ICIC Express Letters
Part B: Applications
Volume **9**, Number **6**, June **2018**

ICIC International ©2018 ISSN 2185-2766
pp. 551–557

LINGUISTIC ANALYSIS BASED ON THE CORPORA OF ENGLISH ABSTRACTS

Minxuan Feng[1], Xiaoshuang Wu[1,*], Mengting Sun[1] and Beibei Xu[2]

[1]School of Chinese Language and Literature
[2]Zhongbei College
Nanjing Normal University
No. 1, Wenyuan Road, Qixia District, Nanjing 210046, P. R. China
{ fennel_2006; xbbbaby2007 }@163.com; *Corresponding author: mona_cry@hotmail.com
1151437559@qq.com

Received December 2017; accepted March 2018

Abstract. *This paper expounds the process of constructing an English abstract corpus. The scope of work is based on the annotation, statistics and analysis of 756 abstracts of the Association of Computational Linguistics Conference's long papers and 284 English abstracts of computational linguistics thesis of the Journal of Chinese Information Processing. A bilingual common term knowledge base with 269 entries, a bilingual common verb knowledge base with 116 entries and a common sentence patterns knowledge base with 325 entries were constructed with corpus tools. Furthermore, an English abstract writing software was designed to help non-native speakers of English to write more normative and idiomatic English abstracts.*

Keywords: English abstract, Language features, Knowledge base, Computational linguistics

1. **Introduction.** Computational linguistics is an emerging interlaced subject, and it is gaining more popularity in the past years. English abstracts play an important role in this academic area, as they condense the information of the research background, purpose, methodologies, results and application, and in [1] we can see the advantage of abstracts. In fact, abstracts have been studied in many specific fields, such as in [2-4], but it is not popular in computational linguistics.

English abstracts can be analyzed from many perspectives. As for the genre of the abstracts, there are three schools in this domain according to [5]: ESP (English for Specific Purposes), EAP (English for Academic Purposes), SFL (Systemic Functional Linguistics), this paper will adopt the theory from ESP represented by Swales and Bhatia, and the theory is also very popular in many studies. In [6,7], it is assumed that a genre comprises a class of communicative events, whose members share some sets of communicative purposes. Swales claimed that genres are the properties of discourse communities and genre analysis is able to reveal the specific cognitive structures from specific discourse, or to interpret discourse structures and communicative function. Bhatia's definition of genre highly accords with Swale. In [8,9], it is believed that genres typically serve socially recognized communicative purposes and that genres are exploited to convey private intentions. In fact, other theories may have different definitions of genre and different viewpoints of genre, but the essential meaning of genre would not differ among the three schools, as is mentioned in [10,11]. And it is common that researchers using the theory in [6], such as the seven moves are used in [12], and the theories proposed in ESP are used in [13].

The dominant theory of type analysis of English abstract is given in [14]. The author categorized them into descriptive abstract, informative abstract, and informative-descriptive abstract. Similarly, in [15], it recommended a classification method for scientific papers, which divided abstracts into informative abstract and descriptive abstract.

Language feature analysis is another perspective in abstract studies. The researchers are interested in verb tenses, verb voices and the first person pronouns in abstracts. For instance, in [16], it was found that there are always "past simple" and "passive" in English abstracts. However, in [17], it is noticed that the most frequent tense in abstracts of scientific papers is "present simple", but it will vary with the move in different scientific areas.

Researchers also analyze the basic information of an abstract, such as the length of an abstract. In [18], it is stated that the length of an abstract of a scientific paper ranges roughly from 80 words to 150 words. In [15], it is pointed out that the length of a scientific paper's abstract should be no more than 250 words. As a matter of fact, different periodicals have different limitations on the length of an English abstract. It is 200 words or so for ACL (Association of Computational Linguistics).

In applications, [19] presents a software system which can assist Japanese software engineers in abstract writing, [20] introduces a method for computational analysis of move structures in abstracts of research articles and [21] builds a CRF for identifying sections in abstracts.

