Extraction of Lexical Bundles used in Natural Language Processing Articles

Chooi Ling GOH
The University of Kitakyushu
Kitakyushu, Japan
Email: goh@kitakyu-u.ac.jp

Yves LEPAGE
Waseda University
Kitakyushu, Japan
Email: yves.lepage@waseda.jp

***Abstract*—Lexical bundles are indispensable for fluent academic writing. They might not constitute complete structural units but they occur very frequently in academic conversations, conference presentations and scientific articles. This paper shows how to collect a large database of lexical bundles from articles in the Natural Language Processing (NLP) domain. We first collect highly frequent N-grams from the ACL-ARC collection of NLP articles and then classify them into true or false lexical bundles using machine learning models trained from a set of manually checked bundles. In a verification experiment, our best model achieves an accuracy of 76 %. Using this model, we extract more than 18,000 lexical bundles from the ACL-ARC corpus, which we publicly release.**

I. Introduction

Lexical bundles are a rather recent concept in corpus linguistics. Although they usually do not make complete structural units, they are frequently used in academic conversations, conference presentations and scientific articles. They have now been studied for two decades, but, before that, linguists only focused on idioms and collocations when teaching students how to compose essays in English. However, it was found that they help greatly in improving the fluency of writing and also contribute to facilitate the readers' comprehension. In a study of theses written in English by Czech students, Dontcheva-Navratilova [1] refers to Wray [2] to stress that:

> formulaicity benefits both the speaker/writer and the listener/reader by facilitating discourse processing and thus enhancing the perception of discourse coherence.

The work presented in this paper is part of a more general work which aims at helping researchers who are not native speakers of English in composing academic articles in English, by offering them writing aids under computer interfaces. Such researchers often encounter problems to compose scientific articles in a fluent manner because their level of proficiency, i.e., their rapid access to a large amount of culturally well-established expressions, is not enough developed. In such a view, a large database of lexical bundles is indispensable so that a computer can assist them.

Our work proceeds as follows. We first carry out a survey on the use of lexical bundles in natural language processing (NLP) scientific articles. For that, we use the ACL Anthology Reference Corpus (ACL-ARC) as our corpus because it is a corpus of recognised NLP articles of high linguistic quality. We collect highly frequent N-gram candidates (N ranging from 3 to 6) from this corpus. A good part of such candidates does not comply to the definition of lexical bundles. Therefore, we resort to machine learning techniques to classify these candidate N-grams into true and false lexical bundles. We use supervised learning to train the models, i.e., we use a list of bundles established in a previous work by Salazar [3] as our training dataset.

The paper is organised as follows. Section II discusses previous work on lexical bundles. It clarifies our goal and illustrates it. Section III describes the selection of the most accurate machine learning method to classify N-gram candidates into true and false lexical bundles. Section IV describes the use of the most accurate learning method to extract a database of lexical bundles from the corpus of scientific articles in NLP. Section V presents the database of lexical bundles and discusses our findings on the extracted lexical bundles.

II. Lexical Bundles

A. Definitions in Linguistic Terms

Idioms are invariable expressions, the meaning of which cannot be predicted from the meaning of the parts. For instance, "a rule of thumb" is an idiom. It refers to "a broadly accurate guide or principle, based on practice rather than theory" according to the Oxford Dictionary of English. As can be seen, the meaning of the word thumb ("the short, thick first digit of the human hand…") does not enter in the above definition.

Collocations are pairs of words which co-occur more frequently than by chance. They are associated together by statistics rather than by their semantics. For instance, the verb "to make" in its different word forms and the noun "decision" make a collocation. Indeed, linguistically, the verb to make is analysed as a light verb for the noun decision. This collocation can be replaced by the single verb "to decide."

By contrast to idioms and collocations, lexical bundles are word forms which co-occur in longer sequences of words. They can be regarded as an extension of collocations. Biber [4] characterise them in the following way:

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

In this definition, three points are made clear. Firstly, high frequency characterises lexical bundles. Secondly, lexical bundles may be structurally complete ("the best of our knowledge"), or incomplete ("at the end of the", 'is due to the fact that"). Thirdly, there may be some variability among them ("thank the anonymous reviewers for their", "thank the reviewers for their", "thanks go to the reviewers who").

