PKU @ CLSciSumm-17: Citation Contextualization

Dongxu Zhang*, Sujian Li

Peking University
zhangdongxuu@gmail.com

Abstract. This report gives a brief introduction of our participation in CL-SciSumm 2017 Task 1A. We demonstrate some data analysis and point out the difficulty of this task. Then we report both unsupervised and supervised methods with their performances on 2016 and 2017 testset, from which efficiency of different features can be estimated.

1 Introduction

Reading scientific articles is necessary but time-consuming for researchers and engineers. Although there are abstracts in papers, readers still find it difficult to understand key contributions of a paper. On one hand, original abstracts usually state in a general but less focused fashion and sometimes they do not contain all aspects of their papers. On the other hand, we might not completely believe contributions written in the abstracts, since they may be over or under-stated by authors, and may not get fully agreements from the research community. Thus, scientific paper summarization aims to automatically captures more detailed and complete contributions of a paper, objectively.

Scientific paper summarization is an NLP sub-task of automatic summarization. Different from traditional general domain such as newswire, scientific article is a special domain with extra features such as citation links, special discourse structures, etc.

To employ different aspects of a paper, [14] analyzed *rhetorical status* of each sentence from original papers to tackle scientific summarization task. [5] pointed out that sentences that cite the paper, which we call *citation sentences* are more semantically consistent and contain more information in contrast to original target papers' abstracts. After that, [11] employed citation sentences for scientific summarization. They clustered sentences from citation papers which cite the target paper and formed the summary using the central sentence of each cluster.

Not long ago, [1] improved the aforementioned method by using *citation context* from the original target paper to produce summaries. *Citation Context* in their work refers to sentences in the reference paper (former) that are most related to the citation sentence from the citation paper (latter). The author pointed out that, although citation sentences are more focused and objective, information dissemination may cause failure to accurately reflect contributions of original papers. This method can be regarded as a combination of classic sentence extraction summarization and citation sentence-based summarization, which does not only remain the original information but also absorbs the consistency and objectiveness of citations.

* This research was conducted during the author's visit at PKU.

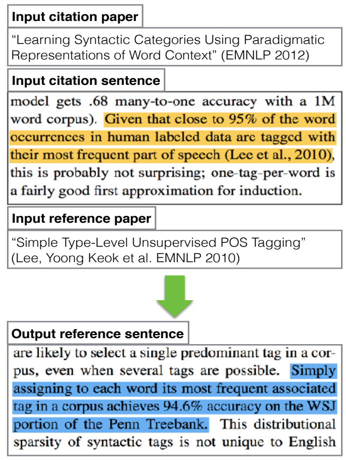

Fig. 1. Task definition of citation contextualization.

In this report, we name the task of finding related sentences from reference papers *Citation Contextualization*, and will focus on analyzing and modeling this task in the rest of this paper. This task was first introduced in TAC 2014 Biomedical Summarization Track [1], and was continued by Computational Linguistic Scientific Summarization (CL-SciSumm) Track continued in 2016 and 2017 [6].

Task Definition As shown in Fig. 1, *Citation Contextualization* has three inputs: the information of citation sentence, citation paper, and target reference paper. These inputs could be distinguished in one word: the citation paper *cited* the reference paper via the citation sentence. This task requires a system to find one or several sentences from the reference paper which can best support the existence of this citation link. Most related work [4, 2, 3, 8] regard this task as a point-wise ranking problem where given these inputs, a system will score each sentence in the reference paper individually and then pick up top-rated sentences as the output. Pair-wise ranking methods with negative sampling were also used by some of the previous work [9].

In this report, we will analyze this task from different angles, then demonstrate our methods and some experimental results. Our system participated in the CL-SciSumm 2017 [6], Task 1A.

