11-2020

Multi-hop inference for question-driven summarization

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

Wenxuan ZHANG

Wai LAM

Multi-hop Inference for Question-driven Summarization*

Yang Deng, Wenxuan Zhang, Wai Lam
The Chinese University of Hong Kong
{ydeng,wxzhang,wlam}@se.cuhk.edu.hk

Abstract

Question-driven summarization has been recently studied as an effective approach to summarizing the source document to produce concise but informative answers for non-factoid questions. In this work, we propose a novel question-driven abstractive summarization method, Multi-hop Selective Generator (MSG), to incorporate multi-hop reasoning into question-driven summarization and, meanwhile, provide justifications for the generated summaries. Specifically, we jointly model the relevance to the question and the interrelation among different sentences via a human-like multi-hop inference module, which captures important sentences for justifying the summarized answer. A gated selective pointer generator network with a multi-view coverage mechanism is designed to integrate diverse information from different perspectives. Experimental results show that the proposed method consistently outperforms state-of-the-art methods on two non-factoid QA datasets, namely WikiHow and PubMedQA.

1 Introduction

Recent years have witnessed several attempts on exploring question-driven summarization, which aims at summarizing the source document with respect to a specific question, to produce a concise but informative answer in non-factoid question answering (QA) (Tomasoni and Huang, 2010; Chan et al., 2012; Song et al., 2017). Unlike factoid QA (Rajpurkar et al., 2016), e.g., "Who is the author of Harry Potter?", whose answer is generally a single phrase or a short sentence with limited information, the answers for non-factoid questions are supposed to be more informative, involving some detailed analysis to explain or justify the final answers, such as questions in community QA (Ishida et al., 2018; Deng et al., 2020a) or explainable QA (Fan et al., 2019; Nakatsuji and Okui, 2020). As the example from PubMedQA (Jin et al., 2019) presented in Figure 1, the answer can be regarded as the summary over the document driven by the reasoning process of the given question.

Most of related studies focus on query-based summarization approaches for summarizing the query-related content from the source document (Shen and Li, 2011; Wang et al., 2013; Cao et al., 2016; Nema et al., 2017). However, these approaches fall short of tackling question-driven summarization problem in QA scenario, since the query-based summarization process is typically based on semantic relevance measurement without a careful reasoning or inference process, which is essential to question-driven summarization. Currently, question-driven summarization is mainly explored by traditional information retrieval methods to select sentences from the source document to construct the final answer (Wang et al., 2014; Song et al., 2017; Yulianti et al., 2018), which heavily rely on hand-crafted features or tedious multi-stage pipelines. Besides, compared to extractive summarization (Cao et al., 2016), abstractive methods (Nema et al., 2017) can produce more coherent and logical summaries to answer the given question. To this end, we study question-driven abstractive summarization to generate natural form of answers by summarizing the source document with respect to a specific question.

To tackle question-driven abstractive summarization, the content selection process for summarization is not only determined by the semantic relevance to the given question, but it also requires a human-like reasoning and inference process to consider the content interrelationship comprehensively and carefully across the whole source text for gener-

* The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Codes: 14200719).

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6734–6744,
November 16–20, 2020.

Question: ***Are human coronaviruses uncommon in patients with gastrointestinal illness?***

Document: <S>Coronaviruses infect numerous animal species causing a variety of illnesses including respiratory, neurologic and enteric disease. <S>***Human coronaviruses (HCoV) are mainly associated with respiratory tract disease but have been implicated in enteric disease.*** <S>***To investigate the frequency of coronaviruses in stool samples from children and adults with gastrointestinal illness by RT-PCR.*** <S>Clinical samples submitted for infectious diarrhea testing were collected from December 2007 through March 2008. <S>**RNA extraction and RT-PCR was performed for stools negative for Clostridium difficile using primer sets against HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1.** <S>Clinical data from samples positive for coronaviruses were reviewed and recorded. <S>Samples from 479 patients were collected including 151 pediatric (< or = 18 years), and 328 adults (>18 years). <S>Of these samples, 4 patients (1.3%, 2 adult; 2 pediatric) screened positive for the presence of a coronavirus. <S>**All detected coronaviruses were identified as HCoV-HKU1.** <S>**No stools screened positive for either HCoV-229E, HCoV-NL63 or HCoV-OC43.** <S>**All HCoV-HKU1 positive samples occurred between mid-January to mid-February.** <S>**Clinical manifestations from HCoV-HKU1 positive patients included diarrhea, emesis and respiratory complaints.** <S>Three (75%) patients were admitted to the hospital with a median length of stay of 6 days. <S>

