ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096137

IMPROVING SENTENCE SIMILARITY ESTIMATION FOR UNSUPERVISED EXTRACTIVE SUMMARIZATION

Shichao Sun[1] *Ruifeng Yuan*[1] *Wenjie Li*[1] *Sujian Li*[2]

[1] The Hong Kong Polytechnic University [2] Peking University
{csssun, csryuan, cswjli}@comp.polyu.edu.hk lisujian@pku.edu.cn

ABSTRACT

Unsupervised extractive summarization aims to extract salient sentences from a document as the summary without labeled data. Recent literatures mostly research how to leverage sentence similarity to rank sentences in the order of salience. However, sentence similarity estimation using pre-trained language models mostly takes little account of document-level information and has a weak correlation with sentence salience ranking. In this paper, we proposed two novel strategies to improve sentence similarity estimation for unsupervised extractive summarization. We use contrastive learning to optimize a document-level objective that sentences from the same document are more similar than those from different documents. Moreover, we use mutual learning to enhance the relationship between sentence similarity estimation and sentence salience ranking, where an extra signal amplifier is used to refine the pivotal information. Experimental results demonstrate the effectiveness of our strategies. [1]

Index Terms— Unsupervised Extractive Summarization, Sentence Similarity, Contrastive Learning, Mutual Learning

1. INTRODUCTION

Text summarization aims to condense a long document into its shorter version while preserving salient content. Existing methods can be divided into two paradigms, i.e., abstractive and extractive. Abstractive methods generate a summary word by word. Extractive methods select salient sentences from a document as the summary. Modern neural network based approaches [1, 2, 3] have achieved promising results, which heavily rely on the large-scale annotated corpus. However, it is unrealistic to expect large-scale and high-quality annotated corpus to be available all the time. It therefore comes as no surprise that unsupervised summarization has attracted much attention [4, 5, 6, 7, 8, 9]. Most attempts are extractive since it is obviously difficult to generate the summary sentences without any reference summary.

Unsupervised extractive summarization is commonly graph-based [5, 8, 9]. These methods contain two stages. The first stage is to obtain a sentence encoder, which encodes a sentence into an embedding. Pre-trained language models like BERT [10] are used in this stage. The second stage is to calculate the salience score via the sentence embedding, and sentences with the highest scores are selected as a summary. In this stage, a document is represented as a graph, where nodes represent sentences. The weight of an edge is the similarity of two adjacent nodes (sentences), which can be estimated by dot product or cosine distance using the sentence embedding. Then the node centrality is calculated as the salience score of a sentence. Most existing methods put more effort into this step while they commonly used pre-trained language sentence encoders. PacSum [5] added the direction information to the degree centrality. FAR [8] incorporated the facet information into PacSum, and DASG [9] augmented PacSum with the distance information.

However, there is a gap between training a sentence encoder and estimating the similarity of two sentences using dot product or cosine distance. The training objectives of a sentence encoder are commonly masked language model [9] and neighboring sentence prediction [5, 8, 9]. There is not any explicit correlation between sentence similarity estimation and dot product or cosine distance of two sentence embeddings. These pre-trained models are not expected to well estimate sentence similarity using dot product or cosine distance. It is also intuitive that by only considering a sentence and its neighboring sentences, the sentence embedding can hardly capture the document-level similarity that sentences from the same document are more similar than those from different documents. Besides, TF-IDF can even outperform BERT in unsupervised extractive summarization as shown in Table 1. It indicates that there is a weak relationship between sentence similarity estimation and sentence salience ranking. This motivated us to explore how to improve sentence similarity estimation for unsupervised extractive summarization.

