushu, Fukuoka, Japan
+81-93-695-3832
goh@kitakyu-u.ac.jp

ABSTRACT

This paper proposes a method to quantify the typicality of lexical bundles in sections of academic articles, specifically in the field of Natural Language Processing papers. Typicality is defined as the product of individual KL-divergence scores and the probability of a bundle to appear in a type of section. An evaluation of our typicality measure against two other baselines shows slight improvements according to the Silhouette coefficient.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Natural language processing**; **Lexical semantics**.

KEYWORDS

lexical bundles; typicality; writing aid

1 INTRODUCTION

1.1 Writing Aids

A *writing aid* is a computer environment, the purpose of which is to assist a person in composing a text. As it is a computer environment, Natural Language Processing (NLP) techniques can be used. Grammatical errors, lack of fluency, improper style, collocation errors, etc. [8] should be dealt with in a standard writing aid.

The purpose of an *academic writing aid* is, more specifically, to assist a researcher in composing a scientific article. As English has become the *de facto lingua franca* of many fields in science, we are concerned here with academic writing aids for composing scientific articles in English. For non-native speakers, the level of knowledge and proficiency in the language, certainly motivates the development of writing aid systems and technologies. However, not only non-native speakers of English, but also experienced writers who are native speakers, may feel a need for such tools.

Scientific texts are characterised by a certain style, one feature of which is the use of *lexical bundles* [1, 13], i.e., frequent sequences of lexical items specific to a certain style or domain. Properly composed texts need to contain proper lexical bundles. Consequently, writers need an access to such bundles, for various reasons like:

- exactness in expression: e.g., what is the correct usage for 'we note/denote something as/by/with'?;
- tip of the tongue phenomenon: e.g., what is a synonymous expression for 'as a whole/to sum up'?;
- search for originality in expression: e.g., what are the various ways of saying 'we thank the reviewers'?;
- etc.

1.2 Ranking Lexical Bundles

Retrieval of lexical bundles has been addressed in previous pieces of work. In [2], a keyword-in-context (KWIC) search engine, called Linggle, is proposed to retrieve lexical bundles for a given query from the entire Web. It can supply information for preposition usage, collocation of a verb with an object, etc. The output is ranked by frequency. Academic word suggestion machine (AWSuM) [1] extracts frequent n-grams from articles from different disciplines and sections of articles, and ranks them by frequency. Other systems like StringNet[2], Just the Words[3], WriteAway[4] search for patterns around syntactic unit. They all give users statistics, mainly frequency of use, for the retrieved candidates.

The metric used to rank lexical bundles in all the above works is thus frequency. This of course reflects the importance of lexical bundles to some extent, so that suggestions are more reliable. It is natural following one of the first definitions of lexical bundles [1]:

> "Lexical bundles are defined as the *most frequent*[5] recurring lexical sequences; however, they are usually not complete structural units, and usually not fixed expressions."

However, *typicality*, i.e., the fact that a lexical bundle is used in the right domain or in the right section of a document, is also an important criterion for suggestion. A scientific article is composed of several typical sections: abstract, introduction, experimental settings, conclusion, etc. which are characterised not only by lexical bundles, but also, e.g., by usage in tense (preferably past tense in the conclusion, sometimes future tense in introduction and present tense elsewhere). In the evaluation of the AWSuM system, it was noticed that the typicality of the lexical bundles for the different sections would provide more accurate suggestions [11].

The goal of this paper is thus to retrieve typical lexical bundles for different sections of scientific papers. Intuitively, typical lexical bundles of a section should of course be more used in that section

ICNLP 2020, July 11–13, 2020, Guangzhou, China

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

[1] http://langtest.jp/awsum/

[2] http://nav4.stringnet.org

[3] http://www.just-the-word.com

[4] http://writeaway.nlpweb.org

[5] The emphasis is ours.

and relatively less used in other sections. Expectedly too, they should rather be not common to all sections. We propose a metric to estimate the typicality of lexical bundles. It is a combination of an individual Kullback-Leibler divergence weight and the probability for the cluster to be classified in the section considered.

The paper is structured as follows. The next section, Section 2, gives a reminder on lexical bundles. Section 3 describes our proposed measure of typicality. Section 4 describes our experiments, i.e., the data used, the settings and the results. It also provides a method to asses the quality of our measure of typicality against two other baseline measures. Section 5 concludes the paper.

