Clustering of Research Papers based on Sentence Roles

Satoshi FUKUDA † Yoichi TOMIURA †

† Graduate School of Information Science and Electrical Engineering, Kyushu University
{s.fukuda, tom}@inf.kyushu-u.ac.jp

Keywords: Sentence role, Clustering, Research paper

1 Introduction

In an academic paper search, particularly a search to confirm the originality of a user's research and to create survey articles, it is important that the search returns comprehensive results related to the user's information need. [1] proposes a method for efficiently selecting relevant research papers from a vast abstract set, which is based on a topic model and search formula created by the user. In this paper, we construct a system that visually expresses categories included in the reduced set of papers from [1], using a clustering based on the user's information need and selecting clusters having relevant papers. To generate the clusters based on the user's information need, it is important to know which structures (such as "background" and "method") in the abstract are relevant to the information need. This is based on the knowledge that if a user searches the papers related to the automatic construction of a thesaurus, he/she will judge whether a paper is relevant from sentences in the abstract describing the research purpose and method. We therefore propose a method using only the sentence content that matches the information need in the clustering.

2 Method

Our method consists of five steps: (1) a user selects role(s) related to the information need; (2) the system extracts sentences playing selected roles in the abstracts; (3) the system vectorizes the sentences extracted in (2) using Doc2Vec; (4) the system clusters vectorized sentences using k-means; (5) the system extracts top-100 phrases from the sentence set in the cluster using C-value [2] and presents clusters with these phrases to the user.

In (2), we construct classifiers for estimating the sentence roles used by the Support Vector Machine (SVM) from the dataset created by [3]. This dataset contains 1,000 abstracts for which three subjects manually judged whether each sentence has the "background," "purpose," "method" or "result" roles. For the construction of each classifier, we used each word as the feature, appearance information (1 if the word appears in the abstract, 0 if not) as the weight for each feature and a second-order polynomial kernel. In addition, we adjusted the number of subjects made positively assessment in order to show the highest performance by SVM in each role. A sentence was positively coded for "background" or "purpose" roles if two or more subjects made this assessment. A sentence was positively coded for "method" and "result" roles if one or more subjects assessed it as such. The recall and precision of the

different classifiers were, respectively, as follows: “background”: 0.863 and 0.863, “purpose”: 0.723 and 0.814, “method”: 0.789 and 0.801, and “result”: 0.737 and 0.821. These were calculated using micro average and 5-fold cross validation. We allow a sentence to be given multiple roles according to [3]. In (4), we set the average number of abstracts for one cluster to 30, 50, 70 and 100. This enables selection of an interpretable clustering result for the user. In (5), the C-value is from the noun phrases of 2 to 8-grams in the sentence set within each cluster.

3 Experimentations

We conducted an experiment on the NTCIR-1 and 2 datasets [4, 5]. These datasets contain 132 search tasks that describe the conditions of research papers that satisfy a particular information need, and Japanese research papers that are rated relevant or non-relevant for each task. We used 8 search tasks acquiring papers proposing new methods for a research. For examples of the used tasks are “new structural analysis of compound nouns” and “automatic construction of thesaurus”. Therefore, in step (1), “purpose” and “method” were selected. We tested our method using the annotated abstracts narrowed down by [1] with the target recall set to 1.000. As training data for the Doc2Vec, we used 444,475 Japanese abstracts in NTCIR datasets. The dimension number and window size were set to 400 and 8, respectively. We used all sentences in an abstract in the clustering as a baseline method and compared both methods by one subject. For the evaluation, we measured select-recall as (number of clusters including relevant abstract(s) selected by the subject) / (number of clusters including relevant abstract(s)), and select-precision as (number of clusters including relevant abstract(s) selected by the subject) / (number of clusters selected by the subject) from the viewpoint of how much the subject could select clusters including relevant abstract(s).

In the experimental results, our method showed 0.448 of recall and 0.864 of precision, while the baseline method showed 0.360 of recall and 0.896 of precision, calculated from the macro average of the 8 tasks, and the select-recall was significantly improved while maintaining a high precision. In the future, we will consider optimal parameter selection for Doc2Vec and further information presentation method to help users to interpret to clusters other than C-value.

References

1. Fukuda, S. et al.: Using Topic Analysis Techniques to Support Comprehensive Research Paper Searches. In: IALP (2017).
2. Frantzi, K. et al.: Extracting Nested Collocations. In: COLING, pp. 41-46 (1996).
3. Yamamoto, T. et al.: Constructing Corpus of Scientific Abstracts Annotated with Sentence Roles. In: IIAI-AAI, pp. 159–162 (2016).
4. Kando, N. et al.: The NTCIR Workshop: The First Evaluation Workshop on Japanese Text Retrieval and Cross-Lingual Information Retrieval. In: IRAL Workshop, pp. 1–7 (1999).
5. Kando. N.: Overview of the Second NTCIR Workshop. In: NTCIR Workshop (2001).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP15H01721.