EE5002 Project Report

Summarizing Scientific Document with Machine Learning Methods

By

CHEN, PEI-HUNG
Matriculation No. A0103184R

Department of Electrical and Computer Engineering
Faculty of Engineering
National University of Singapore

2014

Abstract

In this study, I try to figure out how to use citation sentences to generate a good summary for scientific documents. Scientific documents have its distinct characteristics so it can provide useful information which cannot be found in the source paper. And how to use the citation information is still an open question. I try to simulate the human behaviour during summarization and adopt the cluster-based method to finish this task. The sentence similarity network will be formulated and the regular expression will be applied to do the clustering. The final summary is composed by sentences in individual clusters selected by lexrank scores. The system is tested on 12 ACL Corpus and results show that it can generate good summary providing information in specific categories.

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 a 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 an open problem. In this work, I will review and analyze the previous methods and develop my own solution.

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 (Elkiss et al. 2008) and (Mohammad et al., 2009). 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 two 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. (Teufel. S., 2005) studies the argumentative zoning, a technique for determining the rhetorical status of a sentence, for improved citation; however, this work does not 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. (Teufel. S., 2006) also tries to understand the function of citation sentences and proposes certain features for automatic classification, but again,

it does not really show the effectiveness of using these categories to make a summary. Besides, it requires lots of annotation work. (Mei & Zhai, 2008) use citations to analyze the impact of sentences from original paper, and achieve 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. (Qazvinian and Radev, 2008) 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. (Qazvinian et al, 2010a) define the key-phrases in citation and select the citing sentences with most salient key-phrases. It outperforms (Qazvinian and Radev, 2008) 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 labors 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 (Qazvinian et al, 2010b), 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 becomes a research topic itself. This type of research topic is not well explored. To best of my knowledge, only(Abu-Jbara and Radev, 2012) and (Qazvinian et al, 2010b) have systematic study of this problem; In (Abu-Jbara and Radev, 2012), they try to extract useful information from only single citation sentence; and in (Qazvinian et al, 2010b), they focus on finding out the non-explicit citations. It is shown in (Abu-Jbara and Radev, 2012) 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.

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 sentences S = S1,S2,.....Sn, and the length limit of the final summary 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.

4. Methods

In many AI works, researchers try to build up systems simulating human's behavior. For example, (Kokil et al., 2010) have analyzed the formulation of review papers and tries to figure out how human summarize the scientific article. And based on practical experience, I believe when we are doing summarization, we categorize sentences into different functional groups as topics, methods, experiments, esults, ...etc. And we will consider the best combination between the groups and select appropriate sentences correspondingly o 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.

4.1 Using Regular Expression

As mentioned above, (Qazvinian and Radev, 2008) 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, semantic and functional information will give us a more human-like measure. In (Kokil et al., 2010), 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 human summary, so it will more likely provide us better information to do the clustering. In (Kokil et al., 2010), the authors categorize the rules in detailed. To apply it to the task, I manually select the similar categories and put them together into a group. There are four groups in my regular expression set: Topic, Method, Result and Evaluation. Each of them covers different aspects of the summary so the matched sentences can be clustered by these four types. However, in practice we cannot directly apply the regular expression to the citation sentences. The reason is two-folded. First, if the number of sentences is small, we might not be able to get

enough matched sentences from regular expression to form the final summary. So we cannot only use the matched sentences. Second, these regular expression rules are not able to cover all kinds of features precisely. There are exceptions found in the regular expression rules of (Kokil et al., 2010), and we are not able to define the rules that can be used exactly. So sometimes, the matched sentence will not provide any specific information. For example, here are two citation sentences which is recognized by the keyword "using":

"Results supported our assumption that people using the structured text would judge the tasks easier, although performance was similar on both texts.[12]."

"Virtual environments enjoy considerable interest (Koller et. al., 2010), and this added to our motivation for using them."

