7-2020

Bridging hierarchical and sequential context modeling for question-driven extractive answer summarization

Yang DENG
Singapore Management University, ydeng@smu.edu.sg

Wenxuan ZHANG

Yaliang LI

Min YANG

Wai LAM

See next page for additional authors

Citation

DENG, Yang; ZHANG, Wenxuan; LI, Yaliang; YANG, Min; LAM, Wai; and SHEN, Ying. Bridging hierarchical and sequential context modeling for question-driven extractive answer summarization. (2020). *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, Online, 2020 July 25-30*. 1693-1696.
Available at: https://ink.library.smu.edu.sg/sis_research/9100

Bridging Hierarchical and Sequential Context Modeling for Question-driven Extractive Answer Summarization

Yang Deng[1], Wenxuan Zhang[1], Yaliang Li[2], Min Yang[3], Wai Lam[1], Ying Shen[4,*]
[1]The Chinese University of Hong Kong,[2]Alibaba Group,
[3]Chinese Academy of Sciences, [4]School of Intelligent Systems Engineering, Sun Yat-Sen University
{ydeng,wxzhang,wlam}@se.cuhk.edu.hk,yaliang.li@alibaba-inc.com,
min.yang@siat.ac.cn,sheny76@mail.sysu.edu.cn

ABSTRACT

Non-factoid question answering (QA) is one of the most extensive yet challenging application and research areas of retrieval-based question answering. In particular, answers to non-factoid questions can often be too lengthy and redundant to comprehend, which leads to the great demand on answer sumamrization in non-factoid QA. However, the multi-level interactions between QA pairs and the interrelation among different answer sentences are usually modeled separately on current answer summarization studies. In this paper, we propose a unified model to bridge hierarchical and sequential context modeling for question-driven extractive answer summarization. Specifically, we design a hierarchical compare-aggregate method to integrate the interaction between QA pairs in both word-level and sentence-level into the final question and answer representations. After that, we conduct the question-aware sequential extractor to produce a summary for the lengthy answer. Experimental results show that answer summarization benefits from both hierarchical and sequential context modeling and our method achieves superior performance on WikiHowQA and PubMedQA.

ACM Reference Format:
Yang Deng[1], Wenxuan Zhang[1], Yaliang Li[2], Min Yang[3], Wai Lam[1], Ying Shen[4,*]. 2020. Bridging Hierarchical and Sequential Context Modeling for Question-driven Extractive Answer Summarization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401208

1 INTRODUCTION

Recent years have witnessed many successful applications on non-factoid question answering (QA), such as community QA [3] or explainable QA [4, 7]. However, the original answers, which are usually provided by ordinary users or from long documents, often contain plenty of irrelevant and redundant information. In real-world applications, such lengthy answers may result in the reading difficulties and misunderstandings for other users [12, 15].

Text summarization can be an effective approach to tackle the issue of answer redundancy. In the past studies, answer summarization was mainly explored by traditional information retrieval methods [5, 8, 12, 15] and query-based summarization models [1, 6, 9, 11] in specific. According to the type of summary, these methods can be generally categorized into extractive [1, 5, 8, 11, 12, 15] and abstractive summarization [6, 9]. In this work, we focus on extractive answer summarization, since they are computationally efficient, and can generate more grammatical and coherent summaries [2, 16].

As for non-factoid QA, the answer summary is supposed to be highly related with the concerned question, while the imbalance of information in the question and answer causes difficulties in differentiating the semantic relevancy among answer sentences with the question. However, existing extractive answer summarization studies often underutilize the interaction information between the question and answer during the extraction process [1, 11] or rely heavily on the feature engineering for relevance measurement [5, 12, 15]. Thus, it requires a special design to carefully model the interaction between the question and the original answer to tackle the issue of information imbalance.

Answer sentence selection, which aims at selecting sentences from a set of candidates to answer the question, can be an alternative method for answer summarization. The Compare-Aggregate architecture [10, 13, 14] has been widely adopted to model the interaction between QA pairs, by aggregating comparison signals from low-level elements into high-level representations. Inspired by such idea, we propose to hierarchically model the relevant information between QA pairs in both word and sentence-level to obtain suitable sentence representations for the concerned answer summarization task. On the other hand, current query-based summarization [1, 11] and answer sentence selection methods [10, 13, 14] both fall short to capture the correlation among different sentences in the original answer, which is supposed to be of great importance in extractive summarization settings. Sequential modeling successfully overcomes this issue by taking into account both the current sentence saliency and the information from previous sentences [2, 16].

