Citation Based Collaborative Summarization of Scientific Publications by a New Sentence Similarity Measure

Chengzhe Yuan, Dingding Li(✉), Jia Zhu, Yong Tang,
Shahbaz Wasti, Chaobo He, Hai Liu, and Ronghua Lin

School of Computer Science, South China Normal University,
Guangzhou 510631, Guangdong, China
dingdingli@m.scnu.edu.cn

Abstract. Next-generation network offers unrestricted access for researchers to all kinds of scientific publications, collaborative summarization systems are now being contemplated as a service that can help researchers gain information when they read scientific articles. One way to develop a collaborative summarization system is to measure semantic similarity between sentences to improve its quality. In this paper, we introduce a new sentence similarity measure for summarizing scientific articles with citation context. Our work is based on recent work in document distance metric called the word mover's distance (WMD). Compared to traditional similarity measures, WMD based sentence similarity measure has better performance by capturing the semantic relation between two sentences. Experiments on 2016 version of ACL Anthology Reference Corpus show that our approach outperforms several other baselines by ROUGE metrics.

Keywords: Collaborative summarization · Sentence similarity measure
Citation context

1 Introduction

The next-generation network demands for more collaborative academic services i.e. collaborative text summarizations. The amount of research is being published in different fields of life especially in science has made it more difficult for researchers to up to date their interests. To collaborate their new ideas with existing research, they spend a lot of time in reading scientific articles and making collaborative summaries. These summaries on scientific articles can facilitate researchers to capture the salient ideas of an article more quickly and effortlessly.

There are two types of methods in scientific publications summarization [1]: (1) Abstractive summarization, which attempts to generate novel sentences for summary. A common difficulty with this method is *de novo* meaningful and grammatical summary [2]. (2) Extractive summarization or collaborative summarization extracts key sentences or phrases from source documents and group them into shorter form [3].

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 680–689, 2018.
https://doi.org/10.1007/978-3-030-00916-8_62

According to Elkiss et al. citation contexts usually cover the cited paper in all scenarios such as the research problem, the proposed method, shortcomings and limitations [4]. So different from general free-text extractive summaries, citation based summarization is the set of citing sentences for a given article and this summary is the collaborative summation of other scholars' viewpoints.

Previous work on citation based summarization mainly focused on how to make citation sentences useful. For example, extracting citation context from cited article instead of original article [5] or summarizing articles by detecting common facts in citations [6]. However, the accurate measure of semantic similarity between citation sentences in the process of summarization has been mostly ignored. The absence of semantic similarity measure from summarization process will group the sentences into wrong clusters, this resulted in poor performance for final summarization.

In this paper, we present a WMD [7] based sentence similarity measure to address the aforementioned shortcomings of existing summarization models. Compare to traditional measure of the similarity between two sentences, i.e. (term frequency- inverse document frequency) TF-IDF, (bag of words) BOW, WMD metric is capable of capturing the semantic relation between two sentences by utilizing the high quality of the *word2vec* embedding. Our model generates citation based collaborative summaries in four steps: preprocessing, classifying sentences according to article's discourse structure, semantic clustering and selecting the sentences for final summarization using Maximal Marginal Relevance (MMR). Evaluation results on ACL2016 corpus show that our proposed model outperforms several baselines methods.

2 Related Work

Summarizing scientific publications on the basis of citations is well researched area. According to Elkiss citation-based summary contains more information than abstracts and main contexts [4].

Qazvinian proposed a clustering approach where communities in the citation summary's lexical network are formed and sentences are extracted from separate clusters [8]. Agarwal described a SciSumm system which summarizes document collections that are composed of fragments extracted from co-cited articles [9]. Another similar approach was presented in [6], this approach aims to summarize scientific documents by detecting common facts in citations. Amjad and Dragomir performed a study on how to produce readable summaries [10]. Ronzano and Saggion proposed a platform to automatically extract, enrich and characterize several structural and semantic aspects of scientific publications [11].

Different approaches for the extraction of citation sentences have also been proposed in the past years. Qazvinian proposed a citation based summarization approach which extract important key phrases from a set of citation sentences and build summary by these key phrases [12]. Cohan and Goharian generated summarization by extracting citation context from reference article. This approach overcomes the problem of inconsistency between the citation summary and the article's content by providing context for each citation. Article's inherent discourse model has also been proved helpful in extracting citation sentences [5].

However, above literature do not measure the semantic similarity between sentences in citation based summarization. By utilizing the high quality of the *word2vec* embedding, results from Kusner's work showed that WMD metric works better than any other document distance metrics [7]. Hence, we apply WMD metric to the process of citation based summarization for improving the summarization quality.