And apparently the first one provides certain important information but the second one is useless to us. If we just use the regular expression, it is more likely that we will select some inappropriate sentences; therefore, we need to figure out how to utilize regular expressions in an indirect way.

4.2 Regular Expression on Sentence Similarity Network

The simplest way to use the rules indirectly is to modify the existing structure with the help of regular expression. In the (Qazvinian and Radev, 2008), they build a network by calculating tf-idf similarity score on all the citation sentences, and it provides a good starting point for me to adjust the corresponding weights by using regular expression. The basic idea is that if two citation sentences belong to the same recognized group (ex: topic, method), we will alter the weights between them by certain factor, and we can adjust the weight between two sentences which are not in the same group too. This process will enable the system to change the structure of network and produce different kinds of clusters. And it is more likely to form cluster for each kind of matched sentences. Besides, we can also adjust the lexrank score by a certain factor for the matched sentences to ensure that the matched sentence will be selected. The advantage of this method is obvious. Now, we will be able to use the regular expression and in the meantime we keep other unmatched sentences. Groups are generated in a way which is more similar to human behaviour and we will be able to ensure the matched sentences are more likely to be selected. The system has three parameters to tune: the adjustment of weight between matched sentences, the adjustment of weight between unmatched sentences and the adjustment of lexrank

score of matched sentences. These parameters can be decided by the K-fold validation.

5. Evaluation

5.1 Data

The data we use to evaluate is collected from ACL Anthology. We repeatedly sample at random from ACL Anthology until we have 12 articles with sufficient citations each. The final selection of the data is listed in Table 1:

Article ACL Anthology ID	Number of Citations
A97-1022	11
C08-1122	32
C90-2052	17
E91-1040	17
P06-1110	21
P07-3014	7
P10-1024	9
P97-1059	9
P99-1026	24
W10-4233	26
W11-2821	17
W93-0225	7

Table. 1

The number inside the parentheses is the citing sentences used. Due to the limited resource I have, the size of the data set is not very large. But as in other study (Qazvinian et al, 2010a), the size of data set is also just 25, which implies that it will still provide significant result for well annotated small data set. 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.

5.2 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 quantitatively. 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 which are called model summary in ROUGE; however, it won't give high score to the summary with synonyms of words appearing in the golden summary. In fact, in (Qazvinian and Radev, 2008), 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 try to summarize all 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. I will compare my system against C-lexrank as it is the baseline of my system, and it is important to show the regular expression information can indeed improve the performance.

5.3 Qualitative Analysis

Although the ROUGE score can validate the effect of regular expression and the system, the model summaries are done by limited people which might not be general enough to reflect the best summary we want. That is, the ROUGE score can represent certain meanings but the higher score doesn't necessarily mean a better summary; therefore, besides quantitative result, we also need to analyze it qualitatively. Basically, we would like to know the basic characteristic of a system and what kinds of sentences the system will generate.

6. Result and Discussion

6.1 Qualitative Analysis

SYTSTEM	ROUGE-1	ROUGE-L
C-lexrank	0.33501	0.29175
RE-lexrank	0.34255	0.30370

Table. 2

As we can see in the above table, our performs better than C-lexrank numerically in average of 2 different trails of the 4-fold validation. Although the improvement is not very large, one thing to note is that it is the average value over 12 samples, and some of the samples contain only limited number of citation sentences, so it is impossible to improve those samples. In general, one-fourth of the samples will generate the same summary as C-lexrank and one-fourth of the samples perform slightly worse than C-lexrank, and the remaining half of the samples will perform greatly better than C-lexrank. It is not surprising that our system outperforms C-lexrank as we provides more features on top of C-lexrank, and it is flexible to use these new features. More important thing is still to understand the why the system can generate better result. So we will do the qualitative analysis to study the behavior of the system next.

