An Analytical Study on A Benchmark Corpus Constructed for Related Work Generation

Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang

School of Computer, National University of Defense Technology, Changsha 410073, China
{wangpancheng13, shashali, haifang_zhou, tangjintao, tingwang}@nudt.edu.cn

Abstract. Automatic related work generation aims at producing a related work section for a given scientific paper. Demand for this task replacing a labor-intensive process has substantially increased in recent years. Considering the lack of an open and large-scale dataset for related work generation, we introduce NudtRwG[1], a collection of 2,084 document sets, each with a target paper, a ground truth related work, and the corresponding reference papers. To our knowledge, NudtRwG is the first open, large-scale and high-quality dataset for related work generation. The contribution of this work apart from the dataset is two-fold: firstly, we present a detailed description of the data collection procedure along with an analysis on the characteristics of the dataset; secondly, we conduct an analytical study, investigating the effects of summative sections (abstract, introduction and conclusion) and other sections of reference papers on related work generation. Experiments reveal that the two parts are equally important and other sections should not be ignored. When generating a related work section, researchers should consider not only summative sections, but also other sections of reference papers.

Keywords: Related work generation · Analytical study · Dataset resources

1 Introduction

A related work section is a significant component of a scientific paper. Scholars need to compare their work with previous work and highlight their contributions in this section. A high-quality related work section requires scholars doing a survey of relevant researches by reading amounts of papers, summarizing relevant aspects of these researches and pointing out their weaknesses compared with own work, which tends to be an arduous and time-consuming job for scholars.

In view of this, automatic related work generation is proposed to generate a related work section for a paper being written. The task is defined and pioneered by Hoang and Kan [6], where the input is a target paper excluding the related work section, as well as reference papers of the target paper, and the output is a related work section(example is shown in Figure 1).

[1] https://github.com/NudtRwG/NudtRwG-Dataset/

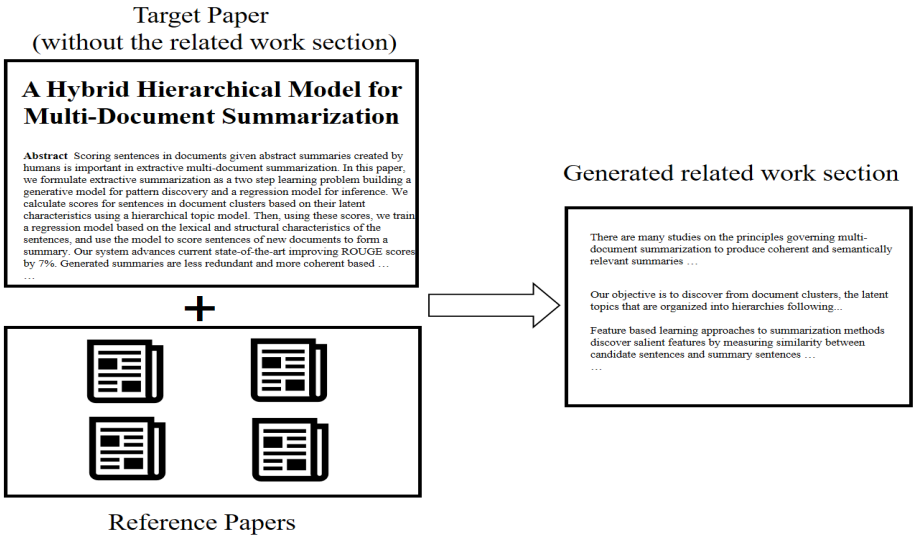

Fig. 1. Example of related work generation, given a target paper and its reference papers.

Some methods [8, 18, 1] have been explored to solve this problem since then. They solve the problem through extractive summarization methods based on their own datasets. The question is, their small and incomplete datasets render it hard to solve this problem (shown in Table 1), not to mention the unavailability of their datasets. These problems make for a fundamental obstacle for automatic related work generation, that is, previous researches cannot be tracked and compared, which is not conducive to this task.

Table 1. Data scales of previous work on automatic related work generation. "#" denotes number of.

Author	#(Document sets)	#(Average reference papers)	whether contain all the reference papers
Hoang[6]	20	10.9	No
Hu[8]	1050	Not Known	No
Widyantoro[19]	50	Not Known	Not Known
Chen[1]	25	10.5	No

Historically, large and realistic datasets have played a crucial role for driving fields forward. To address the need for a large and high-quality dataset for related work generation, we introduce NudtRwG, a collection of 2,084 document sets, each with a target paper, a ground truth related work, and the corresponding reference papers.