3 The Summarization Approach

The framework of our approach is shown in Fig. 1, we describe it in the following four steps:

(1) The Preprocessing step extracts citation sentences from scientific articles as input, then outputs citation semantic similarity between sentences in matrix form by WMD metric.
(2) In the sentences classification step, we use one-vs-rest SVM model with linear kernel to classify sentences into four categories: Introduction, Methods, Results, and Discussion (IMRAD) [13] which is respect to article's discourse structure.
(3) The semantic clustering step takes classified sentences as input, then performs k-means clustering algorithm to cluster sentences into groups by utilizing semantic similarity matrix obtained in step 1.
(4) In the selection of sentences for summary step, we score every sentences in each cluster by MMR strategy (as described in Sect. 3.3) and top ranked sentences are iteratively selected from each IMARD categories until we reach the summary length threshold. Finally, generate the final summary from the selected sentences.

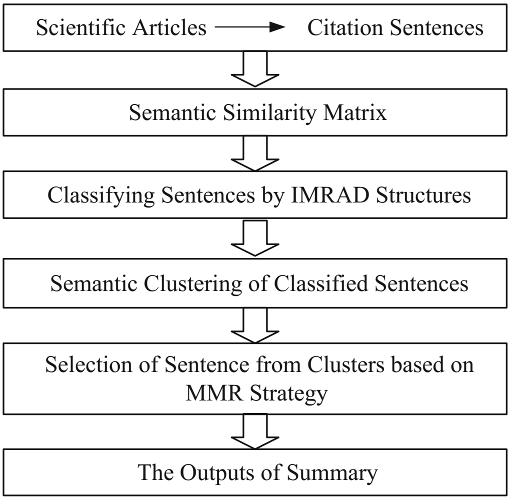
Scientific Articles
Citation Sentences
Semantic Similarity Matrix
Classifying Sentences by IMRAD Structures
Semantic Clustering of Classified Sentences
Selection of Sentence from Clusters based on MMR Strategy
The Outputs of Summary

Fig. 1. The framework of proposed approach

3.1 Sentence Similarity Measure

To get the citation sentences, firstly, we need to extract citation context from scientific article and then identify the scope of citation sentence by annotated reference maker, lastly, find fragments of the sentence which are related to given target reference.

There are two main sentence similarity measures: (1) word overlap measure that compute similarity score based on a number of words shared by two sentences. (2) TF-IDF measure that compute sentence similarity based on term frequency-inverse document frequency [14]. The main drawback of these measures is their failure on capturing the semantic relation between sentences. We employ WMD metric to measure semantic similarity between sentences, WMD metric is defined by the distance between sentences in terms of the vector embedding of the words that make up the sentences [7]. Before implementing WMD metric, we get word-vectors that are semantically synonymous in the embedding space by *word2vec* [15], which is a two-layer neural net that are trained to reconstruct linguistic contexts of words. The main idea of WMD-based sentence similarity (WMD-Similarity) measure is to minimize the amount of distance that "transports" from the embedded words of one sentence to another sentence. WMD-Similarity is defined as follows:

$$D(x_i, x_j) = \min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} \left\| x_i - x_j \right\|_2^p, \text{subject to: } \sum_{j=1}^{n} T_{ij} = d_i^a, \sum_{i=1}^{n} T_{ij} = d_i^b, \forall i,j \quad (1)$$

Where $x_i \in R^d$ is the i^{th} word embedding matrix $X \in R^{d \times n} \cdot \mathbf{d^a}$ and $\mathbf{d^b}$ are the n-dimensional normalized bag-of-words vectors for two sentences, where d_i^a represents the number of word i occurs in $\mathbf{d^a} \cdot \boldsymbol{T}_{ij}$ denotes how much of d_i^a "travels" to d_j^b. The minimum cumulative cost of moving d_i^a to d_j^b can be taken as distance between sentences. The smaller the value of WMD-Similarity measure the higher the semantic similarity between two sentences. For example, we have calculated WMD-Similarity measure between sentences (a) and (b) that is 1.15. Our results show that even having very few common words between them still the sentences are semantically similar to each other. Which proves the efficiency of our proposed sentence similarity measure.