In this work, we bridge the Hierarchical and Sequential Context Modeling (HSCM) for question-driven extractive answer summarization. Specifically, we propose a hierarchical compare-aggregate method to integrate the hierarchical interaction information between question-answer pairs in both word-level and sentence-level into a sequential extractive summarization model. Experimental

* Corresponding Author

This work was financially supported by the Shenzhen General Research Project (No. JCYJ20190808182805919) and a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14204418).

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401208

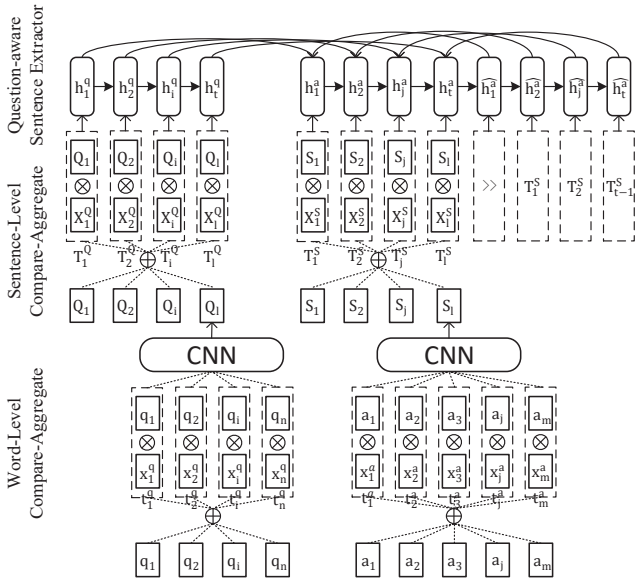

Figure 1: Hierarchical and Sequential Context Modeling for Question-driven Answer Summarization

results show that the proposed method achieves superior performance on two non-fatoid QA datasets.

2 METHOD

Given a question Q and its answer A composed of a list of sentences $[s_1, s_2, ..., s_l]$, the model aims to extract sentences from A to construct a summary Y for the answer. The overall architecture of HSCM model is depicted in Figure 1, which consists of three main components, including Word-level Compare-Aggregate, Sentence-level Compare-Aggregate, and Question-aware Sequential Extractor.

2.1 Word-level Compare-Aggregate

Pre-trianed word embeddings of the question $Q = \{q_1, q_2, ..., q_n\}$ and each sentence s_i in the answer $A = \{s_1, s_2, ..., s_l\}$ are input into the model. We first conduct an attention operation to align the word-level information between the question and the answer sentence, and obtain the attention-weighted vectors for each word for both the question and the answer sentence. Specifically, for the l-th sentence $s_l = \{a_1, a_2, ..., a_m\}$ in the answer A, we have:

$$e_{ij} = q_i \cdot a_j, \tag{1}$$

$$\alpha_{ij}^q = \frac{\exp(e_{ij})}{\sum_{k=1}^m \exp(e_{ik})}, \quad \alpha_{ij}^a = \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{kj})}, \tag{2}$$

$$x_i^q = \sum_{j=1}^m \alpha_{ij}^q a_j, \quad x_j^a = \sum_{i=1}^n \alpha_{ij}^a q_i, \tag{3}$$

where e_{ij} composes the attention matrix. α_{ij}^q and α_{ij}^a are the attention weights. x_i^q and x_j^a are the attention-weighted vectors.

We match each a_j with the corresponding x_j^a for word-level comparison. Here we adopt element-wise multiplication as the comparison function to compute the word-level comparison result:

$$t_j^a = a_j x_j^a, \quad t_i^q = q_i x_i^q. \tag{4}$$

After obtaining the word-level comparison results t_j^a and t_i^q, we finally aggregate these vectors with a convolutional layer:

$$Q = \text{CNN}([t_1^q, ..., t_n^q]), \quad S = \text{CNN}([t_1^a, ..., t_m^a]), \tag{5}$$

where Q and S denote the representations for the question and each answer sentence, respectively. Note that there is a unique question representation Q_l corresponding to each sentence representation S_l in the answer, since the word-level comparison results for each answer sentence are integrated into the question representations.