To the best of our knowledge, NudtRwG is the first open large-scale dataset for automatic related work generation. In contrast to prior datasets, NudtRwG not only has an edge on dataset size, but also on quality. Target papers of NudtRwG are all selected from well-received conferences of computational linguistics and natural language processing, and the average number of citations of target papers reaches 63.59. Hence, from viewpoint of academic community, quality of these ground truth related work is guaranteed. Besides, NudtRwG has more complete reference papers, with 93% of the document sets missing fewer than 6 reference papers.

Based on NudtRwG, we carry out some heuristic explorations of related work generation. We make a thorough inquiry about the lexical characteristics of the ground truth related work with contrast to different sections of reference papers. Experimental result shows that, summative sections (abstract, introduction and conclusion) of reference papers contain most information of the ground truth. However, other sections of reference papers should not be ignored. Further analysis on citation evidence (see section 5.2) of reference papers reveals other sections are competent in becoming candidate for related work generation, depending on the concrete citation purposes. In addition, we apply some general extractive summarization approaches to generate related work, with different sections of reference papers as input. It turns out that, using full texts of reference papers as input to generate related work is on par with using summative sections, which demonstrates the difficulty for extractive summarization approaches to identify salient and relevant information within the scope of full texts. Pointing at this, we propose our suggestions and expect it will be beneficial for researches afterwards.

To sum up, the main contributions of this paper include: (i) the first open, large-scale and high-quality dataset for related work generation, (ii) a detailed description of the data collection procedure along with an analysis on the characteristics of the dataset, (iii) an analytical study on the effects of summative sections (abstract, introduction and conclusion) and other sections of reference papers on related work generation and some heuristic conclusions.

2 Background

Automatic related work generation is pioneered by Hoang and Kan [6]. The authors proposed an automatic related work generation system named ReWoS, which used a given topic hierarchy tree to model the internal topic structure of related work section and strategically extracted sentences for two different contents, general content as well as specific content.

Hu and Wan [8] treated this task as a global optimization problem. They utilized probabilistic latent semantic indexing to group candidate sentences into different topic-biased clusters and applied Support Vector Regression model to score the importance of each sentence. A global optimization framework is proposed to select sentences to generate the related work section based on the former topic clusters and importance scores.

Subsequently, Chen and Zhuge [1] introduced the citation sentences, namely sentences from papers that cite the reference papers, and constructed a graph of representative keywords. Afterwards, they took advantage of a minimum steiner tree to guide the generation by extracting the least number of sentences to cover the discriminated nodes.

More recently, Wang et al. [18] developed a neural data-driven summarizer with a joint context-driven attention mechanism to generate related work section. They constructed a directed graph containing heterogenerous relations among kinds of objects such as papers, authors, keywords and venues, and designed an attention mechanism focusing on the contextual relevance within the target paper being written and the graph. For each candidate sentence, a label of 0 or 1 was assigned after a log-likelihood probability objective being optimized.

3 Dataset Construction

In this section, we describe design considerations and data collection guidelines we follow in the construction of our dataset as well as statistics. We collect our dataset in three stages: target paper collection, reference papers identification and collection, and dataset filtering and replenishment.

Target paper collection. To acquire high-quality articles, we chose papers from main conference of computational linguistics and natural language processing, such as ACL, EMNLP, NAACL, COLING, as the candidate target papers with time span ranging from 2006 to 2017. We first crawled download link for all the target papers from ACL Anthology[2], and then applied an automatic paper download tool to gather the PDF format of all the target papers as per the links. After this stage, we obtained over 3,200 target papers.

Reference papers identification and collection. Next, we converted all the target papers from PDF to text using pdfminer[3]. After this conversion, we screened out papers without a related work section. In the remaining over 2,700 papers, we semi-automatically extracted the list of references using a rule-based method, considering that conferences of computational linguistics and natural language processing often follow the same citation format. We designed specific regular expressions to identify the publication years and split references based on the identified year. Then, we retrieved all the reference papers from Google Scholar and obtain download links. The same paper download strategy was applied to obtain the reference papers. It's worth mentioning that, since some papers were not available in Google Scholar, we neglected these unavailable reference papers.

