Query-focused Scientific Paper Summarization with Localized Sentence Representation

Kazutoshi Shinoda[1] and Akiko Aizawa[2]

[1] The University of Tokyo
shinoda@nii.ac.jp
[2] National Institute of Informatics, The University of Tokyo
aizawa@nii.ac.jp

Abstract. Query-focused summarization enables the extraction of information on a specific "aspect" such as a "proposed approach" or "dataset used." Accordingly, it can be a powerful tool for reviewing scientific literature. In this paper, we present a method for unsupervised extractive query-focused summarization. In our approach, we first calculate word importance scores for each target document using a word-level random walk. Next, we optimize sentence embedding vectors using a Siamese neural network. Here, we utilize localized sentence representations obtained as the weighted average of word embeddings where the weights are determined by the word importance scores. Then, we conduct a sentence-level query-biased random walk to select a sentence to be used as a summary. In our experiments, we constructed a new evaluation dataset for query-focused summarization of scientific papers and showed that our method achieves competitive performance compared to other sentence embeddings.

Keywords: scientific paper mining · extractive summarization · sentence representation

1 Introduction

Query-focused extractive summarization is one of the key technologies needed for finding useful information from a large number of scientific papers [1]. In this paper, we hypothesize that localized sentence representation is beneficial in such summarization tasks where identifying important sentences is crucial. Accordingly, we propose a method for localized sentence representation whereby the importance score of each word in a document is considered in the calculation. The final score of each sentence is determined by considering both the importance of the sentence and its relatedness to a given query.

Considering the diversity and high specificity of scientific papers, we use a graph-based method that, unlike neural networks, does not require a large training dataset for both word- and sentence-level importance calculations. Moreover, we specifically focus on extractive summarization because for scientific knowledge, it is important to preserve the authors' original expressions or opinions in

the generated summaries. The contributions of the present study are summarized as follows: First, we propose an unsupervised method for learning localized sentence representations that takes document-wise word importance into account. Second, we constructed an additional dataset to evaluate query-focused summarization of scientific papers. Third, through our preliminary study, we showed that the sentence representations obtained using the proposed method achieves comparable performance with existing ones.

2 Related Work

Both supervised and unsupervised methods are used for extractive summarization. Among supervised methods, neural models are the mainstream of current research [2, 3]. Among unsupervised methods, conventional graph-based methods are commonly used [4, 5, 17, 19].

Regardless of whether the learning method is supervised or not, sentence representations are generally used as inputs to summarization systems. In most recent studies, sentence representations are calculated based on the embeddings of their constituent words.

In the case of supervised methods, neural networks are often used to learn optimal representations. Different types of neural architectures are applied: recurrent neural networks [7], convolutional neural networks [8, 9], and attention mechanisms [10]. While the supervised methods for sentence representation can learn the importance of the words in a sentence by considering their embedded features, this is usually not the case with the unsupervised methods. For supervised methods, sentence embeddings obtained by simply averaging the embeddings of words in a sentence were shown to be useful [6]. Other unsupervised methods include: IDs [11], skip-thought vectors [12], sequential (denoising) autoencoders [13], and optimizing word embeddings for averaging [14].

Arora et al. proposed sentence embeddings obtained as the weighted average of word embeddings [15]. As a weighting scheme, the frequency-based methods, such as the inverse document frequency (idf) or the inverse sentence frequency (isf), are often used [5]. However, the importance of words depends not only on their frequencies in the entire document collection, but also on their local contexts. Our hypothesis in this paper is that the importance of a particular word could change depending on the topic of the document in which the word appears.

3 Dataset Construction

Hashimoto et al. constructed a dataset of review matrices consisting of scientific papers submitted to NLP shared tasks [1]. The dataset was used for the evaluation of query-focused multi-document summarization where the participant system comparison tables provided by the task organizers were considered as reference summaries. However, the reported summarization performance for the dataset was relatively poor with the ROUGE score ranging from 3 to 9. A major

problem is the ambiguity of the queries. The queries consist of only one or a few words, which makes query expansion extremely difficult without any context.

Based on the above observation, we constructed a new dataset by (1) collecting additional papers from later publications that report the results for the same dataset as the submitted papers, and (2) manually selecting the most relevant sentences for the queries[1]. Thus, existing system descriptions in the review matrices can be used as the "context" for query expansion. In our dataset construction, we selected four shared tasks, 2-3 papers and 1-2 queries, which also appear in the review matrices dataset constructed by Hashimoto et al., for each task. Then, we read the abstracts of the selected papers, found the most relevant chapter for a given query, and chose the most appropriate sentence that has adequate information about the query. The statistics of our new dataset are listed in Table 1.

Table 1. The statistics of our new dataset for query-focused summarization

Shared Task	# of papers	Query	Ave. # of words in the summaries	Ave. # of sentences
CoNLL 2011	3	"Learning Framework", "Markable Identification"	34.7	178.3
CoNLL 2012	3	"Learning Framework", "Markable Identification"	24.2	158.3
CoNLL 2013	3	"Description of Approach"	32.7	129.3
CoNLL 2014	2	"Description of Approach"	27.5	164.5

4 Approach

In our approach, we first obtain word importance scores by conducting a random walk on a word-graph for a given document. Secondly, we train a Siamese neural network to obtain optimal sentence embeddings. Finally, we perform a sentence-level random walk with the acquired sentence representation and an expanded query to select only one sentence for each pair of a paper and a query.