As discussed above, abstracts can be studied from various perspectives. In this paper, we take an empirical approach to the study of abstracts by constructing a corpora annotated with the basic information, types, structures and language features. We collected 756 abstracts of the long papers published from 2010 to 2014 in the *Association of Computational Linguistics Conference* and 284 English abstracts from theses from the *Journal of Chinese Information Processing* that are published from 2010 to 2014. In Section 2, the types of abstracts and the meaning of each type are presented. Section 3 introduces the structure of abstracts, especially the theories from ESP. This section also introduces the annotation schemes of the structures. The language features used in annotation are given in Section 4. In Section 5, the elements introduced in Sections 2, 3 and 4, and some basic information are used to construct the corpus of English abstracts and the knowledge bases. Section 6 explains the three knowledge bases. Section 7 gives the conclusion of the paper.

2. **Type of Abstracts.** We use the theory proposed in [14,15], and the former advocates to classify abstracts into three types according to the functions of abstracts of academic articles: descriptive abstract, informative abstract, and informative-descriptive abstract. Similarly, the latter divided the abstracts of scientific articles into two types: descriptive abstract and informative abstract. Descriptive abstract generally gives a brief statement of the question, the method, the result and the conclusion. Informative abstract tells readers the method, the problem, the conclusion, and hardly ever gives specific statistics. Informative-descriptive abstract is the combination of the two types mentioned above.

However, we discovered two new types in the course of annotating the abstracts, the details are shown in Table 1 and Table 2. One is informative-argumentative abstract, which always gives arguments without providing the method or result about a problem. The other is questioning abstract, which is mostly the question and the answer. However, these two types are isolated cases, which will not be discussed in detail in this essay. And we noticed that the informative abstract is the dominant type.

3. **Structure of Abstracts.** The theories proposed in ESP are used in the present study, such as the IMRD (Introduction, Method, Result, Discussion) and the CARS (Create a Research Space) proposed in [6,7], and the IMRC (Introducing Purpose, Describing Methodology, Summarizing Results and Presenting Conclusions) proposed in [8,9]. In such theories, the core is a move, which is divided by specific communitive purposes, and may contain several steps.

Table 1. Type of the abstracts of long papers from ACL

	Frequency	Effective percentage	Cumulative percentage
descriptive abstract	3	0.4	0.4
questioning abstract	1	0.1	0.5
informative-descriptive abstract	33	4.4	4.9
informative-argumentative abstract	2	0.3	5.2
informative abstract	717	94.8	100.0
total	756	100.0	

Table 2. Type of the abstracts of from the *Journal of Chinese Information Processing*

	Frequency	Effective percentage	Cumulative percentage
descriptive abstract	7	2.5	2.5
informative-descriptive abstract	19	6.7	9.2
informative-argumentative abstract	2	0.7	9.9
informative abstract	256	90.1	100.0
total	284	100.0	

In IMRD and CARS models, the structure of an English abstract is divided into four moves: Introduction, Methodology, Result and Discussion. The move of introduction can be categorized into four steps: Background, Problem, Purpose, Theme. The move of discussion is categorized into four steps: Conclusion, Implication, Limitation, and Application.

It should be noted that the annotation of a move or a step is only applicable for an informative abstract or an informative-descriptive abstract. If descriptive abstracts, informative-argumentative abstracts or questioning abstracts need to be analyzed, we will label them with descriptive, argumentative and questioning separately.

The standard procedure of annotating the structure is as follows. First, read the English abstract and determine the move or the step of every sentence in the abstract; second, evaluate the move or the step that the sentence belongs to by distinguishing the key words or the obvious delimiters in every sentence and understanding the meaning of the sentence; third, use Excel to annotate the move or the step of every sentence from every abstract, and use SPSS for further statistical analysis.

The following is an illustration of the annotating procedure.