[2] https://www.aclweb.org/anthology/
[3] https://pypi.org/project/pdfminer/

Dataset filtering and replenishment. After the above two stages, there were more than 2,700 document sets at hand. For those reference papers that cannot be downloaded automatically, we manually replenished them. It was a laborious work and took us hundreds of hours. Since some document sets cannot meet the requirements of automatic related work generation due to some references recognition errors and download problems, we filtered the document sets whose number of reference papers is less than 10 or whose number of missing reference papers is greater than 5 and the loss rate (the quotient of the number of missing reference papers divided by the number of all reference papers) exceeded 20%. In the end, we obtained 2,084 document sets.

4 Dataset Characteristics

As the first open dataset, NudtRwG has the following characteristics, which make it justified for related work generation.

Large scale. NudtRwG consists of 2,084 target papers and more than 52,000 reference papers. More detailed attributes are presented in Table 2. As can be seen, there are 25.3 reference papers, 8,572.6 sentences and 158,908.9 words per document set on average. Compared with previous work, NudtRwG has a larger scale.

Table 2. We use "#" to denote number. RWS stands for Related Work Section, RPs stands for Reference Papers.

	#(sentences in RWS)	#(words in RWS)	#(RPs)	#(sentences in RPs)	#(words in RPs)
average	24.9	496.4	25.3	8572.6	158908.9
stdev	14.1	289.9	10.8	4553.9	78803.8
min	3	101	5	641	15636
max	59	1180	96	45029	740710

High quality. In our dataset, target papers are all selected from well-received conferences of computational linguistics and natural language processing, such as ACL, EMNLP, NAACL and COLING. These high-quality paper sources make sure the quality of the ground truth related work. For further proof, we investigate the citation number of these target papers. Statistics in Figure 2 shows that, 74.67% of the target papers are cited more than 10 times, indicating that these target papers are widely recognized from perspective of academic community and therefore a high-quality related work section is expected.

High coverage. Another statistic we have done is the integrity of reference papers of NudtRwG. The result is demonstrated in Figure 3. As we can see,

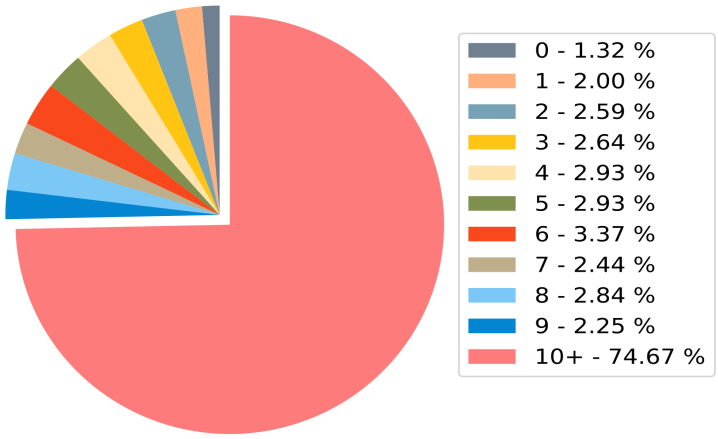

Fig. 2. Citation number distribution of papers in NudtRwG.

13.1% of the document sets cover all the reference papers of the reference list, and 93% of the document sets miss fewer than 6 reference papers. Only taking reference papers cited in related work section into consideration, 59.4% of the document sets contain all the reference papers and over 93% of the sets miss fewer than 3 reference papers. NudtRwG has a more complete list of reference papers for each document set, enabling related work generation task to be free from worrying about the absence of input data.

5 Analytical Study

Summative sections (abstract, introduction and conclusion) of reference papers were used as default input for related work generation in previous work [6, 8]. Notwithstanding, we doubt whether summative sections are sufficiently representative for the task. To investigate the effects of summative sections and other sections of reference papers on related work generation, we conduct the following analytical study on NudtRwG.

5.1 Analysis on Lexical Characteristics

We start with analyzing the lexical characteristics of summative sections and other sections of reference papers. The current ROUGE [11] oriented evaluation metric inspires us that, the N-gram overlaps of reference papers and the ground truth related work determines the upper bound quality of the generated related work. Therefore, we analyze the lexical characteristics by calculating N-gram overlaps between the ground truth related work and different sections of reference papers.