2 Data Analysis

CL-SciSumm 2017 regards all datasets in track 2016 as the training set, and provided extra 10 papers as the evaluation set of this year. To help following researches better understand this task, some data analysis will be presented in this section. There are two annotated datasets for this task: TAC 2014 Biomedical Summarization Track

[1] https://tac.nist.gov//2014/BiomedSumm/

(BioSumm 2014) and Computational Linguistic Scientific Summarization Track 2016 (CL-SciSumm 2016). Since we can only access the training data of BioSumm 2014, we will mainly focus on the CL-SciSumm 2016 dataset in the rest of this paper, and only mentioned BioSumm 2014 in the Human Performance section.

2.1 Annotation Distribution

We did the analysis on CL-SciSumm 2016 dataset, which contain 20 training and development reference papers and 10 test reference papers. Detailed statistics are shown in Table 1.

CL-SciSumm 2016	Train & Dev	Test
Number of reference papers	20	10
Number of citation sentences in total	354	340
Median of citation papers per ref paper	9.0	16.50
Median of citation sentences per ref paper	16.5	23.50
Average length of citation sentences	33.9	34.6
Average length of citation contexts	22.5	26.0

Table 1. CL-SciSumm 2016 data statistics.

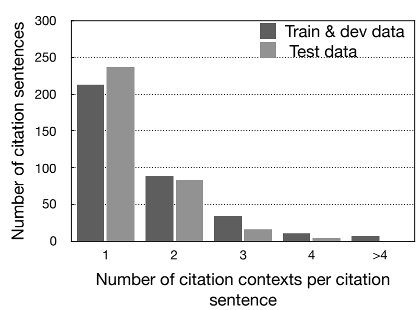

Fig. 2. Distributions of number of citation contexts per citation sentence.

Fig. 2 represents distributions of number of citation contexts per citation sentence. We can find out that most citation sentences are only annotated by one related sentence from the reference paper. This is probably caused by the difficulty of annotation in this task. In this case, most previous work also utilized Rouge to evaluate models' performances. Also it seems to be reasonable to choose top three sentences for each citation. This setup has been followed by most previous work.

Fig. 3 shows the proportions of different section types where citation contexts locate. Since different authors prefer different section names, it is hard to normalize them into a closed name space. So, in this figure, we only choose four standard section names which occupy significant amount of portion, and let rest of names become "other" which mainly contains related work, method and experimental result sections. Notice that, the annotation rule of this track suggests annotators that, if there is no text matches, the title of the reference paper should be chosen. Fig. 4 shows the proportions of different discourse facets of citation contexts. Here, discourse facet type stands for the discourse role of a sentence in the paper.

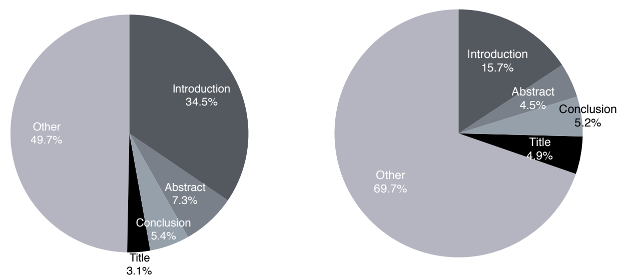

Fig. 3. The proportions of section types citation contexts locate. Left: Training and dev data. Right: Test data.

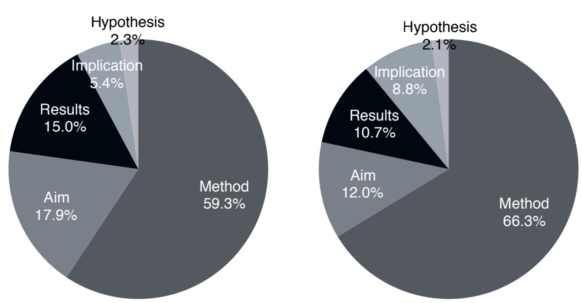

Fig. 4. The proportions of discourse facet types citation contexts belong to. Left: Training and dev data. Right: Test data.