(a) *Other work focuses on lexical simplifications and substitutes difficult words by more common WordNet synonyms or paraphrases found in a predefined dictionary.*

(b) *The literature is rife with attempts to simplify text using mostly hand-crafted syntactic rules aimed at splitting long and complicated sentences into several simpler ones.*

3.2 Sentences Classification and Clustering

When researchers writing scientific articles, they generally follow a standardized structure which is known as introduction, methods, results, and discussion (IMRAD). In order to correctly capture all aspects of the article, we include each category of IMARD in the summarization. We use one-vs-rest SVM model with linear kernel to classify citation sentences to their respective category. To train our classification model, we use all the verbs which are lemmatized from 4,685 annotated sentences (1,705 of Introduction, 1,310 of Methods, 1,391 of Results & 279 of Discussion) and they appear at least twice in the same class, besides auxiliary verbs are excluded.

Clustering Model. After classifying each citation sentence. We apply a two-steps cluster method to divide citation sentences that contain similar information into groups of same topic. In first step, hierarchical clustering is assigned for grouping data into subsets. In second step, each formed subsets in first step will be taken as the input data for k-means clustering. Hierarchical clustering is used for better performance of K-means [16]. In two-step cluster method, the measure of distance between two sentences is weighted by semantic similarity matrix obtained from Sect. 3.1.

3.3 Ranking Model

In previous step, we clustered similar citation sentences into groups. But not all sentences in the same cluster are equally important, for example both sentence (c) and (d) mentioned author Søgaard's work, however, sentence (c) is more important, because it contains more useful information.

(c) *For example, recent work by Søgaard explores data set sub-sampling methods.*
(d) *This idea has been previously explored by Zeman and Resnik and recently by Søgaard.*

The goal of this module is to extract the most representative sentences from each classification. There are various ways of ranking sentences based on their importance, we use a well-known method Maximal Marginal Relevance (MMR) [17] for evaluation. MMR is a linear combination of relevance and novelty scores to rank sentences. The MMR-based ranking score for a sentence are defined by following formula:

$$score(S) \overset{def}{=} \lambda Sim_1(S,R) - (1-\lambda)Sim_2(S,Sum) \tag{2}$$

Where $Sim_1(S,R)$ represents the linear interpolation of similarity of sentence S with all other sentences, $Sim_2(S,Sum)$ is the similarity between sentence S and the sentences already in the summary. $Sim_1(\bullet)$ and $Sim_2(\bullet)$ are WMD-Similarity from formula (1), and we empirically set $\lambda = 0.7$.

4 Experiments

4.1 Dataset

We use ACL Anthology Reference Corpus[1] (ARC) which is a collection of scholarly publications about computational linguistics. It includes all ACL Anthology files up to December 2015, consisting of 22,878 articles. ARC provides all logical document structure and parsed citation information for each article. We randomly select 556

[1] http://acl-arc.comp.nus.edu.sg/.

articles from ARC as our data, each article contains more than 20 citation sentences. After cleaning low-quality sentences (contains less than 5 words) and noise data, we get 17,186 citation sentences in the dataset. We asked two experts in NLP domain read citation sentences and its corresponding cited paper, then they manually create two sets of scientific summary for 10 selected articles, the short summaries of 4 sentences (80–100 words) and longer summaries of 8 sentences (230–250 words).

Evaluation Metrics. We use ROUGE, one of the most popular automatic evaluation metric for evaluation. It automatically measures the quality of a summary by comparing overlapping units such as n-gram, word sequences, and word pairs with ideal summaries created by humans [18]. Specifically, we use ROUGE-N (N-gram Co-Occurrence Statistics) metric for our evaluation. ROUGE-N is defined as follows:

$$\text{ROUGE - N} = \frac{\sum\limits_{S\in\{ReferenceSummaries\}} \sum\limits_{W\in S} f_{match}(W)}{\sum\limits_{S\in\{ReferenceSummaries\}} \sum\limits_{W\in S} f(W)} \tag{3}$$

Where W is the n-gram, $f(\bullet)$ is the count function, $f_{match}(\bullet)$ is the maximum number of n-grams co-occurring in the generated summary and a set of reference summaries, and ROUGE-topic is a novel metric for measuring the topical relation between two documents, this metric is well illustrated in [14].

4.2 Baselines

To evaluate the performance of our approach in citation based summarization, we conducted experiments with 5 widely-used baseline approaches.

- *KL-Sum*. KL-Sum greedily adds sentences to a summary as long as it minimizes the Kullback–Leibler Divergence (KLD). Where KLD is a measure of ‘closeness’ between probability distribution of two documents [19].
- *LexRank*. In this approach, it computes sentence importance based on the concept of eigenvector centrality in a graph representation of sentences [20].
- *LSA*. By capturing main topics of a document, the sentences with most important topics are selected for the summary [21].

Table 1. ROUGE-1, ROUGE-2 and ROUGE-topic recall scores for different approaches.

