Semantic Graph Based Automatic Summarization of Multiple Related Work Sections of Scientific Articles

Nouf Ibrahim Altmami(✉) and Mohamed El Bachir Menai

Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia
naltmami@su.edu.sa, menai@ksu.edu.sa

Abstract. The summarization of scientific articles and particularly their related work sections would support the researchers in their investigation by allowing them to summarize a large number of articles. Scientific articles differ from generic text due to their specific structure and inclusion of citation sentences. Related work sections of scientific articles generally describe the most important facts of prior related work. Automatically summarizing these sections would support research development by speeding up the research process and consequently enhancing research quality. However, these sections may overlap syntactically and semantically. This research proposes to explore the automatic summarization of multiple related work sections. More specifically, the research goals of this work are to reduce the redundancy of citation sentences and enhance the readability of the generated summary by investigating a semantic graph-based approach and cross-document structure theory. These approaches have proven successful in the field of abstractive document summarization.

Keywords: Automatic text summarization · Scientific article · Related work Multi-document · Semantic graph · Cross-document structure theory

1 Introduction

Automatic summarization of scientific articles would be useful for researchers to quickly study and evaluate the sate-of-the-art. However, the most recent articles refer to the same related work and hence a large number of articles is cited in every related work section. Automatically summarizing multiple related work sections would be useful and helpful by reducing the time needed to review a large number of related work sections.

Related work sections have specific characteristics that make them unique. First, they include citation sentences. Second, these sections are short in length, which makes the problem more challenging. Hence, extractive techniques would generate a summary suffering from a lack of readability and coherence. Finally, the overlap between multiple related work sections is an important issue.

Limited research studies addressed related work summarization. Most of these studies have generated a related work section for a target paper by summarizing a set of articles [1–3]. Only one study has tackled the problem of summarization of the related

G. Agre et al. (Eds.): AIMSA 2018, LNAI 11089, pp. 255–259, 2018.
https://doi.org/10.1007/978-3-319-99344-7_23

work section of a single article [4]. Automatically summarizing multiple related work sections of a set of articles on a particular topic would aid the generation of a more comprehensive summary. To the best of our knowledge, no previous research has tackled this particular problem. This research work proposes to address this problem by investigating a semantic graph-based approach and cross-document structure theory (CST).

The remaining of this paper is as follows. Section 2 examines various prior studies in the field of scientific article summarization. The proposed approach is presented in Sect. 3. Finally, Sect. 4 concluded this paper.

2 Related Work

Among the interesting concerns of scientific articles summarization is the generation of research article abstract. Lloret et al. [5] suggested two approaches for this task. The first one is an extractive summarization approach. The second one is based on both extractive and abstractive techniques. Saggion and Lapalme [6] proposed an approach for generating an indicative and informative abstract called Selective Analysis. This type of summarization is not an accurate scientific summary since it stated the contributions in a less focused fashion and general form.

The above-mentioned problems motivated the generation of citation based summaries. Abu-Jbara and Radev [7] tackled some issues related to the readability and coherency of this type of summaries. C-LexRank, a graph based summarizer, is also proposed by Qazvinian and Radev [8] to summarize single scientific article. Chen and Zhuge [9] made additional progress by exploiting a set of terms that co-occur in a set of citations according to the common fact phenomenon.

Related work summarization is a specific instance of scientific article summarization. Hoang and Kan [1] proposed a heuristic system called ReWoS for the automatic generation of a related work section based on a topic hierarchy tree. Chen and Zhuge [2] used citation sentences and performed a comparison of the content of the target article and the content of the citation sentences while Hu and Wan [3] considered this problem as an optimization problem. Widyantoro and Amin [4] proposed a different approach for summarizing a related work section in scientific articles. Their proposed approach consists of two main stages. First, they extracted citation sentences. Then, they categorized these citation sentences into three different classes (i.e., problem, method and conclusion).

3 The Proposed Approach

Our goal is to automatically summarize multiple related work sections while maximizing the readability of the generated summary and minimizing the redundancy of citation sentences. We propose to investigate a hybrid method based on both a semantic graph-based approach and CST. Moreover, different abstractive techniques will be investigated to improve the readability, including multi-sentence compression [10] and language generation [11].

Positive feedback has been obtained when using graph-based approaches in the field of (MDS) [11–13]. However, it suffers from some essential limitations. First, it depends on similarity measures without taking into consideration the semantic relationships among sentences. A second limitation is the lack of diversity of the generated summary due to the ranking algorithms. Thus, we plan to investigate the use of a semantic graph-based approach to cope with the redundancy of citation sentences. Moreover, we will investigate ranking algorithms to take into consideration the semantic similarity. In the other hand, CST has been used to analyze multi-documents to discover semantic relations among their content [14–16]. Based on the particular content of the related work section, CST could help to further reduce redundancy between citation sentences. Different content selection methods will be investigated, including a redundancy operator, general operator [17] and the method proposed by Otterbacher et al. [18]. Following is a small instance of the problem to illustrate the proposed approach.

