Honours Year Project Report

Scientific Document Summarization Exploiting Citing Documents

By

Patrick Chen

Department of Computer Science

School of Computing

National University of Singapore

2014

Honours Year Project Report

Scientific Document Summarization Exploiting Citing Documents

By

Patrick Chen

Department of Computer Science

School of Computing

National University of Singapore

2014

Project No: 0
Advisor: 0
Deliverables:
    Report: 1 Volume

Abstract

Abstract eventually goes here.

0.1 Introduction

The volume of scientific literature has grown rapidly. And there is a need to create an automatic summarization system which enables researchers to digest knowledge in a short time. In this work, we try to use the attributes of scientific document to achieve the goal.

Scientific documents have two distinct characteristics. First, its structure must follow certain conventions. That is, you will definitely see the sections arranged like introduction, previous work, method, experiments, discussion and so on. This provides a significant advantage for a summarization system, in being able to leverage this structure. Second, scientific documents will contain a special dimension unlike other texts – citations. Scientific achievement share an accumulation of knowledge, partially based on learning from other publications by other members in the community, to make their contribution; therefore, many explanations and criticism will be made by the community through citations.

The usage of citation sentences makes scientific document summarization become an more interesting problem to solve. Citation sentences can present certain advanced knowledge processed by human, and generate results that are not available from original article; and how to efficiently use citation sentences to generate summaries is still a open problem. In this work, I will review and analyze the previous methods and develop my own solution.

0.2 Related Work

In fact, using citations to generate a summary is not a new idea, as past research have explored in this direction. The effectiveness of the citation is verified in (1) and (2). Inside both study, they use different measures to point out the fact that there are indeed some valuable information in the citations which cannot be obtained from the source papers. And based on the problem formulation, previous studies can be categorized into 2 types.

In the first type of problem, we assume there is a set of citing sentences and related features provided. The task of this type of problem is figuring out how to select the sentences to formulate the system generated summary. (3) studies the argumentative zoning, a technique for determining the rhetorical status of a sentence, for improved citation; however, this work doesnt consider how to use the identified zoning to construct the summary and the categories of zoning are not diverse enough to cover all kinds of citing sentences. (4) also tries to understand the function of citation sentences and proposes certain features for automatic classification, but again, it doesnt really show the effectiveness of using these categories to make a summary. Besides, it requires lots of annotation work. (5) use citations to analyze the impact of sentences from original paper, and acheieve high performance. But they do not directly use the sentences from the citation context to generate the summary, which will lose certain information as pointed above. (6) model the citing sentences in a graph with many clusters, and pick the most salient sentences by Lexrank score. This method has good performance and it is easy to implement, but it only uses statistical tf-idf score to classify the sentences, which is different from human's behavior. (7) define the keyphrases in citation and select the citing sentences with most keyphrases. It outperforms (6) but again it only reflects the statistical results which will overlook some sentences with semantic meaning.

For this type of problem, the assumption of having a set of citing sentences which are well processed might not be realistic. It usually requires lots of labours to annotate the corresponding labels. Besides, in most of time, we cannot use explicit citation sentence directly. Explicit citation sentence means the sentence containing the citation marker like []. As pointed by (7), sometimes the useful information is in the surroundings. And it will be very difficult to automatically identify the useful part of citations; therefore, how to automatically get good citation sentences become a research topic itself. This type of research topic is not well explored. To best of my knowledge, only (8) and (9) have systematic study of this problem; In (8), they try to extract useful information from only single citation sentence; and in (9), they focus on finding out the non-explicit citations. It is shown in (8) that good citation sentences will result in a better result with the same summarization system; therefore, this type of problem really worth exploring; however, again it requires even more human work to get training data. Due to the limitation of resources I have, in this project, I will focus on the first type of problem, especially those don't require training data.

0.3 Problem Definition

As discussed above, there are different methods to solve the problem using citation sentences. Here, I point out the exact problem to solve by defining it formally. The problem can be formulated as given a set of citing sentencs

S = S1,S2,......Sn, and the length limit of the final sumaary N, we are trying to find out k sentences from S such that words count of these k sentences are smaller than N. We will assume that all sentences in S are well processed and it will also contain non-explicit citations; however, due to the limited resources, the reference scope is not identified so there will be some redundant part in the citing sentence.

0.4 Methods

In many AI works, researchers try to build up systems simulating human's behavior. For example, (10) have analyzed the formulation of review papers and tries to figure out how human summarize the scientific article. And based on personal experience, I believe when we are doing summarization, we catogorize sentences into different functional groups as topics, methods, experiments, results, ...etc. And we will consider the best combination between the groups and select appropriate sentences correspondingly to form the final summary. As we can observe, formulating different groups is the most important step in the whole process. Human might use functional or semantic meanings of sentence to classify the sentences. And in this work, I want to simulate this process and try to find out the useful representations which can generate a good summary.

As mentioned above, (6) classify citing sentences into groups by cosine tf-idf similarities between the sentences. This step is similar to human behaviour in the sense that it will produce different groups; therefore, it's a good start for trying to simulate human decision behaviour. The Lexrank method used in this work is to find out the most salient sentence in a certain cluster. It is easy to implement and has nothing to do with clustering step; therefore, I will keep using this tool once the clustering is done. Instead of using statistical measure, sementical and functional informtaion will give us a more human-like measure. In (10), they have collected many patterns of human summarization in regular expression form. Although it is a rule-based system, it provides many details about the intention of, I will utilize these rules and combine with Lexrank method to summarize the documents.