2.2 Sentence-level Compare-Aggregate

In the sentence-level compare-aggregate, the question is regarded as a whole, while the answer is tokenized into sentences. In order to perform sentence-level compare-aggregate, we compare each sentence in the answer to the given question. Therefore, the question representation Q and the sentence representations S of all the sentences in the answer are input into the sentence-level compare-aggregate layer. Similar to word-level compare-aggregate, the question and the answer sentences are aligned by attention, and sentence-level comparison results are then computed by element-wise multiplication:

$$E_{ij} = Q_i \cdot S_j, \tag{6}$$

$$\omega_{ij}^Q = \frac{\exp(E_{ij})}{\sum_{k=1}^l \exp(E_{ik})}, \quad \omega_{ij}^S = \frac{\exp(E_{ij})}{\sum_{k=1}^l \exp(E_{kj})}, \tag{7}$$

$$X_i^Q = \sum_{j=1}^m \omega_{ij}^Q Q_j, \quad X_j^S = \sum_{i=1}^n \omega_{ij}^S S_i, \tag{8}$$

$$T_j^S = Q_j X_j^S, \quad T_i^Q = S_i X_i^Q, \tag{9}$$

where E_{ij} is the sentence-level attention matrix. ω_{ij}^Q and ω_{ij}^S are the attention weights. X_i^Q and X_j^S are the attention-weighted sentence vectors. T_i^Q and T_j^S are the sentence-level comparison results.

Then we aggregate the sentence-level comparison results into a pair of LSTM to learn sequential representations for the question and the answer:

$$H^q = \text{LSTM}([T_1^Q, ..., T_n^Q]), \quad H^a = \text{LSTM}([T_1^S, ..., T_{m_1}^S]), \tag{10}$$

where H_q and H_a are the final question and answer representations.

2.3 Question-aware Sequential Extractor

We adopt another recurrent neural network as a decoder to label each sentence sequentially, in which the next decoded label takes into account both the encoded document and the previous decoded label. Besides, as the question is of great importance in deciding whether the sentence should be extracted, the encoded question is also integrated into the label prediction:

$$\hat{h}_t^a = \text{LSTM}(s_{t-1}, \hat{h}_{t-1}^a), \tag{11}$$

$$p = \text{MLP}(h_t^a : \hat{h}_t^a : h_t^q). \tag{12}$$

The model is trained to minimize the cross-entropy loss function:

$$L = -\sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log(1 - p_i)], \tag{13}$$

where p and y denote the output of the softmax layer and the ground-truth label.

Table 1: Statictis of Dataset ($Q/A/Y$ denotes Question / Answer / Summary)

	WikiHowQA	PubMedQA
#QA Pairs (train/dev/test)	142K / 19K / 43K	169K / 21K / 21K
Avg Length ($Q/A/Y$)	6.93 / 527 / 67.1	16.3 / 238 / 41.0

3 EXPERIMENT

3.1 Experimental Setup

Dataset. We evaluate the proposed method on two non-factoid QA datasets, WikiHowQA [3] and PubMedQA [7]. WikiHowQA is a community-based QA dataset collected from WikiHow website, in which each sample consists of a question, a set of candidate long answers, and corresponding answer summaries for each answer. We conduct experiments on the positive set of QA pairs as the answer summary generation setting in the original paper. PubMedQA is a conclusion-based biomedical QA dataset collected from PubMed abstracts, in which each instance is composed of a question, a context, and an answer which is the conclusion of the context corresponded to the question. We treat the conclusion as the answer summary in our experiments. The statistics of the WikiHowQA and PubMedQA dataset are shown in Table 1.

Ground-truth Label. Following previous works on extractive summarization [2, 16], we adopt a rule-based greedy approach to label the sentences in the original answer, which is based on the idea that the extracted sentences are supposed to maximize the ROUGE score with respect to the reference summary. We add one sentence to the extracted sentence set at a time to maximize the ROUGE-1 and ROUGE-2 F1 scores with the summary, and stop adding when none of the remaining sentences improves the Rouge score. Finally, the sentences in the extracted sentence set are labeled as the ground-truths for training. In evaluation, we generate summaries by selecting the sentences with the higher scores.

Implementation Details.[1] Pre-trained GloVE embeddings[2] of 100 dimensions are adopted as word embeddings. The learning rate and the dropout rate are set to 0.001 and 0.5 respectively. The hidden unit size of the 2-layer LSTM is set to 150. We train our model in batches with size of 32. All other parameters are randomly initialized from [-0.05, 0.05]. The maximum number of sentences in the original answer and the maximum length of each sentence are both set to be 30. We use a list of convolutional filters with window size of {1,2,3,4,5,6,7}, and the corresponding number of convolutional feature maps are set to be {50,50,100,100,100,100}. We restrict the length of generated summaries within 100 words.

3.2 Experimental Results

3.2.1 ***Comparative Methods.*** In answer summarization task, we can use text summarization methods as well as answer sentence selection methods to extract sentences. Therefore, we compare the proposed method with the state-of-the-art baseline methods on traditional extractive summarization, answer sentence selection, and query-based summarization methods.