4.1 Word-Level Random Walk

The word importance scores are calculated using a random walk on a graph of all words in a document [17, 19]. In this step, we regard words as nodes, and generate a complete graph with each edge assigned a weight based on the similarity between the connected word pair. We recursively compute the probability distribution of N unique words in a document, $\mathbf{p_t^w} = (p_t^w(w_1), ..., p_t^w(w_N))$, until convergence as below:

[1] Our dataset is publicly available. https://github.com/Alab-NII/Q-SciSumm

$$\mathbf{p_t^w} = (d\mathbf{U} + (1-d)\mathbf{B})^{\mathbf{T}}\mathbf{p_{t-1}^w}. \tag{1}$$

Here, $\mathbf{U}$ corresponds to a square matrix with all elements equal $1/N$, t represents the iteration number, and d denotes a "dumping factor." $\mathbf{B}$ is obtained as

$$\mathbf{B}(i,j) = \frac{sim(w_i, w_j)}{\sum_{k=1}^{N} sim(w_i, w_k)} \tag{2}$$

where $sim(w_i, w_j)$ is the similarity score between words w_i and w_j, calculated as the cosine similarity of corresponding word embeddings pretrained using word2vec [16]. The initial probability distribution is set to $\mathbf{p_0^w} = (1/N, 1/N, ...)$. After conversion, we obtain the probability distribution $\mathbf{p^w} = (p^w(w_1), ..., p^w(w_N))$, which we use later as word importance scores. This operation is performed for each scientific paper. Accordingly, different sets of word importance scores are obtained for each paper.

4.2 Siamese Network

Our Siamese network is based on the study by Kenter et al. (2016) [14] that predicts the probability of a sentence pair occurring next to each other in the training data. In their formulation, each word w in a sentence s is first converted into a d-dimensional embedding vector v_w. Then, a localized sentence representation v_s is obtained as the weighted average of the word embeddings using the word importance scores obtained in Section 4.1:

$$v_s = \frac{\sum_{w \in s} p^w(w) v_w}{\sum_{w' \in s} p^w(w')}. \tag{3}$$

The loss function of the network is the categorical cross-entropy:

$$L = -\sum_{s_i \in D} \sum_{s_j \in \{S^+ \cup S^-\}} p(s_i, s_j) \cdot log(p_\theta(s_i, s_j)) \tag{4}$$

where D denotes a document, and S^+ and S^- are positive and negative samples, respectively. The negative samples are randomly chosen from D. In Equation (4), $p(\cdot)$ and $p_\theta(\cdot)$ are the target and the predicted probability distributions, respectively. The predicted probability distribution is obtained using a softmax function over $cos(v_{s_i}, v_{s_j})$. The target probability distribution is given as $p(s_i, s_j) = \frac{1}{|S^+|}$ for $s_j \in S^+$ and 0 for $s_j \in S^-$.

4.3 Sentence-Level Random Walk

In this step, we conduct a sentence-level random walk to select the most appropriate sentence as a summary for a given query [18]. We first expand the query with words that appear more than once in the reference summaries in the review matrices dataset constructed by Hashimoto et al. [1]. Each word in an expanded query is represented as a d-dimensional embedding pretrained using word2vec.

The relevance score between a query q and a sentence s is calculated as:

$$rel(s,q) = \frac{1}{n} sumLargest_n(cos(v_w, v_u)|w \in s, u \in q) \tag{5}$$

where $sumLargest_n$ computes the sum of the top n candidates of $cos(v_w, v_u)$. The embeddings v_w and v_u are obtained using the pretrained word2vec. The probability distribution of sentences $\mathbf{p_t^s}$ is calculated as:

$$\mathbf{p_t^s} = (d\mathbf{Q} + (1-d)\mathbf{B})^T \mathbf{p_{t-1}^s} \tag{6}$$

where $\mathbf{p_0^s}$ is $(1/|D|, 1/|D|, ...)$ and Q is a square matrix whose element is:

$$\mathbf{Q}(i,j) = \frac{rel(s_i,q)}{\sum_{s_j \in D} rel(s_j,q)}. \tag{7}$$

Finally, we conduct a sentence-level random walk until $\mathbf{p_t^s}$ converges. The difference between the word- and sentence-level random walks is that the latter considers query relevance. This type of query-focused random walk can be used for query-focused and extractive summarization [18].

5 Experiments

5.1 Experimental Setup

As noted earlier, we used the word2vec toolkit to obtain word embeddings. We trained word2vec on the scientific papers collected from ACL Anthology[2] using SideNoter [20]. The minimum count of words used for training was four and the size of the embedding vectors was 512. The number of epochs for training was five, and the window size was five.