To address the above issues, we propose two novel strategies to train the sentence encoder. To enable the pre-trained models to get aware of document-level information, we use contrastive learning to optimize a document-level objective that sentences from the same document are more similar than those from different documents. To build the relationship between similarity estimation and dot product of two sentence

[1] Our code is available at: `https://github.com/ShichaoSun/SS4Sum`

embeddings, we define the above sentence similarity as the dot product of two sentence embeddings. Moreover, we use mutual learning [11] to enhance the relationship between sentence similarity estimation and sentence salience ranking. An extra amplifier called Deep Differential Amplifier [12] is used to refine the pivotal information. It learns from the coarse-grained pivotal information that the top 40% ranked sentences are marked as salient sentences, and the bottom 40% ranked sentences are marked as unimportant sentences, where the salience scores are estimated using our sentence encoder. It will output the fine-grained pivotal information that top 3 sentences are marked as salient sentences, and other sentences are marked as unimportant sentences. This fine-grained pivotal information is used to adjust our sentence salience ranking. It means that under the mutual learning framework, the Deep Differential Amplifier learns the salience scores calculated by our sentence similarity estimation. Meanwhile our calculated salience scores are supervised by the predicted results of the Deep Differential Amplifier.

We conduct experiments on two datasets, i.e., **NYT** [13] and **CNNDM** [14]. Experimental results show that our sentence similarity estimation beats other similarity estimation methods in unsupervised extractive summarization. The ablation study demonstrates the effectiveness of our strategies.

2. METHOD

2.1. Similarity Estimation

Let D denote the document consisting of a sequence of sentences $\{s_1, s_2, \cdots, s_n\}$. And e_{ij} denotes the similarity for each sentence pair (s_i, s_j). Their similarity is dot product of their sentence embeddings, which is calculated as follows:

$$e_{ij} = v_i^\top v_j \tag{1}$$

where v_i is the embedding of sentence s_i and v_j is the embedding of sentence s_j. Note that the sentence embedding is generated by a sentence encoder like BERT.

2.2. Contrastive Learning

Contrastive learning is used to optimize the novel objective that the similarity of sentences from the same document are higher than the similarity of sentences from different documents, because sentences of a document contribute to one core. Specifically, we utilize the contrastive learning to increase the dot product of sentences from the same document. Meanwhile, we decrease the dot product of sentences from different sentences (within the same batch). This is because in unsupervised extractive summarization dot product is used to estimate sentence similarity. For the sentences $\{s_1, s_2, \cdots, s_n\}$ in the same batch, where we make s_{2i} and s_{2i+1} belong to the same document, and s_{2i} and s_j ($j \neq 2i$ and $j \neq 2i + 1$) belong to different documents. The contrastive learning loss L_{con} can be calculated as follows:

$$L_{con} = -\log \frac{\exp(v_{2i}^\top v_{2i+1}/\tau)}{\sum\limits_{j \neq 2i} \exp(v_{2i}^\top v_j/\tau)} \tag{2}$$

where τ is a scalar temperature parameter. Note that this loss function can directly optimize sentence similarity.

2.3. Salience Score

The salience score is used to rank sentences of a document, which is calculated by a graph-based algorithm. A document is represented as a graph in which nodes correspond to sentences, and each edge between two sentences is weighted by their similarity. Degree Centrality (DC) is a popular method for calculating the salience score. It is based on the hypothesis that a salient sentence should be similar to the rest sentences in the document. The DC of sentence s_i from a document containing n sentences is calculated as follows:

$$\mathrm{DC}(s_i) = \sum_{j \in \{0, \cdots, i-1, i+1, \cdots, n\}} e_{ij} \tag{3}$$

