Extraction of Key Expressions Indicating the Important Sentence from Article Abstracts

Shuhei Otani
Department of Library Science
Kyushu University
Fukuoka, Japan
otani.shuhei.874@m.kyushu-u.ac.jp

Yoichi Tomiura
Department of informatics
Kyushu University
Fukuoka, Japan

***Abstract*—In this study, we aim to extract key expressions that indicate the important sentence describing the originalities or contributions from article abstracts. The expense of searching academic information increases because of increases in the number of articles, discipline subdivisions, and promotion of interdisciplinary research. Improving the extraction and presentation of the main points from article abstracts will contribute to reducing academic information search expenses. We extracted pseudo-important sentences from each article abstract based on the ratio of the number of words in the identified sentence that appear in the article title to the number of all words in the sentence. After that, we evaluated the ratio of the number of the pseudo-important sentences including each N-gram to the number of all sentences including that N-gram. We then extracted the N-grams with a ratio as high as the key expressions.**

Keywords—Information retrieval of articles; Support of Information retrieval; Key expression; Natural language processing

I. INTRODUCTION

Academic research information continues to be increased by contributions from researchers in developing countries, including China. Researchers face an increase in search results commensurate with the increase in academic research information. Besides, disciplines have become subdivided and interdisciplinary research is promoted by nations. In this environment, information searches that are limited or outside of the researcher's discipline occur. Academic information searches incur large expenses because the researcher may not have enough knowledge of the technical term in that discipline.

In this study, we support the improvement of such an academic information search. Specifically, we extract characteristic expressions from article abstracts (e.g., 'In this paper' or 'We suggest that'; henceforth, we call them key expressions) such that important sentences including these expressions describing originalities or contributions are identified. We finally aim to extract the important parts of sentences estimated to be important using the extracted key expressions. This makes it possible to present search results that emphasize the important parts of article abstracts and recommend keywords to be used as query terms. This process can contribute to reducing the cost of academic information searches.

II. RELATED WORK

Jonnalagadda et al. studied system construction to present important parts of structured abstracts using a medical thesaurus to identify information in the literature for the information needs of the clinician [1]. Structured abstracts are often used in the medical field, but are not used in many other fields.

Nakawatase and Oyama extracted important sentences from article abstracts written in Japanese [2]. At first, they made a list of key expressions such that sentences including them are important for their respective abstracts. Then they extracted important sentences using key expressions. To extract the important sentences that included no key expressions, they made a list of predicates to indicate the important sentences. In this method, the more key expressions that are prepared, the higher the accuracy of extracting important sentences. However, it is expensive to prepare many key expressions. Studies on structuring abstracts [3] and text summarization [4] are related to the extraction of important sentences. The methods used in both techniques with high performance are based on machine learning. Therefore, the cost of preparing training data is a problem. To distinguish an important part from an unimportant part in abstracts is enough to support an academic information search.

III. METHOD FOR EXTRACTION OF KEY EXPRESSIONS

A sentence that has many words in common with an article title is more likely to be an important sentence. We call the sentence that describes the originalities or contributions in an article abstract the *important sentence*. In the experiment described later, 60% of sentences that have many words in common with the title were important sentences. In this study, we call a sentence in an article abstract that has many words in common with the title a *pseudo-important sentence*. For each N-gram (N-word sequence) ω extracted from pseudo-important sentences, we evaluate the ratio of the number of pseudo-important sentences including ω to the number of all sentences including ω. The proposed method is to extract N-grams with a ratio of more than threshold values as key expressions. If a

sentence includes some of the extracted key expressions, the sentence is determined to be an important sentence.

Technical terms specific to fields are expected to appear not only in important sentences, but also in unimportant sentences. Therefore, we can remove such technical terms based on the word's distribution in the abstract.

The details of the procedure for extracting the key expressions are as follows:

(a) We extract pseudo-important sentences in each article A. Let $W_T(A)$ be a set of words that appear in the title of the article A, and $W(s)$ be a set of words that appear in a sentence s in the abstract of the article A. If the following condition is satisfied, we extract the sentence s as a pseudo-important sentence.

$$\frac{|W_T(A) \cap W(s)|}{|W_T(A) \cup W(s)|} \geq \alpha \quad (1)$$

Here, $|S|$ means the number of elements in set S.

(b) We make the set C of N-grams that appear in pseudo-important sentences extracted in step (a). Let $f_I(\omega)$ be the number of pseudo-important sentences including N-gram ω, and $f_{ALL}(\omega)$ be the number of sentences including ω.