① We present an approach for automatically learning to solving algebra word problems. ② Our algorithm reasons across sentence boundaries to construct and solve a system of linear equations, while simultaneously recovering an alignment of the variables and numbers in these equations to the problem text. ③ The learning algorithm uses varied supervision, including either full equations or just the final answers. ④ We evaluate performance on a newly gathered corpus of algebra word problems, demonstrating that the system can correctly answer almost 70% of the questions in the dataset. ⑤ This is, to our knowledge, the first learning result for this task. (ACL, 2014, No.26)

The annotation is given as follows. As for sentence ①, we annotate this as a "Theme" step. In sentence ②, we can find the key words, so we call it a key phrase, which is "our algorithm" in this sentence. It shows us how the algorithm the author proposed is realized. This sentence is therefore annotated as a "Methodology" move. Sentence ③ also states algorithm as we can find the phrase "The learning algorithm". It is also annotated as the "Methodology" move. Sentence ④ states the results and the achievements of the research, which are denoted by key words such as "evaluate", "performance" and the

statistic "70%", so this sentence should be labeled "Result" move. Sentence ⑤ comes after a result move. It can be annotated as a "conclusion" step according to the meaning of the sentence.

Besides, we noticed that there is no necessary one-to-one match between a sentence and a move (or a step). It is also possible that several sentences (always 2 or 3) belong to the same move (or step), like sentence ② and sentence ③ in the example above.

After annotation, we discovered that the most common moves or steps in ACL abstracts are "background, problem, theme, theme + aim, methodology, result". However, for the abstracts from the *Journal of Chinese Information Processing*, there are minute differences, and the most common moves or steps are "background, problem, theme, methodology, result". The abstract from the journal is more likely to introduce the background and the theme, and the domestic researchers could pay more attention to the problem and the aim.

4. **Language Features of Abstracts.** With regard to language features of abstracts, we will present observations about verb tenses, verb voices and the first person pronouns.

Tense is a form of the verb. Different tenses could indicate different time and modes. We annotate four major tenses in our corpus: present simple, past simple, present perfect and past perfect. Besides, there are present progressive, simple future and some others. We note that the most frequent tense in ACL abstracts is "present simple", which is up to 93.6% in the 756 abstracts, the same for the abstracts from the *Journal of Chinese Information Processing*, which is up to 91.3%.

Voice is also a form of the verb. It tells the relation between the subject and the predicate verb. There are "active" and "passive" voices in the corpus. It is noted that "active" is the most common voice in both ACL and the journal afterwards, yet the percentages are slightly different.

In the corpus, we have two labels for the first person pronouns: we and I. And "we" is the most common one based on the statistical result, which is in accordance with previous studies. However, the abstracts from the *Journal of Chinese Information Processing*, "this paper, the paper, this article, this work" take almost half of the subjects used in abstracts.

5. **English Abstract Corpus.** The method of annotation is already presented in the above sections. This section gives a summary of the data, including title, nation, college or institution, the types of abstracts (descriptive abstract, informative abstract, informative-descriptive abstract, informative-argumentative abstract and questioning abstract), number of sentences (sentences were divided by full stops), number of words in abstract, the content of the abstract (break the abstract by full stops), the structure (move or step: Introduction – Background, Problem, Aim, Theme; Methodology; Result; Discussion: Conclusion, Implication, Limitation, Application) and the language features (verb tenses, verb voices and first person pronoun).

Figure 1 gives an example of such summary. This abstract is from a paper of the conference of ACL in 2014, and there are 13 entries of the abstract in total, such as ID, title, type, nationality of the author, institution, number of sentences, number of words, content, structure of each sentence, tense of each sentence, voice of each sentence, first person pronoun of each sentence and a note that noted the things we find special through the annotation.

6. **Knowledge Base for Abstract Writing.** Based on the corpus constructed in the last section, we constructed three knowledge bases to assist abstract writing: the knowledge base of commonly used bilingual terms (Figure 2 is section), the knowledge base of commonly used bilingual verbs (Figure 3 is section) and the knowledge base of commonly used sentence pattern (Figure 4 is section), which offers effectively help for Chinese to write English abstracts and promote international academic communication.