6.2 Qualitative Analysis

After examining all the summaries produced by my system, I figure out there is no guarantee about what kinds of sentences will be generated. The characteristic of summary are basically not static. It is because our system has three parameters to control the final output, and with different partition of K-fold validation, different parameters will be used in the final evaluation; therefore, the output of the system is actually controlled by the training data and it is not fixed. The only certain thing is that my system indeed has ability to form different kinds of clustering, so basically the presence of matched sentences can be controlled by my system. Then it will be valuable to analyze the quality of these sentences matched by regular expression. Besides, we should also investigate if once a specific group of matched sentences are formulated in the network, the lexrank score is able to pick few best sentences from that cluster. These two analyses will let us understand more about the characteristic of the system.

6.2.1 Characteristics of Matched Sentences

ID	Sentence	Selection
P97-1059	, for example, develop methods for modeling the tag disambiguation task by means of a finite-state device.	X
P97-1059	hidden markov models can also be viewed as stochastic finite-state transducers and it is possible to closely approximate them by composing fst in a deterministic way , without a significant loss in accuracy.	O(4)
P97-1059	fst have been used for various tasks including recognising part-of-speech tags .	O
P97-1059	in natural language processing, fsts are used for many basic steps such as part-of speech disambiguation .	O
P97-1059	in principle, such a conversion could be used as an alternative approach to querying hmms, with the advantage that query answering could be done by means of composition of transducers.	O(1)

Table. 3

Model Summary Sentence	Number
This paper describes two algorithms 1 which approximate a Hidden Markov Model (HMM) used for part-of-speech tagging by a finite-state transducer (FST)	1
Since all transducers are approximations of HMMs, they give a lower tagging accuracy than the corresponding HMMs	2
The tagging speed of the transducers is up to five times higher than that of the underlying HMM.	3
The main advantage of transforming an HMM is that the resulting FST can be handled by finite state calculus 1° and thus be directly composed with other transducers which encode tag correction rules and/or perform further steps of text analysis.	4

Table. 4

An example of the procedure of the analysis is shown in the above tables. First, we compare the semantic meaning between citation and the model summary. If there are basically the same semantically, I will consider it to be useful and mark it with the number of matched model sentence. Second, I will go through all sentences which are not marked in the first step again, and judging if each sentence provides useful information or not. In the second step, we will not consider model summary as we

would like to find out the sentences which are actually good but not included in the model. And the result of the selection is shown in the table below.

ID	Number of Matches	Number of Selections	percentage
A97-1022	6	3	50%
C08-1122	20	14	70%
C90-2052	3	2	67%
E91-1040	4	3	75%
P06-1110	10	6	60%
P07-3014	5	3	60%
P10-1024	5	3	60%
P97-1059	5	4	80%
P99-1026	15	10	67%
W10-4233	5	2	40%
W11-2821	9	5	56%
W93-0225	4	2	50%
Overall	91	57	63%

Table. 5

As we can see, only 60 percent of matched sentence will be thought as good citation. Although it is not a very satisfactory result, it still shows that part of the important information can be captured by the matched sentence. Besides, we can also analyze the number of matches from each type of regular expression, and the result is shown in the following table.

Method	Topic	Result	Evaluation
66	22	0	3

Table. 6

According to the experimental result, we can clearly observe that the methods and topics are easier to be recognized by the regular expressions, which means that the system is more likely to provide method and topic related information with citation sentences. Therefore, even though the good citations only occupy 60%, majority of it belong to method and topic so we are still confident in finding out useful information for these two types.

On the other hand, the information as result, experiment settings and system performance seem to be unlikely matched by regular expression, so our system might not be able to summarize these information. This characteristic is also reflected in the phenomena that sometimes the result of validation suggest us to decrease the lexrank scores of matched sentences. In this case, it implies that the model summary contains information different from matched sentences. It is likely that there is more result, evaluation or other kinds of information described in the model summary.

