Measurement Evaluation of Keyword Extraction Based on Topic Coverage

Ryosuke Saga, Hiroshi Kobayashi, Takao Miyamoto, and Hiroshi Tsuji

Graduate School of Engineering, Osaka Prefecture University
1-1, Gakuen-cho, Naka-ku, Sakai-shi, Osaka, Japan
{saga,tsuji}@cs.osakafu-u.ac.jp,
kobayahi@mis.cs.osakafu-u.ac.jp,
aki@kis.osakafu-u.ac.jp

Abstract. This paper proposes a method to measure the performance of keyword extraction based on topic coverage. The answer set of a keyword is required to evaluate keyword extraction by methods such as TF-IDF. However, creating an answer set for a large document is expensive. Thus, this paper proposes a new measurement called topic coverage on the basis of the assumption that the keywords extracted by a superior method can express the topic information efficiently. The experiment using the proceedings of a conference shows the feasibility of our proposed method.

Keywords: keyword extraction, performance evaluation, recall, precision, topic.

1 Introduction

Computerized documents are continuously being generated and used because of the progress of information technology. Furthermore, news articles and magazines from publishers, as well as blogs and tweets in social media, are generated by users daily, and the document set consists of several topics under different genres.

Keyword extraction is one of the most important techniques for comprehending the topics in a document set. Given that these keywords often express topics, we can analogize topics from keywords. The extracted keywords are used not only to determine the topic of documents but also to generate queries for a document set (often called a corpus). Hence, keyword extraction is important in treating a document set.

Keyword extraction can be conducted by several methods, and keyword extraction methods are categorized as supervised or unsupervised. An example of supervised keyword extraction is SVM, which is used for extracting keywords from an answer set [1]. An example of unsupervised keyword extraction is TF-IDF, which uses word frequency in documents [2].

We may choose a keyword extraction method on the basis of performance evaluation. The performance evaluation of a method is based on how well the method extracts keywords or how much the extracted keywords include the answer set. Precision and recall are developed on the basis of these criteria. However, these criteria

C. Stephanidis (Ed.): HCII 2014 Posters, Part I, CCIS 434, pp. 224–227, 2014.

assume that a manually created answer set exists; that is, an answer set needs to be generated for a new document set when evaluating a keyword extraction method. Such tasks require considerable time and effort, particularly for a large document set, such as tweets in social media (i.e., big data). The Mechanical Turk method focuses on creating answer sets; however, this method requires considerable time and resources [4].

Therefore, this paper proposes a measurement, namely, *topic coverage*, to evaluate the performance of a keyword extraction method without an answer set. This measurement is similar to recall except that the proposed method considers the topic.

2 Evaluation of Keyword Extraction Based on Topic Coverage

2.1 Principle and Assumption

Keyword extraction methods are normally evaluated by precision and recall, which have been introduced by several representative books. These methods are based on the contingence table shown in Table 1. For example, precision is calculated as A/(A + B) with the notations in the table, that is, these methods are derived from the correctness and convergence with an answer set assigned by humans.

The proposed method is based on the assumption that *keywords extracted by a superior method can express the topic information efficiently. A better method corresponds to higher topic coverage*. The adequacy of the idea can be proper because keyword extraction itself is used for understanding topics in a document set.

Table 1. Contingence table for recall and precision

	Keywords assigned by humans	Non-keywords assigned by human
Keywords extracted by a method	A	C
Non-keywords extracted by a method	B	D

Thus, if topic coverage is one, the keyword extraction method will cover all topics. If topic coverage is zero, the keyword extraction method will fail to extract keywords from the topic. To implement the proposed method, topic coverage *TC* is defined as follows:

$$TC = \frac{1}{|T|}\sum_{i \in T} \frac{|E_i \cap M_i|}{|E_i|}, \tag{1}$$

where $|E|$ shows the number of elements of set E, T is the set of topics in the document sets, and E_i shows a set of top j keywords in topic i. For convenience, E_i is called topic keywords in this paper. M_i is a set of top k keywords in topic i extracted by a certain method, such as TF-IDF and RIDF.