0.5 Evaluation

Data

The data we use to evaluate is collected from ACL Anthology. We randomly sample 10 articles from the Anthology. Then we need to check if the sampled articles have enough citations. We replace the ones having less than 10 citations with the new sampled articles. And repeat this procedure till all 10 articles have enough citations. The final selection of the the data is listed as follows: P99-1026, W11-2821, W06-3312, P07-3014, C08-1122, P10-1024, C00-1073, N06-1031, J81-3002, J93-1005. The number inside the parentheses is the citing sentences used. Due to the citation processing problem mentioned above, currently the citation sentences are collected by human work, and the reference scope and explicit citation are also identified by myself.

ROUGE score and golden standard summary

In fact, trying to objectively evaluate a subjective task is very difficult. There are many existing evaluations there but none of them is perfect. In this project, we are going to use ROUGE to evaluate the result. ROUGE stands for Recall-Oriented Unders study for Gisting Evaluation. It counts the number of overlapping units such as n-gram, word sequences and word pairs between the system-generated summary and ideal summaries created by humans; however, it won't give high score to the summary with synonyms of words appearing in the golden summary. In fact, in (C-Lexrank), they notice this problem and use pyramid method to evaluate the result. pyramid method requires lots of annotation works to identify certain useful words, called nugget, and try to calculate the coverage of nuggets in the summary. It prefers results using less words to express condensed gists. Indeed, it's hard to find a perfect measure which takes everything into consideration. Due to the limitation of resources, I believe ROUGE is the best measure we should use as it requires only human generated summary and contains less annotation works. In this project, we have 6 group members and everybody will summarize all 10 sampled papers. To make the summarization easier and suitable for other tasks, we summarize the article by extracting whole sentences from the source paper, and the length of the summary is limited around 100 words. It roughly contains 4-5 sentences.

Result and Discussion

C-Lexrank : P07-3014 $- -$

$- -1ROUGE - 1Average_R :$

$0.26598(951ROUGE-1Average_P : 0.25882(951ROUGE-1Average_F : 0.26235(95-------------------------------$ $---------------1ROUGE-2Average_R : 0.05401(951ROUGE-2Average_P : 0.05254(951ROUGE-2Average_F : 0.05326(95--------$ $------------------------------------$ $------1ROUGE-LAverage_R : 0.23143(951ROUGE-LAverage_P : 0.22521(951ROUGE-LAverage_F : 0.22828(95-------------$ $-------------------------------$ My System : P07-3014 $2ROUGE-1Average_R : 0.32124(952ROUGE-1Average_P : 0.36471(952ROUGE-1Average_F : 0.34160(95---------------$ $---------------------------------2ROUGE-2Average_R : 0.08711(952ROUGE-2Average_P : 0.09901(952ROUGE-2Average_F : 0.09268(95-------------------------$ $------------------------2ROUGE-LAverage_R : 0.28670(952ROUGE-LAverage_P : 0.32549(952ROUGE-LAverage_F : 0.30487(95pitodogo@POP-Ubuntu : /SciDoc/src/RELEASE-1.5.5$ C-Lexrank : C08-1122 $--------------------------$ $------------------------1ROUGE-1Average_R : 0.27163(951ROUGE-1Average_P : 0.25323(951ROUGE-1Average_F : 0.26211(95-------------------------------$ $---------------1ROUGE-2Average_R : 0.01745(951ROUGE-2Average_P : 0.01626(951ROUGE-2Average_F : 0.01683(95--------$ $------------------------------------$ $------1ROUGE-LAverage_R : 0.24567(951ROUGE-LAverage_P : 0.22903(951ROUGE-LAverage_F : 0.23706(95-------------$

$------------------------------------$ My System : C08-1122 $2ROUGE-1Average_R$: $0.23356(952ROUGE-1Average_P$: $0.27273(952ROUGE-1Average_F$: $0.25163(95----------------$ $------------------------------------2ROUGE-2Average_R$: $0.03665(952ROUGE-2Average_P$: $0.04286(952ROUGE-2Average_F$: $0.03951(95-----------------------------$ $----------------------------2ROUGE-LAverage_R$: $0.20588(952ROUGE-LAverage_P$: $0.24040(952ROUGE-LAverage_F$: $0.22180(95$

0.6 Conclusion

References

A. F. G. E. D. S. Aaron Elkiss, siwei Shen and D. Radev, “Blind men and elephants: What do citation summaries tell us about a research article?,” *Journal of the American Society for Information Science and Technology*, 2008.

M. E. A. H. P. M. V. Q. D. R. D. Z. Saif Mohammad, Bonnie Dorr, “Using citations to generate surveys of scientific paradigms,” 2009.

S. Teufel, *Argumentative Zoning for improved citation indexing.* Springer, 2005.

Automatic classification of citation function, 2006.

Generating Impact-Based summaries for Scientific Literature, 2008.

V. Qazvinian and D. R. Radev, “Scientific paper summarization using citation summary networks,” in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, (Manchester, UK), pp. 689–696, Coling 2008 Organizing Committee, August 2008.

A. O. Vahed Qazvinian, Dragomir R. Radev, “Citation summarization through keyphrase extraction..,” 2010.

D. R. Amjad Abu-Jbara, “Reference scope identification in citing sentences,” 2012.

A. O. Vahed Qazvinian, Dragomir R. Radev, “Identifying non-explicit citing sentences for citation-based summarization,” 2010.

Imitating Human Literature Review Writing: An Approach to Multi-Document Summarization, 2010.