2 LEXICAL BUNDLES

Although they lack a universally accepted definition, lexical bundles have been shown to play an important role in fluent linguistic production. They help in making sense in a specific context and they strengthen the coherence of the text [7]. Grossly speaking, lexical bundles are an extension of the notion of collocation, They are N-grams which appear more frequently than expected. Lexical bundles can be seen as a kind of formulaic expressions [9]. The following are several well-accepted features [4]:

- Highly frequent and recurrent lexical sequences: this explains why they are often presented by decreasing order of frequency;
- Incomplete structure: they may end up with articles, prepositions, etc.;
- Shared by multiple sources: they are not characteristic of individual style, but of a shared common style.

To know and use lexical bundles is essential for writers who are non-native speakers. Bundles, used as units, keep sentences fluid and is a mark of higher language proficiency, much more than skilled usage of individual words. The work in [1] identified the initial concept of lexical bundles by the observation of language use in classes and textbooks. In [5], the discovered pedagogical implication of using lexical bundles on improving writing quality was demonstrated. [3] showed that published academic papers exhibit the widest range of lexical bundles. Numerous patterns and collocations exist. Texts or speech produced by non-native speakers of English exhibited a small range of lexical bundles, while overusing rare expressions (e.g. ‘in the long run’). All these experiments showed that the quantity of lexical bundles used by authors is a reliable measure of proficiency in the language.

3 A MEASURE OF TYPICALITY

The Kullback–Leibler (KL)-divergence is a measure of how different two distributions are. We rely on it to estimate, for a given bundle, in which types of section it is preferably used. Our measure of typicality combines this estimation with the probability of the bundle to appear in the given type of section. Such a probability can be computed using a classifier.

3.1 Individual KL-Divergence

The Kullback–Leibler divergence, also called relative entropy, is an indicator of the extent to which two probability distributions match [10]. The higher the value, the more different the distributions. The definition is as follows:

$$D_{KL}\left(p \parallel q\right) = \sum_{i=1}^{N} p\left(x_i\right) log\left(\frac{p\left(x_i\right)}{q\left(x_i\right)}\right) \tag{1}$$

where x_i is an event shared by the two distributions, and $p\left(x_i\right)$ and $q\left(x_i\right)$ are the probabilities of event x_i in the two distributions.

The KL-divergence is the sum over all items shared by the two distributions. As we want to assess how common a lexical bundle b_i is to two types of section X and Y, we compute an individual divergence $D_i\left(X \parallel Y\right)$ as follows:

$$D_{b_i}\left(X \parallel Y\right) = p_X\left(b_i\right) \log\left(\frac{p_X\left(b_i\right)}{p_Y\left(b_i\right)}\right) \tag{2}$$

where $p_X(b_i)$ is the probability of bundle b_i to appear in type of section X. This probability is computed from the number of occurrences of each bundle in each type of section, divided by the sum of the number of occurrences over all the types of section.

Now, for a given bundle and a given type of section, we take the average value of the individual KL-divergences over all other types of section to obtain a value which represents how much the given bundle is preferred in the given type of section. Below, N is the total number of types of section.

$$D(b_i, X) = \frac{1}{N-1}\left(\sum_{Y \neq X} D_{b_i}\left(X \parallel Y\right)\right) \tag{3}$$

3.2 Classification Model

To estimate the chance to encounter a given lexical bundle in a given type of section, we train a classifier. To compute a vector representation of N-grams, we use a pre-trained model from Bidirectional Encoder Representation from Transformers (BERT) [6], which combines masked language modelling and next sentence prediction to catch dynamic phrase or sentence-level representation.

Our classification model is a fully connected layer, with cross-entropy as the loss function and with a softmax layer at the top, as our problem is a multi-class classification problem (each type of section is a class). The model takes the vector representation from BERT and outputs a vector of values in the range of $[0; 1]$ which represent the probabilities for the bundle to appear in each type of section. For a given bundle b_i and a type of section X, we call $p(X|b_i)$ the probability for the bundle b_i to be classified in X.

3.3 Typicality Score

Our final score is the combination of the two previous scores. For a given bundle b_i and a given type of section X, we pose:

$$T(b_i, X) = D(b_i, X) \times p(X|b_i) \tag{4}$$

The scores range in the interval $[-1; 1]$. Higher scores mean that the bundle is more typical for the considered type of section.