2.2 Evaluation of Extracted Keyword

We conducted the following steps to evaluate the performance of the proposed keyword extraction method.

Topic Extraction. The first step in evaluating the performance of a keyword extraction method is topic extraction. In this research, we assumed that each document belongs to a topic and does not belong to more than two topics different from a multi-topic model [5]. Thus, we can employ a clustering method to extract topics. Clustering methods include the k-means, k-medoids, and Girvan–Newman algorithm [6][7][8].

Keyword Extraction from Topics. The second step is extracting topic keywords. For example, we can use the TF-IDF index [5] to extract keywords as follows:

$$TFIDF_{ij} = TF_{ij} \log \frac{N_t}{DF_i}, \tag{2}$$

where N_t is the number of documents in topic t, TF_{ij} is the frequency of term i in document j, and DF_j is the document frequency, which is calculated as the number of documents with term i. The j extracted keywords correspond to E_j in Equation (1).

Keyword Extraction by an Evaluated Method and Measurement Calculation. The third step involves extracting keywords that correspond to M_i by using an evaluated method. Hence, k keywords are extracted from each topic i. *Topic coverage* can then be calculated by using M_i and E_i.

3 Experiment

We conducted an experiment to confirm the feasibility and characteristics of the topic coverage. We used 2008 and 2009 NIPS corpuses in this experiment. To verify the proposed measurement, we compared the value of the *topic coverage* with the keywords of abstract in each paper through the correlation between the topic coverage and recall of corpus. If *topic coverage* can be correlated with recall through an answer set, *topic coverage* is useful because it does not require an answer set. The experiment was conducted as follows:

1. N keywords were extracted, and recall for the keywords was calculated.
2. The topic coverage for N keywords in the first step and the correlation between recall and topic coverage were calculated.

Table 2. Correlation between Topic coverage and Recall

# of Topics	Correlation (NIPS in 2007)	Correlation (NIPS in 2008)
7	0.924	0.923
10	0.950	0.933
13	0.978	0.935

We set the number of topics to 7, 10, and 13 in this experiment and calculated the correlations for each topic size to confirm the robustness of this method.

The result of the experiment is shown in Table 2, which shows that topic coverage and recall have high correlation over 0.90. Thus, topic coverage may be used instead of recall.

4 Conclusion

This paper has proposed a new measurement method, namely, topic coverage, to evaluate keyword extraction performance without the use of an answer set. From the experiment, we confirmed that topic coverage and recall have high correlation. However, Keyword extraction from topics depends on the keyword extraction method employed. Thus, the use of preset or prepared keywords as topic keywords, such as in the paper "Keyword Extraction and Performance Evaluation," is better than extracting keywords by a certain method. In the future, the feasibility of this measurement on other corpuses, such as newspaper articles, and its correlation with other measurement will be verified.

Acknowledgement. This research was supported by JSPS KAKENHI Grant Numbers 13370017, 25420448, 23760358.

References

1. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
2. Zhang, K., Xu, H., Tang, J., Li, J.: Keyword Extraction Using Support Vector Machine. In: Yu, J.X., Kitsuregawa, M., Leong, H.-V. (eds.) WAIM 2006. LNCS, vol. 4016, pp. 85–96. Springer, Heidelberg (2006)
3. Salton, G.: Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer. Addison-Wesley Publisher (1988)
4. New York Times: Artificial Intelligence, With Help From the Humans (2007), http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html (accessed in March 2013)
5. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120 (2006)
6. Jayabharathy, J., Kanmani, S., Parveen, A.A.: A Survey of Document Clustering Algorithms with Topic Discovery. Journal of Computing 3, 21–28 (2011)
7. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann (2011)
8. Newman, M.E.: Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems 38, 321–330 (2004)