2.2 Human Performance

Citation contextualization is a difficult task. To evaluate how hard it is, we use the BioSumm 2014 dataset [2] to evaluated human annotators' performance. BioSumm 2014

[2] `https://tac.nist.gov/2014/BiomedSumm/data.html`

dataset presented each citation sentence four different annotation results from four different annotators. Given a citation sentence, each annotator will find up to four related sentences from the target reference paper.

To evaluate human performance, each time we regard one annotator's annotation as the predicted result, and the other three annotations as gold answers. Then we average four annotators' performances as the human performance. We choose results from two recent previous work for comparison.

Models	c-P	c-R	c-F1
Key Word Query Reformulation Method [4]	22.6	29.4	24.1
Biomedical Word Embedding-based Method [2]	23.9	31.2	25.5
Human Performance	32.0	29.3	27.5

Table 2. Human Performance on BioSumm 2014 training set. c-P: character level precision, c-R: character level recall, c-F1: character level F-1 value.

From Table 2, we can see that human performance does not reach 30% F-1 value. Since baseline methods always return top 3 sentences as results, the recall rates are even slightly higher than the human performance.

The result of low consistency among human annotator indicates the difficulty of providing one unique gold answer. It seems to be hard to clearly define the "relatedness" between reference sentence and citation sentence, and different annotators could hold different point of views.

One solution is to define "relatedness" from several more specific dimensions. Another possible solution is to change the current evalution method: For each citation sentence, rather than annotating one gold answer, it might be more reasonable to let annotators score different results from different systems. Though this evaluation method is much more expensive and time-consuming, it should be easier for annotators to distinguish between different results than to pick up answers from the whole reference paper.

3 Methods

3.1 Preprocessing

For each sentence in the corpus, we first replace all reference groups using "TARGETREF" or "NORMALREF". Here, a *Reference Group* is one or multiple citation markers in the same bracket. We recognize reference groups in a citation sentence using bracket pair signs like (,) and [,] and there also required to be at least one two digit number ranging from 00 to 99 or four digit number ranging from 1950 to 2020 inside that bracket pair.

Then, we replace all number with "NUMBER". Finally replace all punctuations with blank space, except for segmentation punctuations such as , . ; ? !

3.2 Search-based Similarity Scoring

We first employ search-based method with different features for sentence pair similarity calculation. For each citation sentence, we choose top 3 most rated sentences in the reference paper as answers.

For TF-IDF methods, we follow the Key Word configuration in [4] where they only kept terms whose IDF are larger than a threshold (2.5 is chosen in practice) to only remain informative words. TF and IDF values are calculated in the sentence level. We employed *CountVectorizer* and *TfidfTransformer* from sklearn [10] to calculate tfidf. And we employ *Word2vec* from gensim [13] to train our word vectors. Both IDF and word embedding are counted using ACL anthology text corpus [12].

The word embedding model refers to [2] which regards sentence similarity as conditional probability between the citation sentence and the reference sentence. They utilized distance from word embedding as the basis of this probability. Word movers distance refers to [7] which measures the distance between two sentences as the minimum distance that the embedded words of one sentence need to move to reach the embedded words of another sentence.

3.3 Supervised Method

we combine both training and development datasets for logistic regression learning using sklearn. Several features are employed, such as similarity of 1-3gram TF-IDF, similarity of 3gram character level TF-IDF, similarity of word embedding-based model [2], similarity of word embedding average, and section type of reference sentence. Then, we also produce top 3 sentences based on the scores from logistic regression classifier.

4 Results on 2016 Test Set

Models	P_{sent}	R_{sent}	F_{sent}
Unigram TF-IDF (vocab = 5k)	9.4	22.2	13.1
1-3gram TF-IDF (1-3gram vocab = 200k)	9.5	22.5	13.3
3gram char TF-IDF (3gram char vocab = 5k)	9.2	21.9	12.9
(vocab = 20k, dim = 300, min_freq_count = 100)			
Word embedding average	6.5	15.5	9.1
Word movers Distance [7]	7.2	17.4	10.1
Word embedding model [2]	7.6	18.7	10.8

Table 3. Performances of search-based methods with different bag-of-word features on 2016 testset.