2.4. Mutual Learning

Mutual learning is used to enhance the relationship between sentence similarity estimation and sentence salience ranking by an extra signal amplifier as illustrated in Figure 1. The left part of this figure is our sentence salience ranking (degree centrality) as described in Section 2.3 with an extra Softmax layer. The output is the ranking score r_i of each sentence s_i. The right part of this figure is a signal amplifier called Deep Differential Amplifier, which is able to amplify the salient signal between the current sentence and other sentences. A feedforward neural network with a Sigmoid layer is used to convert the salient signal to ranking score r_i of each sentence s_i. Note that this signal amplifier is used to refine the pivotal information. This signal amplifier will be supervised by the coarse-grained pivotal information that the top 40% ranked sentences are marked as salient sentences, and the bottom 40% ranked sentences are marked as unimportant sentences, where the salience scores are estimated using degree centrality as shown in the left part of this figure. And this signal amplifier will output the fine-grained score for each sentence that the top 3 sentences are marked as salient sentences, and other sentences are marked as unimportant sentences. This is motivated by the strong ability of generalization. Conversely, this fine-grained score will be used to adjust the salience ranking score of the model in the left part of this figure. We jointly train degree centrality with a softmax layer and the signal amplifier with a sigmoid layer together to construct a mutual learning variant. During this process, the sentence similarity estimation is indirectly optimized to adapt to sentence salience ranking.

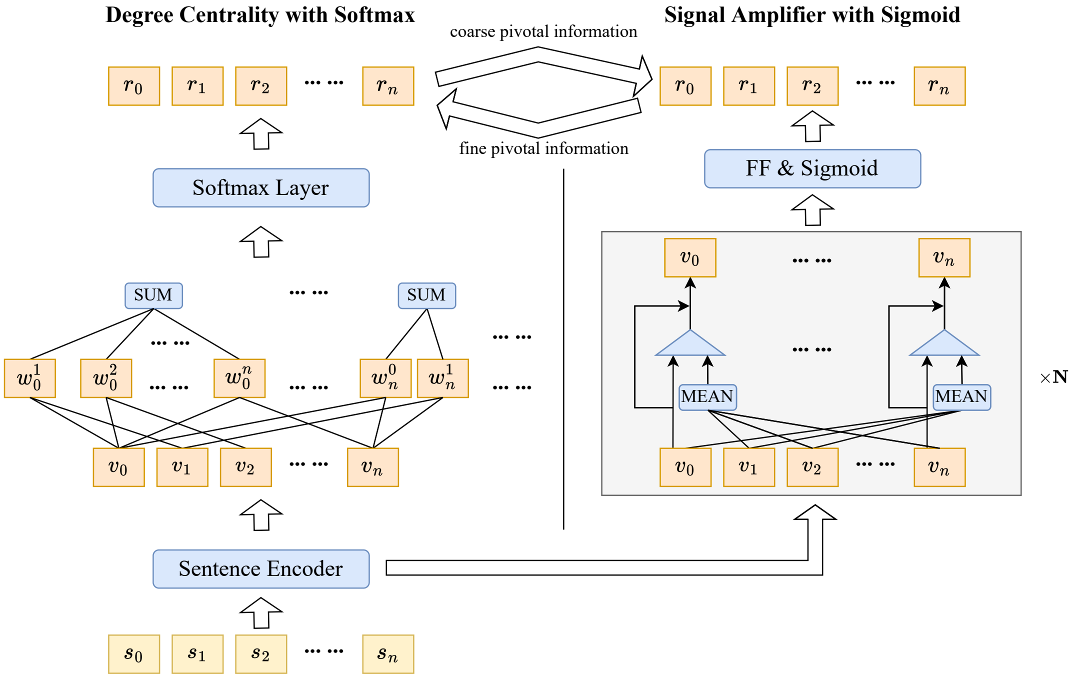
Degree Centrality with Softmax
Signal Amplifier with Sigmoid
coarse pivotal information
fine pivotal information
r_0 r_1 r_2 ⋯ r_n
Softmax Layer
FF & Sigmoid
SUM
SUM
w_0^1 w_0^2 w_0^n w_n^0 w_n^1
v_0 v_1 v_2 ⋯ v_n
MEAN
MEAN
×N
Sentence Encoder
s_0 s_1 s_2 ⋯ s_n

Fig. 1. Mutual Learning for Sentence Embedding Learning.