B. Linguistic and Statistical Perspectives

Initial studies on lexical bundles focused on the conversation and academic prose [4], [5], [6]. Biber et al. [5] find it is surprising that recurrent lexical bundles occur so frequently in academic prose, much more than in conversation. About 21 % of the words in a piece of writing were estimated to occur in a lexical bundle. These initial studies considered word sequences occurring 10 times per million words. Although it is said that lexical bundles might not make complete structural units, in Chapter 13 of *Longman Grammar of Spoken and Written English* [5], lexical bundles are categorised into 14 major structural categories for conversation and 12 major structural categories for academic prose, according to their structural correlates.

Lexical bundles have also been studied in classroom teaching and university textbooks. Biber et al. [7] use a higher frequency cut-off of 40 times per million words. They found that classroom teaching uses a larger set of lexical bundles than conversation, but conversation relies on frequent use of a smaller set of bundles. The same goes for textbooks and academic prose: more different lexical bundles are found in textbooks, but they are used more frequently in academic prose.

Salazar [3] further refined the definition of lexical bundles, using a number of negative criteria to exclude some N-gram candidates. Some of these criteria are: N-grams ending in articles, N-grams composed exclusively of function words, or fragments of other bundles. Such criteria may be subject to discussion and some contradict previous proposals.

Most of previous work identified lexical bundles manually, after automatic extraction of N-grams from a corpus. We propose to select the most accurate machine learning method to classify N-gram candidates into true and false lexical bundles after training on manually checked training data.

C. Use of Lexical Bundles by Non-Native Authors

By contrast to studies which compare professional and students' writing in history and biology [8] in their use of lexical bundles, there has been a range of studies comparing native and non-native writers in various fields. Salazar [3] extracted target lexical bundles from a collection of published articles in biology and biochemistry, and compare their use with a smaller corpus of biomedical research articles written by Spanish scientists, all non-native speakers of English. Safarzadeh et al. [9] showed that Persian-speaking writers employed the same forms of lexical bundles as native speakers, but there were significant differences concerning the nativeness and functions. Based on a comparison with Chinese students' essays, Chen and Baker [10] argue that the lexical bundles found in native experts' writing can be of great help to learners/writers to achieve a more native-like style of academic writing. However, they must be selected and edited carefully before they are integrated into the English as a Second Language (ESL) or English as a Foreign Language (EFL) curricula. Bychkovska and Lee [11] identify some common bundles misused in non-native Chinese students' writing. Ädel and Erman [12] investigated the use of lexical bundles in advanced learner writing by speakers of Swedish in the discipline of linguistics. These previous works have shown that lexical bundles should be used actively in writing fluent academic articles.

D. Use of Lexical Bundles in a Writing Aid for Non-native Authors

Since it has been proven that lexical bundles are frequently used by both native and non-native speakers in academic writing, it is worth studying the lexical bundles used in a specific field. Our intention is to extract a large database of lexical bundles in order to help non-native speakers of English to compose their articles. Our ultimate goal is to integrate these lexical bundles into an academic writing aid.

We envisage several ways for the use of a database of lexical bundles. We mention two of them below:

- Computation of cosine similarity between sequences of words: the word vector representations of the words at the beginning of a sentence composed by a non-native writer may be used to retrieve most similar bundles. For instance, the lexical bundle with the highest cosine similarity to the improper N-gram "By compared to the method", typical of Chinese writers, is "By comparison with the method". This approach may allow a system to propose bundles which exhibit variable forms, like "thank the anonymous reviewers for the" instead of "thank the reviewers for their".
- Access by lemmas: from the lemmatised form of "consisted", a system may be able to identify typical errors like "is consisted of" in texts produced by Chinese authors and propose correct corresponding lexical bundles, like "our method consists in <number> steps".

The examples above show that a database of lexical bundles can improve the writing style of authors who are not native speakers of English, while teaching them correct English expressions at the same time.

III. Classification of N-gram Candidates into True or False Lexical Bundles

A. Training Data for Supervised Learning

Our training data for supervised learning comes from the work by Salazar [3]. The distribution of N-gram candidates is shown in Table II. Out of them, there are 769 true lexical bundles (labeled as Y) and 963 false lexical bundles (labeled as N) according to Salazar's definitions.