There are two possible explanations for this outcome. First, people seldom cite the result numerically. In the 12 citation sentence data we collected, it is true that only few sentences containing result information. But our data set only contains papers in computational linguistics. There might be more result information in other disciplines. Second, some of the results might be disclosed together with the description of method. So part of the results are hided by other types of regular expression. It means that the regular expression currently used is not precise enough to cover all kinds of situation. Here comes an example:

"using an svm, he achieves 50.1*% accuracy on a 20-fold cross-validation for the 5-class barker & szpakowicz dataset. 0.25"*

This sentence will be matched by method category, but in fact it contains the information of test results; however, my system is unable to resolve this conflicts now.

6.2.1 Characteristics of Lexrank Score

In order to study if lexrank score is a good indicator of selecting sentences, we select a set of parameters which tend to generate clusters for matched sentences. And for these clusters composed by matched sentences, I will try to order it by judging the significance of information contained. Then I compare the order with the lexrank scores to see if lexrank score roughly matches the order. In the end, 19 regular expression matched clusters are formed and 12 of them give the desired order by lexrank score. And 7 of them doesn't provide good results. But for 6 out of 7 not satisfied clusters, it actually has all equal lexrank scores as shown in the following table.

ID	Citation Sentence	Lexrank Score
P97-1059	fst have been used for various tasks including recognising part-of-speech tags .	0.5
P97-1059	in natural language processing, fsts are used for many basic steps such as part-of speech disambiguation .	0.5

Table. 7

It is because the lexrank score calculates the tf-idf score between the sentences, so if there is no overlap between two sentences, each sentence will only be similar to itself so each sentence will have the same lexrank score. Under this circumstance, the system is basically picking up sentence randomly from the cluster; therefore, it cannot guarantee to select the best sentences.

So we understand that lexrank score alone is not enough to distinguish all possible clusters; however, since it is easy to recognize under what condition lexrank score is not functional, we can further create another metric to select the sentences when the lexrank score are all equal within the cluster. By combining this new metric with lexrank, the system will be able to select best sentence from the matched-sentences cluster.

7. Conclusion and Future Work

In this study, I demonstrate the scientific summary can be generated by simulating the human decision. By selecting appropriate parameters, the system is able to formulate clustering that each group is composed by specific type of information. And the regular expression can successfully identify appropriate sentences for each group. The lexrank score is also an eligible selector which will give us representative sentences within the cluster. Although the current regular expression is not precise enough to cover all the situations, at least we understand the strength of this method and we can combine it with other methods in the future.

One thing to note is that the result information is hardly found in the citation sentences, but it is also difficult to find it in the original source paper. In the original paper, the result is more likely presented by the forms of graph, table and figures instead of text. And it is very hard to find one single sentence in the original paper to summarize the empirical result.

References

Abu-Jbara, A., & Radev, D. (2012). Reference scope identification in citing sentences, Conference of the North American Chapter of the Association for Computational Linguistics, 2012.

Elkiss, A., siwei Shen, Fader, A., Erkan, G., States, D., & Radev, D. (2008). 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.

Chin-Yew Lin, ROUGE:a Package for Automatic Evaluation of Summaries, Proceedings of the Workshop on Text summarization Branches Out, 2004.
Qiaozhu Mei and ChengXiang Zhai. Generating impact-based summaries for scientific literature. Proceedings of ACL-08: HLT, (2008)

Kokil Jaidka and Christopher Khoo and Jin-Cheon Na. Imitating human literature review writing: An approach to multi-document summarization. Proceedings of ICADL, 2010.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., & Zajic, D. (2009). Using citations to generate surveys of scientific paradigms, , 2009.

Teufel, S. Argumentative zoning for improved citation indexing. Computing Attitude and Affect in Text: Theory and Applications Springer 2005.

Simone Teufel, Advaith Siddharthan and Dan Tidhar. Automatic classification of citation function. Proceedings of EMNLP-06, 2006

Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization us- ing citation summary networks. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) (pp. 689–696), Manchester, UK, August, 2008: Coling 2008 Organizing Committee.

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

Qazvinian, V., Radev, D. R., & Ozgur, A. (2010). Citation summarization through keyphrase extraction.., , 2010.b