The left part of Figure 1 is our sentence salience ranking (degree centrality) as described in Section 2.3 with an extra Softmax layer, which is used to get normalized salience score. It can be calculated as follows:

$$r_i = \frac{\exp(\frac{\mathrm{DC}(s_i)}{\tau(n-1)})}{\sum\limits_{j\in D} \exp(\frac{\mathrm{DC}(s_j)}{\tau(n-1)})} \tag{4}$$

where n is the number of sentences in a document D and τ is a scalar temperature parameter. This salient score will be optimized by the loss function as follows:

$$L_{dc} = -\log \sum_{i\in C} \frac{r_i}{\sum\limits_{j\in D} r_j} \tag{5}$$

where C consists of the salient sentences that the signal amplifier (right part) selects as a summary, i.e., the top 3 salient sentences. It means to train the sentence encoder by using the fine-grained pivotal information from the amplifier.

The right part is the state-of-the-art supervised extractive summarization method, Deep Differential Amplifier [12]. The gray area describes the process of amplifying salient signal, which is iterated for N times as follows:

$$\begin{aligned} F(v_i) &= \mathrm{MLP}(v_i - \mathrm{mean}(\{v_j \mid j \neq i\})) \\ v_i &= F(v_i) + v_i \end{aligned} \tag{6}$$

where v_i is the sentence embedding generated by the sentence encoder with the input of a sentence s_i, MLP is a multilayer perceptron with two layers and the ReLU activation function. And the salience score r_i is calculated as follows:

$$r_i = \mathrm{sigmoid}(\mathbf{w}^\top v_i) \tag{7}$$

where $\mathbf{w}$ is the trainable parameters. It is optimized by using the binary cross entropy as follows:

$$L_{amp} = -y_i \log(r_i) - (1-y_i)\log(1-r_i) \tag{8}$$

where y_i is 1 if the sentence s_i is one of the top-k ranked sentences, and y_i is 0 if sentence s_i is one of the bottom-k ranked sentences according to the salience scores. The salience scores are calculated according to Equation (4). k is around 40% sentence number of a document. This means to train the amplifier by using the coarse-grained pivotal information.

2.5. Training Objective

Finally, the sentence encoder will be trained by summing the above three loss L as follows:

$$L = L_{con} + L_{dc} + L_{amp} \tag{9}$$

3. EXPERIMENT AND ANALYSIS

3.1. Datasets

We evaluate our sentence similarity estimation by testing whether it can improve the performance for unsupervised

Table 1. The Rouge scores on CNNDM and NYT.

	CNNDM			**NYT**		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
TF-IDF	33.00	11.70	29.50	33.20	13.10	29.00
BERT	28.33	8.29	25.38	25.87	7.29	22.28
SimCSE	31.98	10.42	28.52	30.55	10.26	26.38
PacSum BERT	31.42	10.07	28.00	30.99	10.28	26.81
SimBERT (ours)	**35.41**	**13.18**	**31.75**	**35.20**	**14.75**	**30.97**

extractive summarization on two summarization datasets, i.e., **NYT** [13] and **CNNDM** [14]. Our sentence encoder is trained by using the input documents of CNNDM training dataset and NYT training dataset. The final training dataset contains 322,828 documents with more than five sentences. For the sake of quickly validating, we randomly select 500 samples from CNNDM validation dataset and 500 samples from NYT validation dataset.

3.2. Baselines

We choose four different sentence representation methods as baselines. The first one is TF-IDF, whose value of the corresponding dimension is the tf (term frequency) times the idf (inverse document frequency) of the word. The second one uses the original BERT to encode the sentence. The third one is SimCSE [15], which takes an input sentence and predicts itself in a contrastive learning objective, with only dropout used as noise. The fourth one (PacSum BERT) comes from the [5], which captures semantic information by distinguishing context sentences from other sentences. These sentence representations are used to calculate the degree centrality (Equation (3)), and the top 3 sentences are selected as a summary.

3.3. Automatic Evaluation

We automatically evaluate summary quality using Rouge [16], and the experimental results on CNNDM and NYT are presented in Table 1. It shows that our sentence embedding (SimBERT) achieved the best performance on CNNDM and NYT, so it can be proved that our sentence similarity estimation can be more suitable for unsupervised extractive summarization. It can be attributed to the novel training objectives of explicitly capturing document-level sentence similarity and enhancing the relationship between sentence similarity and sentence salience ranking.