(c) We remove the N-gram ω from C such that $f_I(\omega)$ is not in the top β of $f_I(\omega)$ or $f_I(\omega)$ is less than γ. β *and* γ are threshold values.

(d) We remove the N-gram ω from C such that ω does not meet the following condition.

$$\frac{f_I(\omega)}{f_{ALL}(\omega)} \geq \delta \quad (2)$$

This is to remove technical terms and general expressions.

We regard the remaining elements in C as key expressions.

Several articles have titles that do not describe their contents. For example, there is an article titled 'Finding Scientific Topics'. This article performed topic analysis using Latent dirichlet allocation based on Gibbs sampling, but a tangible method or originality cannot be identified from this title. In addition, there are articles that have a rhetoric title. For those articles, we cannot extract important sentence candidates based on the ratio of the number of words in the sentence that appear in the article title to the number of all the words in the sentence. However, we can extract an important sentence candidate using the extracted key expressions.

IV. Experiments

A. Experimental data

We used article abstracts in journals that were classified in 'Computer Science' or 'Analytical Chemistry' as experimental data. They were extracted from the academic information database 'Scopus' provided by Elsevier B.V.

B. Parameter

α, β, and γ were fixed as follows: α=0.3, β=500, and γ=10. δ was changed in the range from 0.1 to 0.5. N was changed in the range from 2 to 4.

α was decided from the result of the preliminary experiment. The precision of determining important sentences was highest when α was 0.3 with a precision of 60.5%. The precision was not improved when we removed stop words or narrowed words to nouns. Then we carried out the experiment described later in a similar way.

β and γ were defined using the following method. We assumed that the number of key expressions was around 300. Therefore, we expected that it was necessary to extract around 500 expressions. In addition, the value of the left-hand side in equation (2) is devoid of reliability for expression of ω with a frequency of once or twice. We could confirm that N-grams such as 'In this paper, we' were included in the set C at the step (c), when we set β=500 and γ=10. We used fixed values (β=500, γ=10) in this experiment. When we build a support system for the academic information search, it will be necessary to find suitable values experimentally.

In this study, we extract the set of key expressions for each pair of values of parameters δ and N, and evaluate the extracted set of key expressions by determining important sentences from abstracts using these parameters.

C. Evaluation data

We used 115 article abstracts as evaluation data in journals that were classified in 'Computer Science' or 'Analytical Chemistry' from Scopus. These article abstracts were different from the experimental data of 10,000 article abstracts. We determined whether each sentence in 115 article abstracts was an important sentence manually. Let these data with the importance label be evaluation data 1.

We removed sentences that satisfied condition equation (1) from evaluation data 1. Let these data be evaluation data 2. By evaluation data 2, we evaluate whether we can extract important sentences from sentences that have less words in common with the title. TABLE I shows the summary of evaluation data 1 and 2.

D. Result

We changed δ=0.1, 0.3, 0.5, and N=2, 3, 4, and extracted key expressions. We show the results by evaluation data 1 in TABLE II and by evaluation data 2 in TABLE III.

At first, we used evaluation data 1. Precision was increased and recall was decreased with increasing values of δ and N. When we set δ=0.5 and N=4, precision was the highest, and the value was 88.2%. At that time, recall was 7.6%. When we set

TABLE I. Evaluation data

	Evaluation data 1	**Evaluation data 2**
Number of sentences	958	916
Number of important sentences	198	171

δ=0.1 and N=2, recall was the highest and the value was 70.0%. At that time, precision was 39.0%. We assumed that a sentence that satisfied condition equation (1) was a pseudo-important sentence. At the preliminary experiment, precision mining of important sentences using this condition was 60.0%. The precision of determining important sentences using the extracted key expressions was more than 60.0% at six of the nine parameter settings.

When we used evaluation data 2, precision and recall for each condition decreased several percent in comparison with using evaluation data 1 in TABLE IV, but the difference is not remarkable. Precisions were more than 60.0% at four of the nine parameter settings.

Therefore, we think that determining important sentences based on the key expressions extracted by the proposed method is effective in respect of precision. However, recall was extremely low, and this problem will need to be improved. We expect that 50% precision of determining important sentences is enough, because extracted important parts are finally confirmed by the user. We will return to this problem in Section V.