	Accuracy		Precision	Recall	F-score	Precision	Recall	F-score
	(10-fold)	(test)	for label Y (true bundles)			for label N (false bundles)		
BernoulliNB	0.723	0.686	0.669	0.593	0.628	0.697	0.762	0.728
SVM	0.749	0.746	0.723	**0.701**	0.712	0.763	0.782	0.773
MLPerceptron	**0.771**	**0.762**	**0.751**	**0.701**	**0.725**	**0.770**	**0.812**	**0.790**

TABLE I
ACCURACY OF THREE MACHINE LEARNING METHODS.

	3-gram	4-gram	5-gram	6-gram
Salazar [3]	1,443	247	34	8

TABLE II
STATISTICS ON THE N-GRAMS COLLECTED BY SALAZAR [3].

	bundles (Y)	not bundles (N)	Total
All	769	963	1,732
Train	575	724	1,299
Test	194	239	433

TABLE III
SUMMARY OF THE TRAINING DATA COLLECTED FROM SALAZAR [3] AND THEIR PROPORTION USED IN OUR EXPERIMENTS

We randomly divide the data into 75% for training and 25% for testing. Table III shows statistics of the experiment data. Salazar's data has a vocabulary of 648 words.

B. Machine Learning Models

We compare three machine learning methods: Bernoulli Naïve Bayes, Support Vector Machines and Multi-layer Perceptron, to build three classification models. We use the implementations provided in the `scikit-learn` package[1]. The main parameters used for training are described below. The details are given in the Appendix for the sake of reproducibility. The three models, which all perform the same task of classifying frequent N-grams into true (Y) or false (N) lexical bundles, using the same features, i.e., bag-of-words from the N-grams, are as follows:

- Naïve Bayes: Multi-variate Bernoulli model performs better than multinomial model in our case, as we only have a small set of data, and hence small vocabulary sizes [13]. Most of the parameters use default values.
- Support Vector Classifier [14]: We have tuned the model with different kernels, C values for margin, degrees and gammas. Linear kernel worked the best for our task. The rest of the parameters are given in the Appendix.
- Multi-layer Perceptron [15]: Tuning was performed on activation functions, solver methods, hidden layer sizes and alpha values. The parameters determined by tuning are given in the Appendix.

C. Selection of the Most Accurate Machine Learning Model

Table I gives the results obtained by the three machine learning methods. The first column is the accuracy for 10-fold cross validation. The second column and after use 1,299 bundles for training and 433 bundles for testing as described in Table III. Evaluation is done with accuracy for all labels and independently on each label (Y or N) for precision, recall and F-score (harmonic mean of precision and recall).

From the experiment results shown in Table I, the Multi-layer Perceptron model gives the best accuracy and achieves about 76% accuracy.

IV. EXTRACTION OF LEXICAL BUNDLES USED IN NLP ARTICLES

A. Data Used and Statistics

Following a trend in natural language processing, we apply our methods to the NLP research field, i.e., we apply NLP methods on NLP data. This has been called NLP4NLP in [16], [17], [18]. We use the ACL Anthology Reference Corpus[2] (ACL-ARC hereafter) as our NLP domain corpus. It is a subset of ACL Anthology[3] which is a digital archive of research papers in the premium conferences in NLP. The English language quality of the papers is reputed.

ACL-ARC consists of publications about computational linguistics and natural language processing from selected conferences and journals since 1979 until 2015. Plain texts are collected from the Omnipage OCR XML files provided. Out of all the 22,878 articles in the corpus, we could exploit the plain text of 21,636 articles. The number of tokens there is 88,006,598, for 578,960 types (i.e., distinct words).

B. Extraction of N-gram Candidates

We extracted highly frequent N-grams from the corpus. The articles only were used for extraction, meaning that we excluded front pages of conferences or workshops. We left titles, authors, addresses and references in the articles. Paragraphs were segmented into sentences and tokenised into words before N-gram extraction. Biber et al. [5] consider word sequences that recur at least ten times per million words in a given register and that spread across at least five different pieces of texts as potential N-gram candidates for bundles. In our approach, there is no difference in register, so we choose to extract all N-grams that occur more than 100 times in the entire corpus.