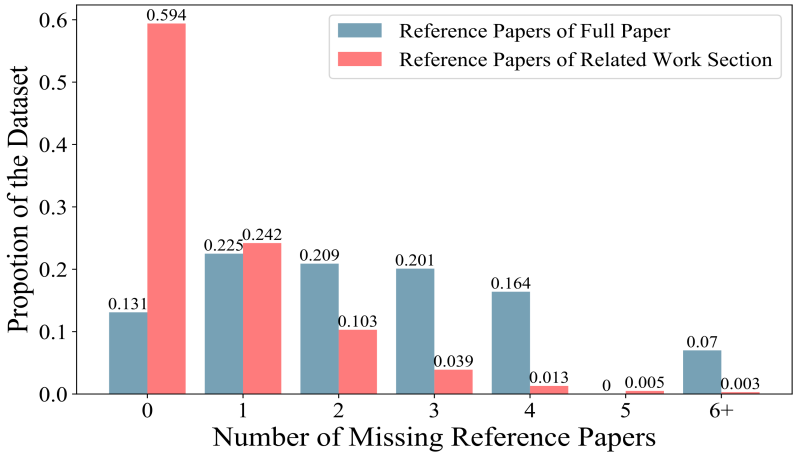

Fig. 3. The integrity of reference papers in full paper or related work section.

ROUGE-1, -2, -SU4 are used to evaluate the overlap score. In addition to reference papers, we also take into consideration contents of the target paper. RPs refers to only using reference papers as input and RPs+TP means supplying additional contents from target paper except the related work section.

Table 3 presents the result. Unsurprisingly, full texts of reference papers contain the most co-occurrence unigrams and bigrams of the ground truth related work, achieving 0.9085 and 0.4252 on ROUGE-1 score and ROUGE-2 score, respectively. Meawhile, adding extra information from target paper increases the ROUGE scores. The result indicates that the complete reference papers cover the most amount of information and information from target paper is indispensable. Second, summative sections are information-condensed parts of reference papers and they work as input in former researches [6, 8]. However, the ROUGE score of summative sections falls behind that of other sections, let alone full texts of reference papers, indicating taking advantage of full texts as input for related work generation should achieve a higher ROUGE score.

5.2 Analysis on Cited Text Spans

To further validate whether sections other than summative sections cover valuable information related to a ground truth, we introduce **C**ited **T**ext **S**pans (CTS), which refers to the fragments of text in the reference paper that most accurately reflect the citation [9]. Therefore, CTS can be considered as citation evidence and has been widely applied in citation-based scientific summarization

Table 3. Rouge results (%) of overlapping units between gold related work and different sections of reference papers and target paper. RPs denotes Reference Papers and TP denotes Target Paper

Contents	ROUGE-1		ROUGE-2		ROUGE-SU4	
	RPs	RPs+TP	RPs	RPs+TP	RPs	RPs+TP
Full Texts	90.85	92.42	42.52	47.02	52.78	56.52
Abstract	51.40	67.82	12.20	23.64	17.50	28.56
Introduction	79.40	81.88	27.96	33.26	36.70	40.28
Conclusion	39.20	60.21	8.20	19.86	12.58	24.58
Abstract+Introduction	80.79	83.01	29.50	34.98	38.25	42.13
Abstract+Introduction+Conclusion	81.43	83.87	30.10	35.32	38.90	42.78
Other Sections	89.19	90.12	38.90	39.83	49.20	50.10

[2, 20]. Here, we utilize CTS to locat incoming citations in reference papers. An obvious section distribution of CTS is expected via this investigation.

We artificially select 150 citations from related work section of target papers in NudtRwG and manually mark CTS of the given citations in corresponding reference papers. The annotation rule complies with that of TAC (Text Analysis Conference) 2014 Biomedical Summarization track[4]. We split sections of a paper into abstract, introduction, conclusion and other sections(method, experiment). One example of citation and CTS is shown in Figure 4.