Answer: ***Coronaviruses as a group are not commonly identified in stool samples of patients presenting with gastrointestinal illness.*** **HCoV-HKU1 can be identified in stool samples from children and adults with gastrointestinal disease, with most individuals having respiratory findings as well. No stool samples screened positive for HCoV-NL63, HCoV-229E, or HCoV-OC43.**

Figure 1: An example from PubMedQA. The **highlighted** sentences illustrate the inference process when humans answer the given question. ***Italic*** represents direct matching sentences from the question. **Underlined** and **wavy-underlined** represent sentences inferred by 2nd-hop and 3rd-hop reasoning, respectively, to justify the answer.

ating the summary. For instance, in Figure 1, given the specific question, there are several **highlighted** sentences required to be concentrated for conducting summarization so as to generate the answer. It leads to the necessity of measuring the importance of each sentence, instead of regarding the source text as an undifferentiated whole. Among these **highlighted** sentences, only the ***italic*** sentences are directly related to the given question, while other **highlighted** sentences need to be inferred from their interrelationships with other sentences. In other words, the generated summary is likely to lose important information, if we only focus on the semantically relevant content to the given question. Moreover, it can be observed that one-time inference sometimes is insufficient for collecting all the required information for producing a summary. In this example, the answer is summarized from both the ***1st-hop*** and **3rd-hop** inference sentences in the document, indicating the importance of multi-hop reasoning for content selection in question-driven summarization.

In this work, we propose a question-driven abstractive summarization model, namely Multi-hop Selective Generator (MSG), which incorporates multi-hop inference to summarize abstractive answers over the source document for non-factoid questions. Concretely, the document is regarded as a hierarchical text structure to be assessed with the importance degree in both word- and sentence-level for content selection. Then we develop a multi-hop inference module to enable human-like multi-hop reasoning in question-driven summarization, which considers the semantic relevance to the question as well as the information consistency among different sentences. Finally, a gated selective pointer generator network with multi-view coverage mechanism is proposed to generate a concise but informative summary as the answer to the given question.

The main contributions of this paper can be summarized as follows: (1) We propose a novel question-driven abstractive summarization model for generating answers in non-factoid QA, which incorporates multi-hop reasoning to infer the important content for facilitating answer generation; (2) We propose a multi-view coverage mechanism to address the repetition issue along with the multi-view pointer network and generate informative answers; (3) Experimental results show that the proposed method achieves state-of-the-art performance on WikiHow and PubMedQA datasets, and it is able to provide justification sentences as the evidence for the answer.

2 Related Works

Query-based Summarization. Early works on query-based summarization focus on extracting query-related sentences to construct the summary (Lin et al., 2010; Shen and Li, 2011), which are later improved by exploiting sentence compression on the extracted sentences (Wang et al., 2013; Li and Li, 2014). Recently, some data-driven neural abstractive models are proposed to generate natural form of summaries with respect to the given query (Nema et al., 2017; Hasselqvist et al., 2017). However, current studies on query-based abstractive summarization are restricted by the lack of large-scale datasets (Baumel et al., 2016; Nema et al., 2017). One the other hand, some researchers spark a new pave of question-driven summarization in non-factoid QA (Song et al., 2017; Yulianti

et al., 2018; Deng et al., 2020b), which requires the ability of reasoning or inference for supporting summarization, not merely relevance measurement, and also preserves remarkable testbeds of large-scale datasets.