At the beginning, we used the whole content of articles for extraction. However, noise is caused by conference or proceeding names appearing in reference sections and in

[1] https://scikit-learn.org/

[2] https://acl-arc.comp.nus.edu.sg/

[3] https://aclanthology.coli.uni-saarland.de/

	3-gram	4-gram	5-gram	6-gram	Total
Filtered	43,523	11,720	2,262	554	58,059
True lexical bundles (Y)	12,895	4,355	836	188	18,274
Newly discovered	12,437	4,293	833	188	17,751
Existing in Salazar's data	430	62	3	0	495
Mis-classified	28	0	0	0	28
Percentage	29.6%	37.2%	37.0%	33.9%	31.5%
False lexical bundles (N)	30,628	7,365	1,426	366	39,785
Percentage	70.4%	62.8%	63.0%	66.1%	68.5%

TABLE IV
LEXICAL BUNDLES EXTRACTED FROM ACL-ARC. 'MIS-CLASSIFIED' ARE N-GRAM CANDIDATES WHICH ARE FALSE LEXICAL BUNDLES IN THE TRAINING DATA.

	3-gram	4-gram	5-gram	6-gram
With References				
Extracted	68,045	27,598	10,077	4,988
Filtered	48,181	14,448	4,173	1,964
Without References				
Extracted	59,590	21,372	5,825	1,910
Filtered	43,523	11,720	2,262	554

TABLE V
STATISTICS ON THE N-GRAMS EXTRACTED FROM ACL-ARC ARTICLES WITH AND WITHOUT REFERENCES. FILTERED N-GRAMS ARE THOSE WITHOUT PUNCTUATIONS AND ARABIC NUMBERS.

footers on almost every first page of an article. We could exclude the reference sections but were unable to consistently recognise footers using the XML text files of the output of Omnipage OCR. So we excluded the references but left the footers. As there is also noise caused by punctuations and numbers, we filtered out N-grams containing punctuations and Arabic numbers. Table V gives the statistics of the N-grams left for consideration as a result of our cleaning process.

C. Classification into True and False Lexical Bundles

Relying on the results of Section III-C, we use the Multi-layer Perceptron model to classify the frequent N-grams extracted from ACL-ARC (filtered without references). A version of this model was trained from all available training data (1,732 bundles) with the parameters determined previously.

Table IV shows the result of the application of this model on our data. In total, 31.5 % of the N-grams are labeled as true lexical bundles and 68.5 % are false lexical bundles. For 3-gram candidates, only about 30 % of them are classified as true lexical bundles. Out of these bundles, 430 bundles exist in the training data, 28 of them were mis-classified, i.e., they are false lexical bundles according to the training data, and 12,437 bundles are new. About 37 % of the 4-gram and 5-gram candidates are labeled as true lexical bundles, about 34 % for the 6-gram candidates. 62 and 3 lexical bundles respectively already exist in the training data for 4-gram and 5-gram candidates. No 4-, 5- and 6-gram true lexical bundle was mis-classified.

V. A DATABASE OF LEXICAL BUNDLES USED IN NLP ARTICLES

In this section, we analyse some of the bundles classified by the Multi-layer Perceptron model: existing lexical bundles, new lexical bundles and false bundles. Let us stress that we do not make Salazar's criteria our absolute standard because our goal is not a characterisation of lexical bundles, but the creation of a database of lexical bundles for use in an academic writing aid. For instance, N-grams included in larger bundles or N-grams ending with articles are rejected by Salazar, but they may well be of great utility in our system.

A. Lexical Bundles Existing in Salazar's Data

The lexical bundles automatically extracted from ACL-ARC which also exist in our training data from [3] are listed in Table VI. The top-5 bundles with highest frequencies occur more than the number of papers in the corpus. Putting it in another way, these lexical bundles are used in almost every paper. They are: "the number of", "in order to", "a set of", "in this paper" and "as well as". Bundles which occur more than half of the number of papers are: "in terms of", "the use of", "with respect to" and "on the other hand", etc. Such bundles should be used actively when writing academic articles and a writing aid system should absolutely recommend them to a non-native author.