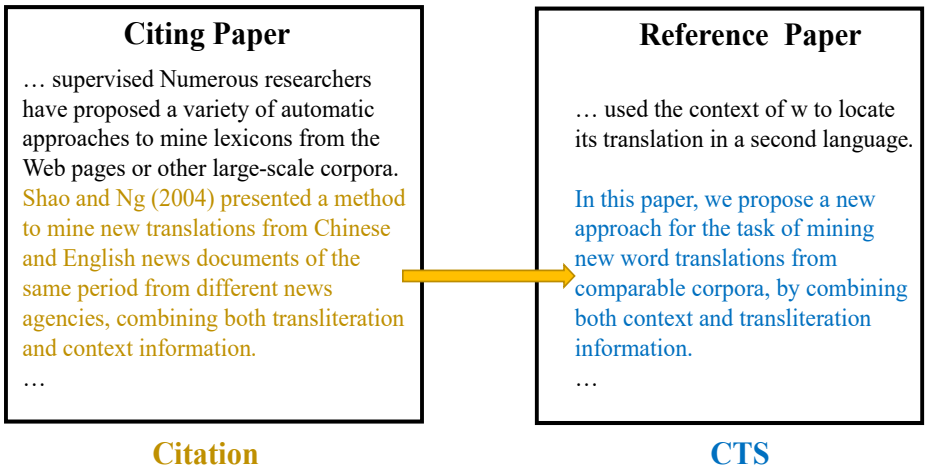

Fig. 4. An example of citation and its corresponding CTS.

The section distribution of CTS is shown in Table 4. We can find that abstract and introduction are important citation sources, with approximately three fifths

[4] https://tac.nist.gov//2014/BiomedSumm/index.html

of CTS are selected from these sections. In contrast, conclusion is less important in terms of the citation evidence. The statistics, along with former analysis,

Table 4. Section distribution of CTS in reference papers for citations.

Section Type	Abstract	Introduction	Conclusion	Other sections
Citations Number	20	66	4	60

demonstrate why previous researches prefer these summative sections as input. However, we can also see that the remaining two fifths of CTS are selected from other sections. More detailed statistical result is, 36 CTS are from method-relevant sections and 16 CTS are from experiment-relevant sections. The result indicates that, full texts of reference papers are indispensable for related work generation, not just the summative sections.

5.3 Experiment on Extractive Models

Next, we conduct experiments on current extractive summarization methods to investigate the influence of different sections on related work generation. The generated related work summaries are truncated to the same length of the ground truth.

We implement five extractive models, including:

Lexrank: Lexrank [3] is a graph-based summary approach inspired by Pagerank. A similarity graph $G(V, E)$ is constructed where V and E are the set of sentences and edges, respectively. An edge e_{ij} is drawn between sentence v_i and v_j if and only if the cosine similarity between them is above a given threshold. Sentences are scored according to their Pagerank score in G.

Sumbasic: Sumbasic [13] is a frequency-based summarizer. Each candidate sentence S is assigned a score $Score(S)$ reflecting how many high-frequency words it contains, where $Score(S)$ is calculated as an average of unigram probabilities of words of sentence S.

ICSI: ICSI [4] is a global linear optimization framework that has been identified as one of the state-of-the-art by [7]. It extracts a summary by solving a maximum coverage problem considering the most frequent bigrams in the source documents.

JS-Gen: JS-Gen [14] presents an optimization framework for extractive multi-document summarization. It optimizes JS divergence with a genetic algorithm.

TopicSum: TopicSum [5] is a generative probabilistic model. It is a hierarchical LDA style model and presumes that each word is generated by a single topic which can be a corpus-wide background distribution over common words, a distribution of documen-specific words or a distribution of the core content of a given cluster.

Table 5. Rouge results (%) of the generated related work of different models using full texts as input and summative sections as input, respectively.

Models	Full Texts			Abstract+Introduction+Conclusion		
	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-1	ROUGE-2	ROUGE-SU4
Lexrank	39.55	7.61	14.91	39.19	7.62	14.52
Sumbasic	38.03	6.00	13.36	38.08	6.17	13.34
TopicSum	38.98	6.79	14.11	38.73	6.35	13.95
ICSI	**40.33**	**8.51**	**15.41**	40.13	8.47	15.12
JS-Gen	38.05	6.45	13.52	38.08	6.67	13.52

We take full texts and summative sections as input, respectively. Table 5 reports the evaluation over ROUGE metric.

From the table, the performance with summative sections as input is comparable to that with full texts as input. One possible reason is, while full texts of reference papers cover more information relevant to a target paper, they inevitably carry more redundant and confusing information than summative sections. In addition, the extractive models we select are suitable for general multi-document summarization, they may be incapable of identifying target paper-relevant sentences in other sections. The same drawback shows up in [6, 8]. The authors concentrate on summative sections and therefore ignore valuable information in other sections.