Non-factoid Question Answering. Different from factoid QA that can be tackled by extracting answer spans (Rajpurkar et al., 2016) or generating short sentences (Nguyen et al., 2016; Kociský et al., 2018), non-factoid QA aims at producing relatively informative and complete answers. In the past studies, non-factoid QA focused on retrieval-based methods, such as answer sentence selection (Nakov et al., 2015) or answer ranking (Zhang et al., 2020). Recently, several efforts have been made on tackling long-answer generative question answering over supporting documents, which targets on questions that require detailed explanations (Fan et al., 2019). This kind of QA problem contains a large proportion of non-factoid questions, such as "how" or "why" type questions (Koupaee and Wang, 2018; Ishida et al., 2018; Deng et al., 2020a). Besides, some studies aim at generating a conclusion for the concerned question (Jin et al., 2019; Nakatsuji and Okui, 2020). Fan et al. (2019) propose a multi-task Seq2Seq model with the concatenation of the question and support documents to generate long-form answers. Iida et al. (2019) and Nakatsuji and Okui (2020) incorporate some background knowledge into Seq2Seq model for why questions and conclusion-centric questions. Some latest works (Feldman and El-Yaniv, 2019; Yadav et al., 2019; Nishida et al., 2019a) attempt to provide evidence or justifications for human-understandable explanation of the multi-hop inference process in factoid QA, where the inferred evidences are only treated as the middle steps for finding the answer. However, in non-factoid QA, the intermediate output is also important to form a complete answer, which requires a bridge between the multi-hop inference and summarization.

3 Proposed Framework

We propose a question-driven abstractive summarization model, namely Multi-hop Selective Generator (MSG). The overview of MSG is depicted in Figure 2, which consists of three main components: (1) *Co-attentive Encoder* (Section 3.1), (2) *Multi-hop Inference Module* (Section 3.2), and (3) *Gated Selective Generator* (Section 3.3). Moreover, *Multi-view Coverage Loss* is integrated to the overall training procedure (Section 3.4).

3.1 Co-attentive Encoder

Pre-trianed word embeddings, E_q and E_{s_i}, of the question q and each sentence s_i in the document $D = \{s_1, s_2, ..., s_n\}$ are input into the model. We first encode the question and each sentence in the document by a Bi-LSTM (Bidirectional Long Short-Term Memory Networks) shared encoder to learn the word-level contextual information, $H_q, H_{s_i} \in \mathbb{R}^{l \times d_h}$, where l and d_h denotes the sentence length and the dimension of the encoder output respectively. The overall word-level representations H_d for the document is sequentially concatenated by $[H_{s_1}, H_{s_2}, ..., H_{s_n}]$.

We compute the attention weights to align the word-level information between the question and the document sentences, and obtain the attention-weighted vectors of each word for both the question and the document sentences. For the question q and the i-th sentence s_i in the document D, we have:

$$O_{qs_i} = \tanh\left(H_q^T U H_{s_i}\right), \quad (1)$$

$$\alpha_{q_i} = \text{softmax}(\text{Max}(O_{qs_i})), \quad (2)$$

$$\alpha_{s_i} = \text{softmax}(\text{Max}({O_{qs_i}}^T)), \quad (3)$$

where $U \in \mathbb{R}^{d_h \times d_h}$ is the attention matrix to be learned; α_{q_i} and α_{s_i} are co-attention weights for the question and i-th sentence in the document.

We conduct dot product between the attention vectors and the word-level representations to generate the sentence representations for the question and the document:

$$M_q = \frac{1}{n} \sum_{i=1}^{n} H_q^T \alpha_{q_i} \quad (4)$$

$$M_s = [H_{s_1}^T \alpha_{s_1} : ... : H_{s_n}^T \alpha_{s_n}], \quad (5)$$

where M_q and M_s denote the sentence-level representations for the question and the document.

3.2 Multi-hop Inference Module

Multi-hop Inference Module measures the degree of importance for each sentence in the document to generate the answer, through a multi-hop reasoning procedure, which contains two kinds of inference units: Attentive Unit and MAR Unit.

3.2.1 Attentive Unit

Attentive Unit basically measures the matching degree between each sentence in the document and the given question by the following vanilla attention mechanism:

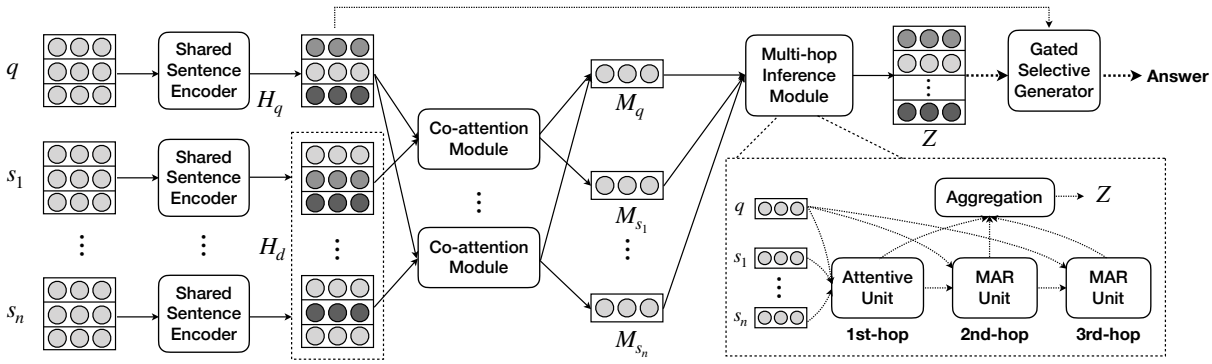


Figure 2: The overview of Multi-hop Selective Generator (MSG).