B. Newly Discovered Lexical Bundles

The automatic extraction of lexical bundles from the ACL-ARC corpus allowed us to discover a large number of new lexical bundles, proper to NLP articles. The top-20 new ones are shown in Table VII for each length of N-gram. Although we excluded references, there is still some noise caused by the names of conferences on the first page of articles. Apart from that, the rest of the newly discovered lexical bundles look good: "the set of", "in this section", "it is possible to" and "to be able to", etc. It is obvious that the longer the N-grams, the more difficult to get possible lexical bundles. Many of them are just noise. However, we can still obtain some good lexical bundles such as "due to the fact that", "in the same way as", "it is important to note that", "it is interesting to note that", etc. for N equal to or greater than 5. These lexical bundles can surely contribute in helping non-native speakers in composing academic articles.

C. False Negatives

We also examined the false bundles automatically classified as such by our Multi-layer Perceptron. Table VIII shows the top-20 most frequent false bundles for each length of N-gram. Some could possibly be considered as true lexical bundles. This is no surprise because of the relatively low accuracy (76 %) of the Multi-layer Perceptron model. It looks like the longer the N-grams, the lower the accuracy. This can be attributed to the small amount of training data for longer N-grams. More training data for longer N-grams should be collected so as to improve the accuracy.

VI. Conclusion

The use of lexical bundles is essential for fluent academic writing, Non-native speakers of English usually lack the capability of using bundles in their writing. In this paper, we extracted lexical bundles from the ACL-ARC corpus, a large corpus of scientific articles in the NLP domain, so that they can be used as a reference for writing NLP articles in English. Around 32 % of highly frequent N-grams were classified as true lexical bundles using a supervised machine learning model. This amounts to 18,000 new lexical bundles which we make publicly available[4].

As for future work, we intend to collect more training data from different sources so as to improve the accuracy of our machine learning models. More features can be added to the machine learning models in addition to our current approach which relies essentially on bags-of-words. We also think of a way to build a model that is tolerant to unknown words, as we only have a small-sized vocabulary with which it is difficult to deal with unknown words.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K11446.

References

[1] O. Dontcheva-Navratilova, “Lexical bundles in academic texts by non-native speakers,” *Brno Studies in English*, vol. 38, pp. 37–58, January 2012. [Online]. Available: https://www.researchgate.net/publication/272852564_Lexical_Bundles_in_Academic_Texts_by_Non-native_Speakers

[2] A. Wray, “Formulaic sequences in second language teaching: principle and practice,” *Applied Linguistics*, vol. 21, no. 4, pp. 487–489, 2000.

[3] D. J. L. Salazar, “Lexical bundles in scientific English: A corpus-based study of native and non-native writing,” Ph.D. dissertation, Universitat de Barcelona, 2011.

[4] D. Biber and S. Conrad, “Lexical bundles in conversation and academic prose,” in *Out of corpora: Studies in honour of Stig Johansson*, H. H. . S. O. (Eds.), Ed. Amsterdam: Rodopi, 1999, pp. 181–190.

[5] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English.* Longman, 1999, ch. 13, pp. 990–1024.

[6] S. Conrad and D. Biber, “The frequency and use of lexical bundles in conversation and academic prose,” *Lexicographica*, vol. 20, pp. 56–71, 2004.

[7] D. Biber, S. Conrad, and V. Cortes, “*If you look at . . .*: Lexical bundles in university teaching and textbooks,” *Applied Linguistics*, vol. 25, no. 3, pp. 371–405, 2004.

[8] V. Cortes, “Lexical bundles in published and student disciplinary writing: Examples from history and biology,” *English for Specific Purposes*, vol. 23, no. 4, pp. 397–423, 2004.

[9] M. M. Safarzadeh, A. Monfared, and M. Sarfeju, “Native and non-native use of lexical bundles in discussion section of political science articles,” *Iranian Journal of Applied Language Studies*, vol. 5, no. 2, pp. 137–166, 2013.

[10] Y.-H. Chen and P. Baker, “Lexical bundles in L1 and L2 academic writing,” *Language Learning & Technology*, vol. 14, no. 2, pp. 30–49, 2010.

[11] T. Bychkovska and J. J. Lee, “*At the same time*: Lexical bundles in L1 and L2 university student argumentative writing,” *Journal of English for Academic Purposes*, vol. 30, pp. 38–52, 2017.