Besides, it should be noted that TF-IDF can perform better than the other BERT variants except SimBERT. It is counterintuitive because the pre-training based representation (like BERT) always outperforms statistic based representation (like TF-IDF). It indicates that current pre-training based sentence encoder is not good at capturing sentence similarity. This result motivated us to explore how to estimate sentence similarity for unsupervised extractive summarization.

Table 2. The results of ablation study.

	CNNDM and NYT		
	R1	R2	RL
SimBERT	35.36	13.62	31.54
- mutual learning	31.68	10.50	28.04
- contrastive learning	27.54	7.96	24.42

3.4. Ablation Study

The ablation study is conducted on the merged testing dataset of CNNDM and NYT to evaluate the contribution of mutual learning (L_{dc} and L_{amp}) and contrastive learning (L_{con}). The Rouge scores are given in Table 2. It tells that mutual learning and contrastive learning are complementary since we can achieve better results by using both of them. Besides, it can be found that the performance will degrade a lot without contrastive learning. It can indicate that the sentence similarity in the document level is important. The degradation without mutual learning can show that there is a weak relationship between sentence similarity and sentence salience ranking.

4. CONCLUSION

In this paper, we proposed two novel strategies to improve sentence similarity estimation for unsupervised extractive summarization. We use contrastive learning to optimize a document-level objective that sentences from the same document are more similar than those from different documents. Moreover, we use mutual learning to enhance the relationship between sentence similarity estimation and sentence salience ranking, where an extra signal amplifier is used to refine the pivotal information. Experimental results show that our sentence similarity estimation beats other similarity estimation methods in unsupervised extractive summarization. The ablation study demonstrates the effectiveness of our strategies.

5. ACKNOWLEDGEMENT

This paper was supported by Research Grants Council of Hong Kong (PolyU/15203617 and PolyU/5210919), National Natural Science Foundation of China (61672445, 62076212).

6. REFERENCES

[1] Abigail See, Peter J. Liu, and Christopher D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017, pp. 1073–1083, Association for Computational Linguistics.

[2] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang, "Extractive summarization as text matching," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 6197–6208, Association for Computational Linguistics.

[3] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig, "BRIO: Bringing order to abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, May 2022, pp. 2890–2903, Association for Computational Linguistics.

[4] Rada Mihalcea and Paul Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, July 2004, pp. 404–411, Association for Computational Linguistics.

[5] Hao Zheng and Mirella Lapata, "Sentence centrality revisited for unsupervised summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 6236–6247, Association for Computational Linguistics.

[6] Shusheng Xu, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou, "Unsupervised extractive summarization by pre-training hierarchical transformers," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 1784–1795, Association for Computational Linguistics.

[7] Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve, "TED: A pretrained unsupervised summarization model with theme modeling and denoising," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 1865–1874, Association for Computational Linguistics.

[8] Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li, "Improving unsupervised extractive summarization with facet-aware modeling," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 1685–1697, Association for Computational Linguistics.

[9] Jingzhou Liu, Dominic J. D. Hughes, and Yiming Yang, *Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs*, p. 2313–2317, Association for Computing Machinery, New York, NY, USA, 2021.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[11] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu, "Deep mutual learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[12] Ruipeng Jia, Yanan Cao, Fang Fang, Yuchen Zhou, Zheng Fang, Yanbing Liu, and Shi Wang, "Deep differential amplifier for extractive summarization," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Aug. 2021, pp. 366–376, Association for Computational Linguistics.

[13] Evan Sandhaus, "The new york times annotated corpus," *Linguistic Data Consortium, Philadelphia*, vol. 6, no. 12, pp. e26752, 2008.

[14] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom, "Teaching machines to read and comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2015, NIPS'15, p. 1693–1701, MIT Press.

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 6894–6910, Association for Computational Linguistics.

[16] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.