ID	Title	Type	Nationality of the Primary Author	College/ Institution	Number of Sentences	Number of Words	content	Structure(move)	tense	voice	first person pronoun	note
2014001	Learning Ensembles of Structured Prediction Rules	informative-descriptive	America	Google research	3	71	We present a series of algorithms with theoretical guarantees for learning accurate ensembles of several structured prediction rules for which no prior knowledge is assumed.	Theme+Aim	present simple	active	we	The result is reported in the paper, but that part is not reported in abstract.
							This includes a number of randomized and deterministic algorithms devised by converting on-line learning algorithms to batch ones, and a boostingstyle algorithm applicable in the context of structured prediction with a large number of labels.	methodology	present simple	active	we	
							We also report the results of extensive experiments with these algorithms.	descriptive	present simple	active	we	

FIGURE 1. The annotation of an abstract

CRF——Conditional Random Fields 条件随机场
CTB——Chinese Treebank 中文树库
HMM——Hidden Markov Model 隐马尔科夫模型
IE——Information Extraction 信息提取
ILP——Integer Linear Programming 整数线性规划
IR——Information Retrieval 信息检索
KB——Knowledge Base 知识库
ML——Machine Learning 机器学习
MT——Machine Translation 机器翻译
NER——Named Entity Recognition 命名实体识别

FIGURE 2. The example of abbreviated terms

address 解决	describe 介绍	focus on 将目光投向	lack 缺乏
analyze 分析	determine 决定	hamper 阻碍	learn 学习
build 构建	explore 探索	ignore 忽视	limit 限制
compare 比较	face 面临	introduce 介绍	propose 提出
demonstrate 证明	fail to 未能	investigate 研究	present 提出

FIGURE 3. The example of ordinary verbs from introduction move

1. A variety of …… algorithms have been applied to ……., but often …….
2. An important …… task in the …… is to …….
3. Existing research has studied …… on certain issues, opposing …….
4. Previous work has investigated the ……., essentially treating them as …….
5. Previous work has shown that ……., where the …… is encouraged, but not require to ……., can substantially improve …… performance.

FIGURE 4. The example of sentence pattern templates from introduction move

First, we choose the standard abstracts from the ACL abstracts that we annotated, and then we use a software (AntConc) to grab the terms and verbs. As for the sentence pattern, it is not suitable to use machine learning models, so manual selecting is used.

We obtained 269 commonly used terms (including 31 abbreviated terms, 179 nominal terms and 59 phrase terms), 116 commonly used ordinary verbs (31 in introduction move, 81 in methodology move, 22 in result move and 9 in discussion move, it is noticed that a verb may exist in several moves), 325 sentence pattern templates (109 in introduction move, 132 in methodology move, 70 in result move and 14 in discussion move).

7. **Conclusion.** We took 1040 abstracts as research object, constructing them as two corpora (in Section 5) with annotation (in Sections 2, 3, 4), such as type, structure, language feature and some basic information. And based on the two corpora we construct three knowledge bases (in Section 6) for writing assistance.

There are three main types, and we found two new types and the type of an abstract might better be informative after we got the corpora. And in structure part, we have four moves, it appears that introduction, methodology and result are the necessary move in the structure of an abstract, and we should pay attention to background, theme, problem and aim in the introduction move, especially the background. As for the language features, finally we got that the present simple and active verb should be held essential, and it is appropriate to use "we" as the first person pronoun in an abstract. And about the basic information, we found out that the length of an abstract should be within 4-6 sentences or 90-140 words.