[1] https://github.com/dengyang17/hscm
[2] http://nlp.stanford.edu/data/glove.6B.zip

Table 2: Method Comparisons for Answer Summarization. * Abstractive methods. † The result reported from [3].

Models	WikiHowQA			PubMedQA		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**
Lead3	24.66†	5.56†	22.67†	30.86	9.77	21.15
NeuralSum [2]	27.01†	6.78†	25.10†	30.94	9.65	22.36
NeuSum [16]	26.78†	6.88†	25.14†	30.96	9.73	22.52
BiMPM [14]	24.77	6.11	22.82	30.92	9.55	24.43
CA [13]	24.51	5.98	22.64	31.16	9.58	24.49
COALA [10]	26.12	6.24	23.66	31.58	9.78	25.55
MMR [8]	26.78	6.05	23.56	30.11	9.02	24.39
AttSum [1]	26.36	6.29	24.01	31.24	9.77	25.34
QS* [6]	27.14	7.57	25.13	32.53	11.05	26.66
SD_2* [9]	26.65†	6.92†	24.77†	32.26	10.53	26.02
QPGN* [3]	27.32†	7.98†	25.46†	-	-	-
HSCM	**27.84**	**7.75**	**25.85**	**32.34**	**10.07**	**25.98**

Three extractive text summarization methods are adopted for comparison: **Lead3**, **NeuralSum** [2], **NeuSum** [16], which only conduct traditional text summarization task without question information. Besides, three answer selection models are adopted for comparison: **BiMPM** [14], **CA** [13], **COALA** [10], which selects sentences as the summary from the original answer by measuring the relevance degree between question and each answer sentence. Finally, we compare with five feature-free query-based summarization methods, including two extractive: **MMR** [8], **AttSum** [1], and three abstractive methods: **QS** [6], **SD_2** [9], **QPGN** [3]. These methods take into account the question information for guiding the answer summarization process.

3.2.2 ***Answer Summarzation Results.*** The experimental results on WikiHowQA and PubMedQA are presented in Table 2. There are several notable observations:

(1) The proposed method, HSCM, substantially and consistently outperforms all the extractive summarization methods and answer selection methods by a noticeable margin on both two datasets. This result demonstrates the superiority of taking into account both sequential context information in the answer and relevant information with the question for extractive answer summarization.

(2) Traditional extractive summarization methods perform better than answer selection methods on WikiHowQA, which is contrary to the performance on PubMedQA. We conjecture that the contextual information attaches more importance on WikiHowQA, while the similarity measurement is more useful on the other.

(3) HSCM achieves competitive performance with state-of-the-art abstractive query-based summarization methods in terms of ROUGE scores. More importantly, as an extractive method, HSCM is more computationally efficient, whose model parameters are 10x less and inference speed is 1,000x faster than those abstractive methods (e.g., QS, SD_2), and can generate more grammatically correct and coherent answer summaries.

3.3 Analysis

3.3.1 ***Ablation Study.*** In this section, we conduct ablation study to validate the effectiveness of different components of the proposed model. First, in order to analyze the effect of sequential context

Table 3: Ablation Study

Models	WikiHowQA			PubMedQA		
	R1	R2	RL	R1	R2	RL
HSCM	**27.8**	**7.8**	**25.9**	**32.3**	**10.1**	**26.0**
- sequential decoder	26.0	6.3	24.1	31.7	9.8	25.6
- compare-aggregate	26.8	7.0	25.0	30.9	9.6	23.9
- word-level	27.1	7.1	25.2	31.4	9.7	24.8
- sentence-level	27.5	7.3	25.4	31.7	9.8	25.2

modeling, we set apart the LSTM decoder and just perform hierarchical sentence pair modeling (w/o "sequential decoder"). In addition, we discard the compare-aggregate layers and only keep the question-aware sequential extractor model, and conduct experiments in terms of removing the word-level or sentence-level compare-aggregate layer to evaluate the effect of the hierarchical context modeling.

The ablation test results in Table 3 show that both hierarchical and sequential context modeling contribute to the final performance of answer summarization. As for WikiHowQA, the performance suffers a larger decrease when discarding the sequential decoder, indicating that the sequential context modeling plays a more important role on WikiHowQA, since there is rich context information preserved in the original answer with such length (over 500 words). Conversely, we observe that, as for PubMedQA, the hierarchical interaction modeling contributes more to the performance of answer summarization, as the question provides adequate information to facilitate the interactive context modeling between question and answer sentences. This analysis further explains the observation (2) in Section 3.2.2. In addition, both word-level and sentence-level compare-aggregate make contributions to the improvement, which validates the effectiveness of the hierarchical compare-aggregate method to encode the multi-level interaction between QA pairs.