Metric	Our System	KL	LexRank	LSA	SumBasic	TextRank
Rouge-1–4	**0.556**	0.456	0.507	0.540	0.359	0.485
Rouge-2–4	**0.447**	0.316	0.363	0.409	0.158	0.342
Rouge-topic-4	**0.463**	0.313	0.387	0.432	0.211	0.366
Rouge-1–8	**0.697**	0.576	0.577	0.625	0.571	0.646
Rouge-2–8	**0.626**	0.454	0.449	0.512	0.407	0.525
Rouge-topic-8	**0.618**	0.454	0.475	0.540	0.454	0.553

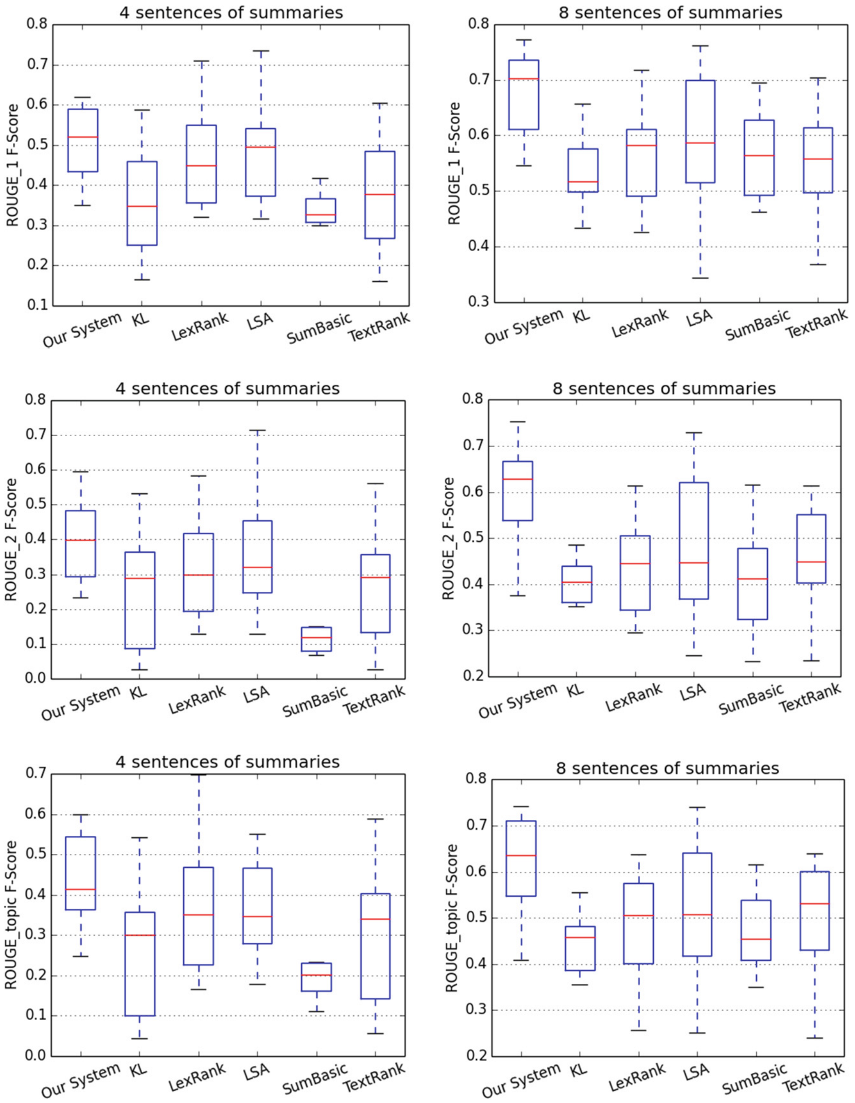

Fig. 2. The comparison of different approaches on Rouge-1, Rouge-2 and Rouge-topic F-scores.

- *TextRank.* It is a graph-based ranking model for sentence extraction and summarization using Levenshtein distance as relation between text units [22].
- *SumBasic.* SumBasic is a generic summarization system that exploits frequency information exclusively [23].

5 Results and Discussion

In this section, we present the results of our method with other baselines in ROUGE scores. The citation context for all the methods were removed numeric values, stop words and citation markers. We compute the ROUGE scores based on 2 * 10 gold standard summaries.

Table 1 shows the average ROUGE recall scores for all summarization approaches of 10 articles in dataset, metric “Rouge-*-4” and “Rouge-*-8” mean ROUGE scores on short summaries (4 sentences) and long summaries (8 sentences).