[12] A. Ädel and B. Erman, “Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach,” *English for Specific Purposes*, vol. 31, no. 2, pp. 81–92, 2012.

[13] A. McCallum and K. Nigam, “A comparison of event models for Naïve Bayes text classification,” in *Proceedings of AAAI/ICML-98 Workshop on Learning for Text Categorization*, 1998, pp. 41–48.

[14] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] G. E. Hinton, “Connectionist learning procedures,” *Artificial intelligence*, vol. 40, no. 1–3, pp. 185–234, 1989.

[16] G. Francopoulo, J.-J. Mariani, and P. Paroubek, “NLP4NLP: Applying NLP to scientific corpora about written and spoken language processing,” in *Proceedings of the First Workshop on Mining Scientific Papers: CLBib@ISSI*, 2015, pp. 5–11.

[17] J. Mariani, G. Francopoulo, and P. Paroubek, “The NLP4NLP corpus (I): 50 years of publication, collaboration and citation in speech and language processing,” *Frontiers in Research Metrics and Analytics*, vol. 3, p. 36, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/frma.2018.00036

[18] J. Mariani, G. Francopoulo, P. Paroubek, and F. Vernier, “The NLP4NLP corpus (II): 50 years of research in speech and language processing,” *Frontiers in Research Metrics and Analytics*, vol. 3, p. 37, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/frma.2018.00037

Appendix

For reproducibility of experiments, we provide hereafter the parameters of the machine learning models of Section III-B.

```
BernoulliNB(alpha=1.0, binarize=0.0,
  class_prior=None, fit_prior=True)

SVC(C=1.0, cache_size=200,
  class_weight=None, coef0=0.0,
  decision_function_shape='ovr',
  degree=3, gamma='auto',
  kernel='linear',
  max_iter=-1, probability=False,
  random_state=None, shrinking=True,
  tol=0.001, verbose=False)

MLPClassifier(activation='relu',
  alpha=0.1, batch_size='auto',
  beta_1=0.9, beta_2=0.999,
  early_stopping=False, epsilon=1e-08,
  hidden_layer_sizes=(200,),
  learning_rate='constant',
  learning_rate_init=0.001,
  max_iter=300, momentum=0.9,
  nesterovs_momentum=True,
  power_t=0.5, random_state=None,
  shuffle=True, solver='lbfgs',
  tol=0.0001, validation_fraction=0.1,
  verbose=False, warm_start=False)
```

[4] http://lepage-lab.ips.waseda.ac.jp/ > Projects > Kakenhi Kiban C 18K11446 > Experimental Data

3-gram	the number of (54625), in order to (39700), a set of (33402), in this paper (32174), as well as (27462), in terms of (21066), the use of (18984), with respect to (14845), can be used (14827), a number of (14342), is based on (12387), the results of (12141), be used to (11697), in the same (10642), shown in figure (10472), shown in table (10441), in this case (9175), as shown in (9090), in addition to (9046), the context of (8425)
4-gram	on the other hand (11355), in the case of (8019), in the context of (6202), on the basis of (5761), the total number of (5346), are shown in table (4696), at the same time (3953), a large number of (3844), as shown in figure (3177), at the end of (3108), is shown in figure (3088), it is important to (2759), a wide range of (2320), by the fact that (1689), as a function of (1605), as a result of (1603), in the number of (1525), we have shown that (1367), at the level of (1361), it should be noted (1300)
5-gram	it should be noted that (1141), has been shown to be (544), it has been shown that (451)
6-gram	None

TABLE VI
TOP-20 LEXICAL BUNDLES EXTRACTED FROM ACL-ARC WHICH EXIST IN THE TRAINING DATA. NUMBER OF OCCURRENCES IN PARENTHESES. DOUBLE UNDERLINE FOR LEXICAL BUNDLES WHICH OCCUR MORE TIMES THAN THE NUMBER OF ARTICLES. SINGLE UNDERLINE FOR MORE THAN HALF.