3.3.2 ***Case Study.*** In order to intuitively observe the advantage of the proposed method, we randomly choose one example from WikiHowQA dataset to show the extractive summarization results. We compare the proposed method with one extractive summarization method, NeuralSum, and one answer selection method, CA. As shown in the Figure 2, NeuralSum tends to extract sentences which are important and informative in the context, while CA selects those sentences which are highly related to the question. As for the proposed method, it not only captures the sentence saliency but also consider the interaction with question information, which is more suitable in the answer summarization task.

4 CONCLUSIONS

In this paper, we bridge hierarchical and sequential context modeling for answer summarization to address the answer redundancy issue in non-factoid QA. We propose a hierarchical compare-aggregate method to encode the multi-level interaction information between question and answer, and then employ sequential learning to extract answer sentences for constructing the answer summary with the guidance of question. Experimental results show that the proposed method achieves superior performance on answer summarization for two non-factoid QA dataset, WikiHowQA and PubMedQA.

QUESTION: How to Resolve International Investment Disputes?

ANSWER: International investment disputes are very complicated. **Generally, you should already have a lawyer who represents you or your company in regular business disputes. they may have an international dispute practice.** If your current lawyer can not handle the dispute, then he or she should be able to find a specialist in international arbitration or international investment disputes. In a large law firm, there should be a group of lawyers who specialize in this field. **If your regular lawyer works in a smaller firm, then he or she could find lawyers in a larger firm who could represent you.** You could also get referrals by talking with other businesses that have been involved in international investment disputes. They can tell you whether they would recommend their lawyer. Because of the money at stake in international disputes, you should carefully screen your attorney to make sure that he or she has sufficient experience resolving international investment disputes. **For example, you should try to get the following information at the consultation: how many international disputes they have handled.** You will want someone who has handled several international arbitrations or trials. the size of the disputes they have handled. If your dispute is very large (say over \$ 100 million), then you will want someone experienced in large-scale arbitrations or trials. **Whether they think arbitration or a lawsuit is the best option. Be sure to bring copies of your contracts so that the lawyers can see whether or not arbitration is an option.** ...

SUMMARY: Find a lawyer. Attend a consultation. Discuss whether you want to arbitrate or sue at all.

Figure 2: Case Study. Bold / underlined / shadowed sentences are selected by HSCM / CA / NeuralSum, respectively.

REFERENCES

[1] Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. AttSum: Joint Learning of Focusing and Summarization with Neural Attention. In *COLING*. 547–556.

[2] Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. In *ACL*.

[3] Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2020. Joint Learning of Answer Selection and Answer Summary Generation in Community Question Answering. In *AAAI*.

[4] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *ACL*. 3558–3567.

[5] Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised Query-Focused Multi-Document Summarization using the Cross Entropy Method. In *SIGIR*. 961–964.

[6] Johan Hasselqvist, Niklas Helmertz, and Mikael Kågebäck. 2017. Query-Based Abstractive Summarization Using Neural Networks. *CoRR* abs/1712.06100 (2017).

[7] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *EMNLP-IJCNLP*. 2567–2577.

[8] Jimmy J. Lin, Nitin Madnani, and Bonnie J. Dorr. 2010. Putting the User in the Loop: Interactive Maximal Marginal Relevance for Query-Focused Summarization. In *HLT-NAACL*. 305–308.

[9] Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *ACL*. 1063–1072.

[10] Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. 2019. COALA: A Neural Coverage-Based Approach for Long Answer Selection with Small Data. In *AAAI*. 6932–6939.

[11] Mittul Singh, Arunav Mishra, Youssef Oualil, Klaus Berberich, and Dietrich Klakow. 2018. Long-Span Language Models for Query-Focused Unsupervised Extractive Text Summarization. In *ECIR*. 657–664.

[12] Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing Answers in Non-Factoid Community Question-Answering. In *WSDM*. 405–414.

[13] Shuohang Wang and Jing Jiang. 2017. A Compare-Aggregate Model for Matching Text Sequences. In *ICLR*.

[14] Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. In *IJCAI*. 4144–4150.

[15] Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2018. Document Summarization for Answering Non-Factoid Queries. *IEEE Trans. Knowl. Data Eng.* 30, 1 (2018), 15–28.

[16] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *ACL*. 654–663.