From Table 1 and Fig. 2, it is clear that semantic similarity measure based models (System & LSA) performed better than TF-IDF based models. With WMD based sentence similarity model, our system achieves better results on all the ROUGE metrics in both short and long summaries and has improved the performance in summarizing scientific articles. In longer summaries, the performance gap is even wider between our approach and others.

Among baseline approaches, KL and SumBasic performed below average, due to their failure on selecting optimal sentences for summarization. However, LSA performed slightly better than baseliners. LexRank and TextRank showed almost similar performance. The main idea behind these two unsupervised approaches is to find sentences which are very similar to each other therefore diversity in summarization is not considered in both approaches. Our approach can address this problem by selecting important sentences from IMARD category of the article. We attribute the competitive performance of our approach to its accurate measure of semantic similarity between sentences which improve the quality of clustering and ranking model in the process of summarization.

6 Conclusion

In this paper, we proposed a four steps approach based on WMD-similarity to measure semantic similarity between sentences for scientific article summarization. This will generate a semantically collaborative summary. The first step pre-processes citation sentences and obtain semantic similarity between sentences by WMD-similarity metric. In the second step, we classify the clusters based on article’s discourse structure: IMARD. In the third step, we employ a two-step clustering algorithm by semantic similarity matrix obtained in step 1. In the last step, we rank citation sentences within four categories by MMR strategy. Our experiments show that our proposed approach effectively achieved improvement over several baselines. In future, we will expand the application of our approach to other types of publications.

Acknowledgements. This work was funded by the National Natural Science Foundation of China under grant number 61772211 and 61502180, by the Natural Science Foundation of Guangdong Province, China under grant number 2014A030310238, 2015B010129009, 2016A030313441 and 2017A030303074, by the Special Fund for Key Program of Science and Technology of Guangdong Province, China under grant number 2014B010116002, by the Pearl River S&T Nova Program of Guangzhou under grant number 201710010189 and 201604016007.

References

1. Nenkova, A., McKeown, K.: Automatic summarization. Found Trends® Inf. Retr. **5**(2–3), 103–233 (2011)
2. Khan, A., Salim, N.: A review on abstractive summarization methods. J. Theor. Appl. Inf. Technol. **59**(1), 64–72 (2014)
3. Gupta, V., Lehal, G.S.: A survey of text summarization extractive techniques. J. Emerg. Technol. Web Intell. **2**(3), 258–268 (2010)
4. Elkiss, A., et al.: Blind men and elephants: what do citation summaries tell us about a research article? J. Assoc. Inf. Sci. Technol. **59**(1), 51–62 (2008)
5. Cohan, A., Goharian, N.: Scientific article summarization using citation-context and article's discourse structure (2017). ArXiv preprint arXiv:1704.06619
6. Chen, J., Zhuge, H.: Summarization of scientific documents by detecting common facts in citations. Future Gener. Comput. Syst. **32**, 246–252 (2014)
7. Kusner, M., et al.: From word embeddings to document distances. In: International Conference on Machine Learning (2015)
8. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1. Association for Computational Linguistics (2008)
9. Agarwal, N., et al.: Towards multi-document summarization of scientific articles: making interesting comparisons with SciSumm. In: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages. Association for Computational Linguistics (2011)
10. 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, vol. 1. Association for Computational Linguistics (2011)
11. Ronzano, F., Saggion, H.: Knowledge extraction and modeling from scientific publications. In: González-Beltrán, A., Osborne, F., Peroni, S. (eds.) SAVE-SD 2016. LNCS, vol. 9792, pp. 11–25. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-53637-8_2
12. Qazvinian, V., Radev, D.R., Özgür, A.: Citation summarization through keyphrase extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics (2010)
13. Sollaci, L.B., Pereira, M.G.: The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. J. Med. Library Assoc. **92**(3), 364 (2004)
14. Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 305–316. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85836-2_29
15. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method (2014). ArXiv preprint arXiv:1402.3722
16. Lu, J.-F., et al.: Hierarchical initialization approach for K-Means clustering. Pattern Recognit. Lett. **29**(6), 787–795 (2008)
17. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (1998)
18. Lin, C.-Y., Och, F.: Looking for a few good metrics: ROUGE and its evaluation. In: NTCIR Workshop (2004)

19. 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. Association for Computational Linguistics (2009)
20. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)
21. Steinberger, J., Křišťan, M.: LSA-based multi-document summarization. In: Proceedings of 8th International Workshop on Systems and Control (2007)
22. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: EMNLP (2004)
23. Nenkova, A., Vanderwende, L.: The impact of frequency on summarization. Technical report MSR-TR-2005-101. Microsoft Research, Redmond, Washington (2005)