TABLE II. EVALUATION 1

N	*δ*	Number of expressions	Precision (percent)	Recall (percent)
2	0.1	332	39.0	70.2
2	0.3	75	55. 6	17.7
2	0.5	11	72.7	4.0
3	0.1	220	52.6	40.9
3	0.3	110	64. 2	17.2
3	0.5	37	85.7	9.1
4	0.1	90	69.8	18.7
4	0.3	60	80.7	12.6
4	0.5	35	88.2	7.6

TABLE III. EVALUATION 2

N	*δ*	Number of expressions	Precision (percent)	Recall (percent)
2	0.1	332	35.9	66.7
2	0.3	75	46.9	13.5
2	0.5	11	40.0	1.2
3	0.1	220	48.1	36.8
3	0.3	110	56.1	13.5
3	0.5	37	75.0	5.3
4	0.1	90	68.3	16.4
4	0.3	60	77.3	9.9
4	0.5	35	77.8	4.1

TABLE IV. COMPARISON OF EVALUATION 1 WITH EVALUATION 2

Parameter	Precision (evaluation .1)	Precision (evaluation.2)
N=2, δ =0.1	39	35.9
N=2, δ =0.3	55.6	46.9
N=2, δ =0.5	72.7	40.0
N=3, δ =0.1	52.6	48.1
N=3, δ =0.3	64.2	56.1
N=3, δ =0.5	85.7	75
N=4, δ =0.1	69.8	68.3
N=4, δ =0.3	80.7	77.3
N=4, δ =0.5	88.2	77.8

E. Extracted key expressions

In each condition, TABLES V, VI, and VII show parts of the extracted key expressions such that the precision of determining important sentences using them is high.

When we set δ=0.1 or 0.3, we could extract expressions such as (‘in', ‘this', ‘study', ‘a‘) and (‘scheme’, ‘based’, ‘on‘), which were expected key expressions. On the other hand, expressions such as (‘atomic’, ‘absorption’) and (‘and‘, ‘rapid’) were extracted too. Though we succeeded in determining an important sentence using those key expressions, it was sheer luck. When we set δ=0.5, we could extract some useful expressions such as (‘method’, ‘was’, ‘developed’, ‘for’), but we cannot extract many useful key expressions such as (‘in’, ‘this’, ‘study’, ‘a‘). Technical terms specific to a particular field such as (‘liquid’, ‘chromatography-tandem’,

TABLE V. EXTRACTED KEY EXPRESSIONS (δ=0.1)

2-gram	3-gram	4-gram
(‘based’, ‘on’)	(‘a’, ‘sensitive’, ‘and’)	(‘in’, ‘this’, ‘study’, ‘a’)
(‘for’, ‘determining’)	(‘based’, ‘on’, ‘the’)	(‘for’, ‘the’, ‘first’, ‘time’)
(‘highly’, ‘sensitive’)	(‘in’, ‘human’, ‘plasma’)	(‘method’, ‘was’, ‘successfully’, ‘applied’)
(‘analysis’, ‘of’)	(‘internal’, ‘standard’, ‘is’)	(‘method’, ‘was’, ‘successfully’, ‘applied’)
(‘atomic’, ‘absorption’)	(‘scheme’, ‘based’, ‘on’)	(‘proposed’, ‘in’, ‘this’, ‘paper’)

TABLE VI. EXTRACTED KEY EXPRESSIONS (δ=0.3)

2-gram	3-gram	4-gram
(‘and’, ‘rapid’)	(‘developed’, ‘and’, ‘validated’)	(‘this’, ‘study’, ‘was’, ‘to’)
(‘determination’, ‘of’)	(‘method’, ‘has’, ‘been’)	(‘was’, ‘developed’, ‘for’, ‘the’)
(‘problem’, ‘with’)	(‘for’, ‘the’, ‘quantitative’)	(‘high’, ‘performance’, ‘liquid’, ‘chromatography’)
(‘simultaneous’, ‘determination’)	(‘is’, ‘proposed’, ‘in’)	(‘in’, ‘this’, ‘study’, ‘a’)
(‘we’, ‘investigated’)	(‘study’, ‘was’, ‘to’)	(‘method’, ‘was’, ‘developed’, ‘for’)

TABLE VII. EXTRACTED KEY EXPRESSIONS (δ=0.5)

2-gram	3-gram	4-gram
(‘spectrophotometric’, ‘method’)	(‘been’, ‘developed’, ‘for’)	(‘a’, ‘simple’, ‘and’, ‘sensitive’)
(‘for’, ‘simultaneous’)	(‘simultaneous’, ‘determination’, ‘of’)	(‘developed’, ‘and’, ‘validated’, ‘for’)
(‘spectrometry’, ‘method’)	(‘and’, ‘validated’, ‘for’)	(‘method’, ‘was’, ‘developed’, ‘for’)
(‘chromatography-tandem’, ‘mass’)	(‘capillary’, ‘zone’, ‘electrophoresis’)	(‘validated’, ‘for’, ‘the’, ‘simultaneous’)
(‘liquid’, ‘chromatography-tandem’)	(‘in’, ‘human’, ‘plasma’)	(‘liquid’, ‘chromatography-tandem’, ‘mass’, ‘spectrometry’)