Table. 3 shows performances of search-based methods with different bag-of-words models. It shows that ngram tf-idf is slightly better than unigram tfidf. Surprisingly, character level ngram model performs quite well. We haven't successfully reproduced

Models	P_{sent}	R_{sent}	F_{sent}
All feature + logistic regression	11.6	27.5	16.2
- ngram TF-IDF (ngram vocab = 200k)	11.0	26.6	15.5
- 3gram char TF-IDF (3gram char vocab = 5k)	9.7	23.2	13.6
- Word embedding model	11.3	27.2	15.9
- Word embedding average	9.8	23.3	13.7
- section type	10.6	25.1	14.9

Table 4. Performances of supervised method with different features on 2016 testset.

the results from [2]. And in our experiments, it seems that performances of simple word embedding-based alignment methods are in average worse than TF-IDF methods.

Table 4 shows performance of the supervised model trained on training and development set of CL-Scisumm 2016. Slightly different from results reported in [3], it seems that supervised model performs much better than previous unsupervised methods. After removing each feature individually, we can see that ngram character-level TF-IDF , average word embedding similarity and section type features contribute most to classification.

Finally, we employed all hyper-parameter setups and trained the classifier on 2017 training data, and produced results on 2017 test data as our submission. The results are listed in Table 5. First two rows represent the results of unsupervised systems. And the last row stands for the supervised model.

Models	P_{sent}	R_{sent}	F_{sent}
1-3gram TF-IDF (vocab = 200k)	5.9	14.1	8.4
Word embedding model (vocab = 20k, dim = 300)	5.6	13.3	7.9
All feature + logistic regression	8.4	19.1	11.7

Table 5. Results of system submissions on CL-SciSumm 2017 testset.

5 Conclusion

This paper is a system report on CL-SciSumm 2017 Task 1A. We analyzed annotation datasets and found out the difficulty of this task: since “relatedness” between reference and citation sentences is quite hard to define, both annotation process and question modeling become hard.

We also briefly introduced our system’s methods and their performances on CL-SciSumm 2016 dataset. From results, it indicates the effectiveness of supervised methods and shows contributions of different features.

References

1. Cohan, A., Goharian, N.: Scientific article summarization using citation-context and article's discourse structure. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 390–400. Association for Computational Linguistics, Lisbon, Portugal (September 2015), http://aclweb.org/anthology/D15-1045
2. Cohan, A., Goharian, N.: Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 17 (2017), http://doi.acm.org/10.1145/3077136.3080740
3. Cohan, A., Goharian, N.: Scientific document summarization via citation contextualization and scientific discourse. International Journal on Digital Libraries pp. 1–17 (2017), http://dx.doi.org/10.1007/s00799-017-0216-8
4. Cohan, A., Soldaini, L., Goharian, N.: Matching citation text and cited spans in biomedical literature: a search-oriented approach. In: HLT-NAACL (2015)
5. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D.J., Radev, D.R.: Blind men and elephants: What do citation summaries tell us about a research article? JASIST 59, 51–62 (2008)
6. Jaidka, K., Chandrasekaran, M.K., Jain, D., Kan, M.Y.: Overview of the cl-scisumm 2017 shared task. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017). Tokyo, Japan (2017)
7. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning. pp. 957–966 (2015)
8. Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., Cong, X., Peng, H.: Cist system for cl-scisumm 2016 shared task. In: BIRNDL@ JCDL. pp. 156–167 (2016)
9. Nomoto, T.: Neal: A neurally enhanced approach to linking citation and reference. In: BIRNDL@ JCDL. pp. 168–174 (2016)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
11. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 689–696. Association for Computational Linguistics (2008)
12. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The ACL anthology network corpus. In: Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries. Singapore (2009)
13. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
14. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Computational linguistics 28(4), 409–445 (2002)