Table 1: Number of N-gram and number of sentences in the training set and the test test

	ACL-ARC	
	N-grams	Sents
All items	17,782,741	392,555
Training set	13,337,056	294,416
Test set	4,445,685	98,139

4 EXPERIMENTS

4.1 Dataset

The ACL Anthology Reference Corpus (ACL-ARC)[6] is a corpus of scholarly publications published by or in association with the Association for Computational Linguistics (ACL). It comprises 21,520 articles from year 1979 to year 2015.

Our study concentrates on three types of sections: abstract, introduction, conclusion. From the provided ParsCit structured XML files, we could extract 19,385 abstracts, 16,473 introductions and 16,317 conclusions. Our dataset is thus well-balanced, The smaller number of introductions and conclusions is due to articles in which the section may not exist or cannot be extracted based on the given title.

We generate candidate lexical bundles from the above data and restrain ourselves to 3-grams, 4-grams and 5-grams.

We use a pre-trained BERT model [7] for our classifier. Table 1 shows the number of sentences and N-grams used for training and testing the accuracy of the model.

4.2 Evaluation

We assess the quality of our typicality measure by comparison to simple ranking by frequency or Kullback-Leibler divergence.

It is natural to assert that a lexical bundle typical of a section should appear in that section with high frequency, while it should comparatively appear less often in other types of section. In all the above ranking schemes (typicality measure, frequency and KL–divergence) the higher value, the more typical the lexical bundle should be. For each type of sections, higher values should reflect a closer distance to the centroid of the values for the type of section under consideration. The Silhouette coefficient measures interpretation and validation of consistency within clusters of data [12]. It is defined in Formula (5).

$$s\left(b_i, X\right) = \frac{\text{Intra}(b_i, X) - \text{Inter}(b_i, X)}{\max\left\{\text{Intra}(b_i, X), \text{Inter}(b_i, X)\right\}} \tag{5}$$

Intra (b_i, X) is the average distance of lexical bundle b_i to other lexical bundles in the same type of section X, Inter (b_i, X) is the average distance of bundle b_i to all other lexical bundles in all other sections. The distance we consider is the score of the lexical bundle (frequency, Kullback-Leibler divergence or typicality measure), normalised in the range $[0; 1]$.

The Silhouette coefficient is in the range of $[-1, 1]$. A bundle appearing in only one section has a Silhouette coefficient of 1 for that type of section. Silhouette coefficient of -1 means a bundle frequently appearing in every type of section.

Table 2: Accuracy of classification of lexical bundles into types of sections (abstract, introduction and conclusion)

	ACL-ARC	
	N-grams	Sents
Accuracy (%)	66.9	75.3

Table 3: Average Silhouette coefficient for ranking according to three measures, over all lexical bundles

Frequency	0.6097
KL–divergence	0.6145
Typicality measure (ours)	0.6233

4.3 Results

The accuracy of the classification model presented in Section 3.2 is shown in Table 2. It is more than three quarters for entire sentences but falls down to two-thirds for N-grams, i.e., candidate lexical bundles. This relatively low score may be explained by the number of lexical bundles shared by the three types of sections.

Tables 4 show the top 5 lexical bundles typical of each type of section, abstract, introduction and conclusion, plus 5 other typical lexical bundles drawn at random. The proposed measure of typicality ranks typical lexical bundles in seemingly the right sections. The top 5 typical lexical bundles are quite convincing.

By comparing the column Typicality score and Freq. in these tables, one sees that the typicality score ranks lexical bundles differently than a simple ranking by frequency. The question of the quality of these rankings is evaluated by the Silhouette coefficient that we presented in Section 4.2. As baselines, to compare our typicality measure, we use the simple ranking by frequency or KL-divergence. Table 3 reports the average values over all lexical bundles for all types of sections for the three scores we considered: frequency, KL–divergence and typicality measure. The Silhouette value for our proposed typicality measure is slightly higher than for the two other scores. This points at the fact that our score better identifies typical lexical bundles.

5 CONCLUSION

We proposed a measure for typicality of lexical bundles in different sections of documents. The measure is the product of individual KL-divergence scores and the probability of a lexical bundle to appear in a type of section.

We applied our method on scientific papers from the domain of natural language processing. The types of sections we considered were abstract, introduction and conclusion. Our proposed measure is able to better rank typical lexical bundles used in the right sections in higher positions than a simple ranking by frequency. This impression is confirmed by the Silhouette coefficient scores against two baseline rankings.