‘mass’, ‘spectrometry’) were extracted as key expressions. The key expressions (‘liquid’, ‘chromatography-tandem’, ‘mass’, ‘spectrometry’) were used in combination with expressions such as (ADJECTIVE, ‘method’, ‘of’). We would rather extract N-grams such as (ADJECTIVE, ‘method’, ‘of’) as key expressions than (‘liquid’, ‘chromatography-tandem’, ‘mass’, ‘spectrometry’), but ADJECTIVE words were various, e.g., ‘new’, ‘sensitive’, or ‘validated’. Each expression’s frequency was low. These expressions did not satisfy the condition (2).

V. DISCUSSION

We set equation (2) as one of the conditions for key expressions so that we could remove technical terms and general expressions from candidates. When we set δ=0.5, the precision of determining important sentences using the extracted key expressions was around 70–80%. However, recall was very low. One method for precision is to remove expressions such as 'in this paper, we show', which are expected key expressions under the above conditions. When we set δ=0.1, recall of important sentences was increased, but at the same time, precision decreased to 39–69%. These two improvements can keep the precision high and increase the search accuracy of recalled information.

The ratio of pseudo-important sentences that were actually important sentences was 60%. Pseudo-important sentences extracted in step (a) in the proposed method are only a small part of all important sentences as evaluation data 1 and 2 show. We expect that we can keep the precision high and increase the search accuracy of recalled information by increasing pseudo-important sentences using the extracted key expressions. At first we extract key expressions by the method described in Section III where we set δ = 0.5, e.g., so that the precision of determining important sentences is high. Next, we add sentences that were determined as important sentences using the extracted key expressions into pseudo-important sentences. Then we execute steps (b), (c), and (d) again to extract key expressions.

The other is to increase the fields of the article. The reason we set the condition equation (2) was to remove technical terms from candidates, but when we set δ=0.5, we extracted technical terms such as (‘liquid’, ‘chromatography-tandem’, ‘mass’, ‘spectrometry’) as key expressions. In this experiment, we used the two fields, 'Computer Science' and 'Analytical Chemistry'. The appearance frequencies of technical terms specific to the field were expected to decrease relatively when we increased the intended fields. Thus, we expect that we can remove technical terms by increasing fields without the condition of equation (2). In this case, we have to pay attention to the selection of fields. Of course, technical terms in 'Information Engineering' are different from technical terms in 'Humanities'. However, the difference is not only in technical terms, but also in the culture of academic writing, e.g., structure of abstracts and expressions. Therefore, the frequencies of key expressions to be extracted were also relatively decreasing. To avoid this problem, we should select fields where technical terms are different, but the culture of academic writing does not show substantial differences between fields, e.g., 'Natural Language Processing' and 'Computer Networks'.

VI. CONCLUSION

In this study, we aimed to extract key expressions that indicate the important sentence from article abstracts. We performed an experiment with the following method. We regard a sentence that shares many words with the article title as a pseudo-important sentence. We extracted key expressions from article abstracts based on the ratio of the number of the pseudo-important sentences and include them with the number of all sentences including these key expressions. We showed that the precision of determining an important sentence using the extracted key expressions was high, but that low information recall is a problem. In future research, we will attempt the two improvements described in the Discussion.

REFERENCES

[1] S. R. Jonnalagadda, G. Del Fiol, R. Medlin, C. Weir, M. Fiszman, J. Mostafa, and H. Liu. "Automatically Extracting Sentences from Medline Citations to Support Clinicians’ Information Needs." *Journal of the American Medical Informatics Association* 20, no. 5 (2013): 995-1000.

[2] H. Nakawatase, K. Oyama. Extraction of Sentences that Summarize the Main Point of Research Abstract[in Japanese]. The Japanese Society for Artificial Intelligence Conference Paper（Information Compilation 6th）, No.TETDM-01-SIG-IC-06-03, pp.13-16 (2011)

[3] S. Teufel. The structure of scientific articles: applications to citation indexing and summarization. Center for the Study of Language and Information, 2010, xii, 518 p.p.

[4] E. Lloret. and M. Palomar. "Text Summarisation in Progress: A Literature Review." *Artificial Intelligence Review* 37, no. 1 (2012): 1-41