3-gram	the set of (23235), association for computational (13809), in this section (9710), is used to (9272), in the following (9062), described in section (8020), a list of (8007), would like to (7225), of this paper (6610), in natural language (6290), a sequence of (6241), of a word (6137), in this work (5979), in our experiments (5798), are used to (5769), is defined as (5453), of the data (5381), in other words (5184), we need to (5081), of a sentence (5006)
4-gram	association for computational linguistics (12315), can be used to (7423), in this paper we (6526), is the number of (6484), we would like to (5006), it is possible to (4445), in the form of (3967), in terms of the (3723), is the set of (3602), would like to thank (3188), to be able to (3082), in the next section (3052), in proceedings of the (2944), can be found in (2914), of the association for (2893), we can see that (2850), is a set of (2808), in this section we (2721), meeting of the association (2524), in the previous section (2517)
5-gram	of the association for computational (2807), annual meeting of the association (2450), meeting of the association for (2355), paper is organized as follows (2227), due to the fact that (1973), we would like to thank (1845), empirical methods in natural language (1553), conference on empirical methods in (1550), on empirical methods in natural (1525), results are shown in table (1465), methods in natural language pages (1152), this paper is organized as (1099), this work was supported by (1024), is the total number of (1017), n is the number of (996), was supported in part by (984), can be seen as a (965), in the same way as (947), the number of words in (932), in this paper we present (882)
6-gram	meeting of the association for computational (2330), annual meeting of the association for (2286), on empirical methods in natural language (1524), conference on empirical methods in natural (1523), of the association for computational linguistics (1352), of the association for computational pages (1277), empirical methods in natural language pages (1146), this paper is organized as follows (1070), it is important to note that (747), of this paper is organized as (728), of the paper is organized as (694), where n is the number of (559), work was supported in part by (503), rest of the paper is organized (502), the rest of this paper is (499), this work was supported in part (493), it is interesting to note that (463), divided by the total number of (456), of the north american chapter of (456), in order to be able to (422)

TABLE VII
TOP-20 NEW LEXICAL BUNDLES EXTRACTED FROM ACL-ARC. NUMBER OF OCCURRENCES IN PARENTHESES. CONFERENCE NAMES, NOT TO BE CONSIDERED AS LEXICAL BUNDLES, ARE UNDERLINED.

3-gram	based on the (26276), one of the (21129), the performance of (17239), we use the (14812), proceedings of the (14718), the fact that (14675), part of the (14026), there is a (13562), the training data (13373), due to the (12934), on the other (12831), according to the (12673), each of the (11775), the other hand (11437), can not be (11351), there is no (10564), in the training (10269), words in the (10253), the case of (9933), used in the (9902)
4-gram	the performance of the (6390), as well as the (6331), the size of the (6130), with respect to the (5834), is based on the (4771), for each of the (4167), in the training data (4106), the results of the (3904), the rest of the (3857), the fact that the (3850), the quality of the (3725), department of computer science (3508), in addition to the (3468), the association for computational (3291), the length of the (3168), annual meeting of the (3072), the output of the (2858), the end of the (2804), to the fact that (2739), the user s (2694)
5-gram	the association for computational linguistics (1831), on the basis of the (1673), at the end of the (1464), the best of our knowledge (1421), to the best of our (1419), the association for computational pages (1277), in the context of the (1237), in the case of the (1231), the anonymous reviewers for their (1220), the state of the art (1171), the paper is organized as (1145), the results are shown in (1110), in such a way that (1109), the rest of the paper (1104), we can see that the (962), of computer science university of (941), the remainder of this paper (906), in the form of a (877), of the words in the (876), • • • • • (858)
6-gram	to the best of our knowledge (1400), the paper is organized as follows (1128), the results are shown in table (788), are those of the authors and (751), department of computer science university of (750), the rest of the paper is (710), and do not necessarily reflect the (634), is due to the fact that (626), thank the anonymous reviewers for their (622), the authors would like to thank (619), • • • • • • (595), due to the fact that the (581), from the point of view of (578), of the authors and do not (553), those of the authors and do (549), the authors and do not necessarily (525), the defense advanced research projects agency (524), the number of words in the (521), this work was supported by the (508), the remainder of this paper is (507)

TABLE VIII
TOP-20 N-GRAMS CLASSIFIED AS FALSE BUNDLES. NUMBER OF OCCURRENCES IN PARENTHESES. FALSE NEGATIVES ARE UNDERLINED.