For the Siamese network, we trained our network with two positive and five negative samples for each sentence. The number of epochs for training was one. In our query expansion, we added at most eight words from the reference summaries. For both the word- and sentence-level random walks, the dumping factor was set to 0.3. All these parameters other than those related to word2vec were tuned using the review matrices dataset [1]. The two dumping factors were chosen from $\{0.1, 0.2, ..., 1.0\}$ and the number of added queries was chosen from $\{1, 2, ..., 10\}$.

We compared our sentence representation with four other baselines.

- **TFIDF**: a v-dimensional vector whose elements are the dot products of the term frequency and the inverse document frequency of each word.
- **Average WE**: obtained by simply averaging pretrained word embeddings in a sentence.
- **Siamese CBOW**: proposed by [14]. In this approach, a simple average of word embeddings is used as a sentence representation used to train a Siamese network.

[2] http://aclweb.org/anthology/

– **IDF Weighted Siamese Network**: obtained by replacing the word importance scores in our approach with the IDF scores. Note that the IDF scores are common for all documents.

5.2 Results

In the evaluation, we calculated the ROUGE-1, ROUGE-2, and ROUGE-SU4 scores which are the automatic evaluation metrics [21] for each summary output. The average performance for each method is shown in Table 2.

As shown in Table 2, surface-level word matching, used by TFIDF, is not sufficient for our sentence similarity calculation. Among the other four embedding-based methods, our approach achieved the best score for all the metrics. Although our dataset is still small and limited, the results indicate that considering context-dependent word importance scores proved effective in our task. In addition, the performance of simply averaging pretrained word embeddings was relatively low; thus, utilizing a Siamese network had a signifiant impact on this task.

Table 2. Results on our new review matrices dataset.

Sentence Representation	ROUGE-1	ROUGE-2	ROUGE-SU4
TFIDF	17.07	5.48	5.03
Average WE	26.73	8.52	9.75
Siamese CBOW	33.32	18.29	18.65
IDF weighted	28.67	12.33	12.53
Our approach	**35.01**	**18.59**	**19.12**

6 Conclusion

In this study, we constructed a dataset for the summarization of scientific papers, and evaluated the proposed approach on our dataset. In terms of mining scientific papers by query-focused multi-document summarization, our localized sentence representations were proved to be useful. However, in certain cases, the output summaries for different queries were the same, indicating that the dumping factor of sentence-level random walk should be more flexibly adjusted. In addition, we observed the cases where the query expansion fails to capture the original intension. Future work includes the improvement of the query expansion and the relevance score calculation between a sentence and a query. We will also consider to incorporate more advanced graph-based algorithms such as [4, 5].

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 16K12546 and CREST JP-MJCR1513.

References

1. Hashimoto, H., Shinoda, K., Yokono, H., Aizawa, A.: Automatic Generation of Review Matrices as Multi-document Summarization of Scientific Papers. In: Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 69-82. (2017)
2. Cheng, J., Lapata, M.: Neural Summarization by Extracting Sentences and Words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 484-494. (2016)
3. Ren, P., Chen, Z., Ren, Z., Wei, F., Ma, J., Rijke, M.: Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 95-104. (2017)
4. Parveen, D., Ramsl, H., Strube, M.: Topical Coherence for Graph-based Extractive Summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1949-1954. (2015)
5. Wang, K., Liu, T., Sui, Z., Chang, B.: Affinity-Preserving Random Walk for Multi-Document Summarization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 210-220, (2017)
6. Wieting, J., Bansal, M., Gimpel, K., Livescu, K.: Towards Universal Paraphrastic Sentence Embeddings. In: Proceedings of the 4th International Conference on Learning Representations. (2016)
7. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 1556-1566. (2015)
8. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A Convolutional Neural Network for Modelling Sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 655-665. (2014)
9. Gan, Z., Pu, Y., Henao, R., Li, C., He, X., Carin, L.: Learning Generic Sentence Representations Using Convolutional Neural Networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2390-2400. (2017)
10. Lin, Z., Feng, M., Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A Structured Self-attentive Sentence Embedding. In: Proceedings of the 5th International Conference on Learning Representations. (2017)
11. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on Machine Learning. (2014)
12. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S.: Skip-thought vectors. In: Advances in Neural Information Processing Systems. (2015)
13. Hill, F., Cho, K., Korhonen, A.: Learning Distributed Representations of Sentences from Unlabelled Data. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1367-1377. (2016)
14. Kenter, T., Borisov, A., Rijke, M.: Siamese CBOW: Optimizing word embeddings for sentence representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 941-951. (2016)
15. Arora, S., Liang, Y., Ma, T.: A Simple But Tough-to-beat Baseline For Sentence Embeddings. In: Proceedings of the 5th International Conference on Learning Representations. (2017)

16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. (2013)
17. Erkan, G., Radev, D.R.: LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22:457-479. (2004)
18. Otterbacher, J., Erkan, G., Radev, D.R.: Using Random Walks for Question-focused Sentence Retrieval. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 915-922, 2005.
19. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. (2004)
20. Abekawa, T., Aizawa, A.: SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation. In: Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations, pp. 136-140. (2016)
21. Lin, C.: ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics workshop on Text Summarization Branches Out, pp. 74-81. (2004)