A part of the related work section of paper [1]:	*A part of the related work section of paper* [2]:
"Further, Mei and Zhai (2008) and Qazvinian and Radev (2008) utilized citation information in creating summaries for a single scientific article in computational linguistics domain."	"Based on the finding, Qazvinian and Radev employ the citations to create the summary for the scientific paper [3, 5]."

Reference Qazvinian and Radev (2008) in paper [1] is cited as [5] in paper [2] and the two text spans have the same information content. Thus, the result of the proposed approach should be similar to:

Mei and Zhai [1] *utilized citation information in creating summaries for a single scientific article in computational linguistics domain. Qazvinian and Radev* [4, 6] *employed the citations to create the summary for the scientific paper.*

The main steps of the proposed approach are:

- A preprocessing step to identify the same reference in each related work section and to reduce them to one format for example IEEE format.
- A Graph: to represent the relations between the references and their citation sentences.
- CST to analyze the different citation sentences of the same reference in order to discover semantic relations among their content.
- Content selection: the final step is summary extraction by transforming the graph into smaller one while preserving its properties.

The main objectives of this research are summarized in the following points:

- Summarizing multiple related work sections of scientific articles while enhancing the readability of the generated summary and minimizing the redundancy of citation sentences.
- Proposing a hybrid method based on both semantic graph based approach and CST.

- Finding the semantic relationships between different contents in order to not influence the discourse meaning.
- Examining different abstractive techniques to hopefully improve the readability.
- Building our own dataset for the summarization of related work sections. According to our first investigation, we have not found a benchmark dataset available online for the summarization of related work sections. However, we have been able to obtain the data set used in [4] to evaluate summaries of related work sections. This dataset is composed of a collection of 20 article sets, and each set contains different reference articles that need to be summarized to generate a related work section.

4 Conclusion

In this paper, we took the first step towards summarizing multiple related work sections of scientific articles. We outlined a hybrid approach which consists of combining semantic graphs and CST. It aims at minimizing the redundancy of citation sentences and improving the readability of the generated summary by investigating different abstractive and content selection techniques.

References

1. Hoang, C.D.V., Kan, M.-Y.: Towards automated related work summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters 2010, pp. 427–435 (2010)
2. Chen, J., Zhuge, H.: Summarization of related work through citations. In: 12th International Conference on Semantics, Knowledge and Grids (SKG) 2016, pp. 54–61. IEEE (2016)
3. Hu, Y., Wan, X.: Automatic generation of related work sections in scientific papers: an optimization approach. In: EMNLP 2014, pp. 1624–1633 (2014)
4. Widyantoro, D.H., Amin, I.: Citation sentence identification and classification for related work summarization. In: International Conference on Advanced Computer Science and Information Systems (ICACSIS) 2014, pp. 291–296 (2014)
5. Lloret, E., Romá-Ferri, M.T., Palomar, M.: COMPENDIUM: a text summarization system for generating abstracts of research papers. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 3–14. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22327-3_2
6. Saggion, H., Lapalme, G.: Selective analysis for automatic abstracting: evaluating indicativeness and acceptability. In: Content-Based Multimedia Information Access-Volume 1, pp. 747–764 (2000)
7. Abu-Jbara, A., Radev, D.: Coherent citation-based summarization of scientific papers. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 500–509 (2011)
8. 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 (2008)
9. Chen, J., Zhuge, H.: Summarization of scientific documents by detecting common facts in citations. Future Gener. Comput. Syst. **32**, 246–252 (2014)

10. Banerjee, S., Mitra, P., Sugiyama, K.: Multi-document abstractive summarization using ILP based multi-sentence compression. arXiv preprint arXiv:1609.07034 (2016)
11. Atif, K., Salim, N., Kumar, Y.: Genetic semantic graph approach for multi-document abstractive summarization. In: Fifth International Conference on Digital Information Processing and Communications (ICDIPC) 2015. IEEE (2015)
12. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)
13. Ganesan, K., Zhai, C., Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd International Conference on Computational Linguistics 2010, pp. 340–348 (2010)
14. Zhang, Z., Otterbacher, J., Radev, D.: Learning cross-document structural relationships using boosting. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management 2003, pp. 124–130 (2003)
15. del Rosario Castro Jorge, M.L., Pardo, T.A.S.: Experiments with CST-based multidocument summarization. In: Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing 2010, pp. 74–82 (2010)
16. Zhang, Z., Blair-Goldensohn, S., Radev, D.R.: Towards CST-enhanced summarization. In: AAAI/IAAI (2002)
17. Radev, D.R.: A common theory of information fusion from multiple text sources step one: cross-document structure. In: Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue-Volume 10, pp. 74–83 (2000)
18. Otterbacher, J.C., Radev, D.R., Luo, A.: Revisions that improve cohesion in multi-document summaries: a preliminary study. In: Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4, pp. 27–36 (2002)