Non-native speakers of English usually lack the capability of using lexical bundles, which leads to improper use of expressions

[6] https://acl-arc.comp.nus.edu.sg
[7] https://github.com/huggingface/transformers

Table 4: Top 5 selected lexical bundles typical in abstract plus 5 other lexical bundles selected at random

Bundles in abstract	Freq.	KL-divergence ($\times 10^{-3}$)	Classification probability	Typicality score ($\times 10^{-3}$)
in this paper we present	375	22.548	0.493	11.107
we present a	1059	17.434	0.613	10.686
in this paper we propose	355	18.119	0.503	9.108
this paper presents a	304	12.305	0.614	7.557
this paper we present a	228	14.732	0.503	7.403
⋮				
statistical machine translation	387	3.087	0.398	1.227
present an algorithm for	19	1.324	0.865	1.145
we present an algorithm that	10	1.258	0.860	1.082
paper describes an approach	14	0.881	0.817	0.720
among the participating systems	8	0.404	0.739	0.298
⋮				
Bundles in introduction				
natural language processing nlp	445	25.624	0.597	15.293
in recent years	413	16.153	0.858	13.854
is the task of	236	12.893	0.784	10.102
in natural language processing	361	19.809	0.472	9.342
natural language processing	1221	17.472	0.484	8.451
⋮				
in many natural language processing	58	8.682	0.541	4.695
is one of the	179	8.487	0.518	4.400
in the field	105	3.186	0.519	1.654
tasks in natural language processing	24	2.870	0.634	1.818
the meaning of	96	2.858	0.484	1.382
⋮				
Bundles in conclusion				
in this paper we have	561	127.218	0.559	71.123
we have presented a	539	65.054	0.998	64.914
we have shown that	462	54.391	0.990	53.871
we have presented	976	51.488	0.998	51.389
in this paper we presented	231	51.589	0.994	51.267
⋮				
we have studied	28	0.723	0.979	0.708
the results presented	25	0.624	0.998	0.623
have described a method	10	0.440	0.998	0.439
this work we presented an	5	0.318	0.997	0.317
results are promising	19	0.435	0.685	0.298
⋮				

in their writing. In the frame of an academic writing aid to help non-native speakers in writing scientific papers in English, the typicality measure that we proposed will be useful for retrieving more relevant lexical bundles by suggesting typical lexical bundles for the sections the writers are composing.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP18K11 446 entitled: “Natural language processing for academic writing in English”.

REFERENCES

[1] Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. *If you look at…*: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25, 3 (09 2004), 371–405. https://doi.org/10.1093/applin/25.3.371 arXiv:https://academic.oup.com/applij/article-pdf/25/3/371/431268/250371.pdf

[2] Joanne Boisson, Ting-Hui Kao, Jian-Cheng Wu, Tzu-Hsi Yen, and Jason S. Chang. 2013. Linggle: a Web-scale Linguistic Search Engine for Words in Context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol. System Demonstrations. Association for Computational Linguistics, Sofia, Bulgaria, 139–144. https://www.aclweb.org/anthology/P13-4024

[3] Y. H. Chen and P. Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology* 14, 2 (1 6 2010), 30–49. https://www.researchgate.net/publication/45681690_Lexical_Bundles_in_L1_and_L2_Academic_Writing

[4] Susan Conrad and D. Biber. 2004. The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica* 20 (01 2004), 56–71.

[5] Viviana Cortes. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23, 4 (2004), 397 – 423. https://doi.org/10.1016/j.esp.2003.12.001

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27, 1 (2008), 4 – 21. https://doi.org/10.1016/j.esp.2007.06.001

[8] Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the Rough: Generating Fluent Sentences from Early-Stage Drafts for Academic Writing Assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 40–53. https://doi.org/10.18653/v1/W19-8606

[9] Kenichi Iwatsuki and Akiko Aizawa. 2018. Using Formulaic Expressions in Writing Assistance Systems. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2678–2689. https://www.aclweb.org/anthology/C18-1227

[10] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Ann. Math. Statist.* 22, 1 (03 1951), 79–86. https://doi.org/10.1214/aoms/1177729694

[11] Atsushi Mizumoto. 2017. Initial Evaluation of AWSuM: A Pilot Study. *Vocabulary Learning and Instruction* 6, 2 (dec 2017), 46–51. https://ci.nii.ac.jp/naid/120006408125/en/

[12] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[13] Danica Salazar. 2014. *Lexical bundles in native and non-native scientific writing*. Studies in corpus linguistics, Vol. 65. John Benjamins Publishing Company.