There are also research results that are not presented here due to limited space such as the basic information, the mistakes in these abstracts, and so did the software we designed which was based on the knowledge bases we constructed, this software contains the bases that mentioned earlier and some summaries we made. We hope it could help Chinese to write English abstracts and promote international academic communication. And we wish the findings could help non-native speakers to write a more normative English abstract for computational linguistics thesis and promote the impact of Chinese linguistics and technological achievements at the international level. Besides, we expect that our knowledge bases could be adopted in automated essay scoring system, such as PEG (Project Essay Grade), IEA (Intelligent Essay Assessor) and E-rater (Electronic Essay Rater). And in future, we hope that we could use more methods to gain the terms as the method we take now is a bit subjective and we would like to collect more abstracts of ACL articles to enlarge the bases.

Acknowledgment. This work is partially supported by "A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the National Languages Committee (YB135-23) and the project for Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Sciences (2017STD006). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] A. Letchford, T. Preis and H. S. Moat, The advantage of simple paper abstracts, *Journal of Informetrics*, vol.10, no.1, pp.1-8, 2016.

[2] B. Behnam and F. Golpour, A genre analysis of English and Iranian research articles abstracts in applied linguistics and mathematics, *International Journal of Applied Linguistics & English Literature*, vol.3, no.5, pp.173-179, 2014.

[3] M. Spiroski, Analysis of abstracts from the medical theses written in Macedonian language and proposal of standards for abstract preparation, *Macedonian Journal of Medical Sciences*, vol.7, no.1, pp.5-10, 2014.

[4] S. Can, E. Karabacak and J. Qin, Structure of moves in research article abstracts in applied linguistics, *Publications*, vol.4, no.3, p.23, 2016.

[5] S. Hyon, Genre in three traditions: Implications for ESL, *TESOL Quarterly*, 1996.

[6] J. Swales, *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press, Cambridge, 1990.

[7] J. Swales, *Genre Analysis – English in Academic and Research Settings*, Shanghai Foreign Language Educational Press, Shanghai, 2001.
[8] V. K. Bhatia, *Analysing Genre: Language Use in Professional Settings*, Longman, London & New York, 1993.
[9] V. K. Bhatia, Genre-mixing in academic introductions, *English for Specific Purposes*, vol.16, no.3, pp.181-195, 1997.
[10] H. Kay and T. Dudley-Evans, Genre: What teachers think, *ELT Journal*, 1998.
[11] T. Dudley-Evans and M. J. St. John, *Development in ESP: A Multi-Disciplinary Approach*, Cambridge University Press, Cambridge, 1998.
[12] H. P. Liu, Linguistic features for paper abstract of international retrieval journal – Case study of SCI and SSCI journal article, *Journal of Chang'an University*, 2012.
[13] H. Marefat and S. Mohammadzadeh, Genre analysis of literature research article abstracts: A cross-linguistic, cross-cultural study, *Applied Research on English Language*, 2013.
[14] E. T. Cremmins, *The Art of Abstracting*, ISI Press, Philadelphia, 1982.
[15] R. Day and B. Gastel, *How to Write and Publish a Scientific Paper*, Cambridge University Press, Cambridge, 2012.
[16] B. Melander, J. M. Swales and K. M. Fredrickson, Journal abstracts from three academic fields in the United States and Sweden: National or disciplinary proclivities?, in *Intellectual Styles and Cross-Cultural Communication*, A. Duszak (ed.), Berlin, Mouton De Gruyter, 1997.
[17] D. Ge and R. Yang, Genre analysis of academic paper abstracts, *Modern Foreign Languages Quarterly*, 2005.
[18] H. Glasman-Deal, *Science Research Writing for Non-Native Speakers of English*, Imperial College Press, 2010.
[19] M. Narita, Constructing a tagged EJ parallel corpus for assisting Japanese software engineers in writing English abstracts, *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pp.1187-1191, 2000.
[20] J. C. Wu, Y. C. Chang, H. C. Liou and J. S. Chang, Computational analysis of move structures in academic abstracts, *COLING/ACL on Interactive Presentation Sessions*, pp.41-44, 2006.
[21] K. Hirohata, N. Okazaki and S. Ananiadou, Identifying sections in scientific abstracts using conditional random fields, *Proc. of the IJCNLP*, 2008.