5.4 Set out with Full Texts

Considering that current general summarization approaches have difficulty in distinguishing salient and relevant sentences in full texts of reference papers, a reasonable suggestion is taking advantage of citation sentences which cite reference papers to locate salient information in reference papers [17, 2]. Such method has been extensively used in scientific summarization [15, 16] and survey generation [12, 10]. Additionally, CTS-based summarization can be considered for related work generation, as they provide more detailed and precise information about reference papers than citations alone. They may help to mark valuable information in full texts according to viewpoint from academic community.

Another suggestion is to model content relevance of target paper and reference papers in an efficient way. A feasible way is to utilize generative probabilistic models to catch target paper-relevant contents in reference papers. Furthermore, abstractive approaches are also encouraged for related work generation, as former discussion indicates full texts of reference papers contain almost all of the unigrams in a ground truth related work.

6 Conclusion

Towards the goal of automatic related work generation, we construct the NudtRwG dataset, a collection of 2084 document sets. Based on NudtRwG,

we conduct an analytical study on the effects of summative sections (abstract, introduction and conclusion) and other sections of reference papers on related work generation. We find, different from previous researches, other sections apart from summative sections are also of vital importance for related work generation. What really matters is how to identify those target paper-relevant and salient information throughout full texts.

NudtRwG is the first open, large-scale and high-quality dataset for related work generation. We have made our dataset freely available to encourage the research of related work generation. At the same time, we hope our analyses on this task will enlighten more expressive models.

Acknowledgements

The research is supported by the National Key Research and Development Program of China (2018YFB1004502) and the National Natural Science Foundation of China (61532001, 61303190)

References

1. Chen, J., Zhuge, H.: Automatic generation of related work through summarizing citations. Concurrency and Computation: Practice and Experience **31**(3), e4261 (2016)
2. Cohan, A., Goharian, N.: Scientific document summarization via citation contextualization and scientific discourse. International Journal on Digital Libraries **19**(2-3), 287–303 (2018)
3. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004)
4. Gillick, D., Favre, B.: A scalable global model for summarization. In: Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing. pp. 10–18. Association for Computational Linguistics (2009)
5. Haghighi, A., Vanderwende, L.: Exploring content models for multi-document summarization. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 362–370. Association for Computational Linguistics (2009)
6. Hoang, C.D.V., Kan, M.Y.: Towards automated related work summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 427–435. Association for Computational Linguistics (2010)
7. Hong, K., Conroy, J.M., Favre, B., Kulesza, A., Lin, H., Nenkova, A.: A repository of state of the art and competitive baseline summaries for generic news summarization. In: LREC. pp. 1608–1616 (2014)
8. Hu, Y., Wan, X.: Automatic generation of related work sections in scientific papers: an optimization approach. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1624–1633 (2014)
9. Jaidka, K., Chandrasekaran, M.K., Elizalde, B.F., Jha, R., Jones, C., Kan, M.Y., Khanna, A., Molla-Aliod, D., Radev, D.R., Ronzano, F., et al.: The computational linguistics summarization pilot task (2014)

10. Jha, R., Finegan-Dollak, C., King, B., Coke, R., Radev, D.: Content models for survey generation: a factoid-based evaluation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). vol. 1, pp. 441–450 (2015)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
12. Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., Zajic, D.: Using citations to generate surveys of scientific paradigms. In: Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics. pp. 584–592. Association for Computational Linguistics (2009)
13. Nenkova, A., Vanderwende, L.: The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005 **101** (2005)
14. Peyrard, M., Eckle-Kohler, J.: A general optimization framework for multi-document summarization using genetic algorithms and swarm intelligence. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 247–257 (2016)
15. 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)
16. Qazvinian, V., Radev, D.R., Mohammad, S.M., Dorr, B., Zajic, D., Whidby, M., Moon, T.: Generating extractive summaries of scientific paradigms. Journal of Artificial Intelligence Research **46**, 165–201 (2013)
17. Wang, P., Li, S., Wang, T., Zhou, H., Tang, J.: Nudt@ clscisumm-18. In: BIRNDL@ SIGIR. pp. 102–113 (2018)
18. Wang, Y., Liu, X., Gao, Z.: Neural related work summarization with a joint context-driven attention mechanism. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1776–1786 (2018)
19. Widyantoro, D.H., Amin, I.: Citation sentence identification and classification for related work summarization. In: 2014 International Conference on Advanced Computer Science and Information System. pp. 291–296. IEEE (2014)
20. Yasunaga, M., Kasai, J., Zhang, R., Dan, A.R.F.I.L., Radev, F.D.R.: Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks (2019)