$$m_{dq} = \tanh(M_s W_m M_q), \quad (6)$$

$$\alpha_s = \text{softmax}(\omega_m^T m_{d_q}), \quad (7)$$

$$\textbf{Attentive}(M_s, M_q) = M_s \odot \alpha_s, \quad (8)$$

where W_m and ω_m are the attention matrices to be learned. α_s is the sentence-level attention weight which measures the matching degree of each document sentence with the given question. $\odot$ denotes the element-wise product for obtaining the attentive sentence-level representations for the document.

3.2.2 MAR Unit

Maximal Marginal Relevance (MMR) is an IR model that can be adopted to measure the query-relevancy and information-redundancy simultaneously for extractive summarization (Carbonell and Goldstein, 1998). However, as for the content selection in abstractive summarization, the relevance to both the question and the other sentences in the document should be taken into consideration for a high recall of selecting necessary content. Thus, we propose Maximal Absolute Relevance (MAR) to select highly salient sentences for generating the summary, which is formulated as:

$$\begin{aligned} mar_i = &\lambda \text{Sim}_1(M_{s_i}, M_q) + \\ &(1-\lambda) \max_{s_j \in D, j \neq i} \text{Sim}_2(M_{s_i}, M_{s_j}), \end{aligned} \quad (9)$$

where λ is a hyper-parameter for balancing the question-relevancy and information-consistency measurement. The relevance to the question is calculated by:

$$\text{Sim}_1(M_{s_i}, M_q) = M_{s_i} U_1 M_q, \quad (10)$$

where U_1 is a similarity matrix to be learned. We apply an attention mechanism over other sentences in the document to choose the highest relevance score, which can be regarded as the reasoning procedure where the next-hop justification sentences are supposed to be highly related to the last-hop justification sentences.

$$e_{ij} = \tanh(M_{s_i} U_2 M_{s_j}), \quad (11)$$

$$\text{Sim}_2(M_{s_i}, M_{s_j}) = \frac{exp(e_{ij})}{\sum_j exp(e_{ij})}, \quad (12)$$

where U_1 is a similarity matrix to be learned.

Then the weighted sentence representations are computed by the element-wise product of the original sentence representations and the MAR scores gated by a sigmoid function denoted as σ:

$$\textbf{MAR}(M_s, M_q) = M_s \odot \sigma(mar). \quad (13)$$

Overall, MAR Unit assigns higher weights to sentences in two situations: (i) Those sentences are correlated to the given question, due to the first term in Equation 9, (ii) Those sentences are consistent with the highly weighted justification sentences from the last hop, due to the second term.

3.2.3 Reasoning Procedure

In accordance with human-like multi-hop inference procedure, the first hop is supposed to capture the semantic-relevant sentences to the given question. Then the subsequent hops should consider not only the relevance to the question, but also the information-consistency with the previous attended sentences. Hence, the Attentive Unit is adopted as the 1st-hop inference unit, while the MAR Unit is served as the kth-hop unit, where $k > 1$. Before each hop, a Bi-LSTM layer is employed to refine the input sentence representation. For instance, a 3-hop inference procedure is as follows:

$$M_s^{(1)} = \textbf{Attentive}(\text{Bi-LSTM}(M_s), M_q), \quad (14)$$

$$M_s^{(2)} = \textbf{MAR}(\text{Bi-LSTM}(M_s^{(1)}), M_q), \quad (15)$$

$$M_s^{(3)} = \textbf{MAR}(\text{Bi-LSTM}(M_s^{(2)}), M_q). \quad (16)$$

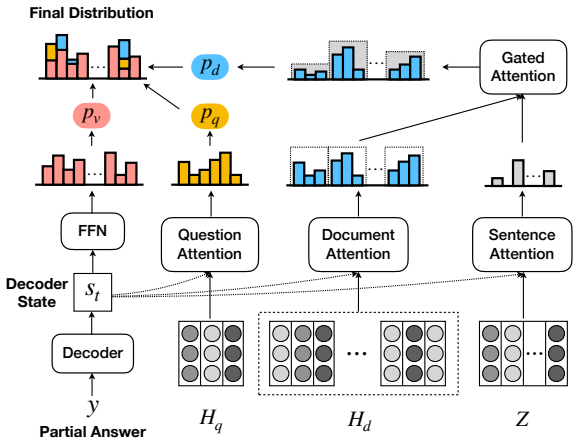


Figure 3: Gated Selective Pointer-Generator Network.

Then, we merge the 3-hop sentence representations, $\hat{M}_s = [M_s^{(1)}, M_s^{(2)}, M_s^{(3)}]$, via the following attention mechanism:

$$\alpha_h = \text{softmax}(\omega_h^T \tanh(W_h \hat{M}_s)), \quad (17)$$

$$Z = \hat{M}_s^T \alpha_h, \quad (18)$$

where W_h and ω_h are attention matrices to be learned. Z is the final sentence-level document representation for justifying the importance degree of each sentence in the decoding phase.

3.3 Gated Selective Generator

We obtain the word-level representations H_q and H_d for the question and document, respectively, from the encoding phase, and the sentence-level document representation Z via the multi-hop inference module. Figure 3 depicts the Gated Selective Pointer Generator Network in MSG.

A unidirectional LSTM is adopted as the decoder. At each step t, the decoder produces hidden state s_t with the input of the previous word w_{t-1}. The attention for each word in the question and the document, α_t^q and α_t^d, are generated by:

$$e_t^{q_j} = \omega_t^{qT} \tanh(W_q H_{q_j} + W_{qs} s_t + b_q), \quad (19)$$

$$\alpha_t^q = \text{softmax}(e_t^q), \quad (20)$$

$$e_t^{d_i} = \omega_t^{dT} \tanh(W_d H_{d_i} + W_{ds} s_t + b_d), \quad (21)$$

$$\alpha_t^d = \text{softmax}(e_t^d), \quad (22)$$

where W_q, W_{qs}, W_d, W_{ds}, ω_t^q, ω_t^d, b_q, b_d are parameters to be learned.

Then, we incorporate the multi-hop inference results Z to compute the gated attention weights β_t for each sentence in the document:

$$\beta_t = \sigma(\omega_t^{sT} \tanh(W_s Z_k + W_{ss} s_t + b_s)), \quad (23)$$

where W_s, W_{ss}, ω_t^s, b_s are parameters to be learned. We re-weight the word-level document attention scores α^d gated by the sentence-level document attention scores β to attend important justification sentences along with the decoding process:

$$\hat{\alpha}_t^{d_i} = \frac{\alpha_t^{d_i} \beta_{t,d_i \in s_k}}{\sum_i \alpha_t^{d_i} \beta_{t,d_i \in s_k}}. \quad (24)$$

Thus, the re-weighted word-level document attention $\hat{\alpha}^d$ naturally blends with the results from the multi-hop inference module to enhance the influence of those important justification sentences.

Finally, a multi-view pointer-generator architecture is designed to generate answers with multi-hop inference results as well as handle the multi-perspective out-of-vocabulary (OOV) issue. Such approach enables MSG to copy words from the question and be aware of the differential importance degree of different sentences in the document.

The attention weights α_t^q and $\hat{\alpha}_t^d$ are used to compute context vectors c_t^q and c_t^d as the probability distribution over the source words:

$$c_t^q = H_q^T \alpha_t^q, \quad c_t^d = H_d^T \hat{\alpha}_t^d. \quad (25)$$

The context vector aggregates the information from the source text for the current step. We concatenate the context vector with the decoder state s_t and pass through a linear layer to generate the answer representation h_t^s:

$$h_t^s = W_1[s_t : c_t^q : c_t^d] + b_1, \quad (26)$$

where W_1 and b_1 are parameters to be learned.

Then, the probability distribution P^v over the fixed vocabulary is obtained by passing the answer representation h_t^s through a softmax layer:

$$P^v(y_t) = \text{softmax}(W_2 h_t^s + b_2), \quad (27)$$

where W_2 and b_2 are parameters to be learned.

The final probability distribution of y_t is obtained from three views of word distributions:

$$P^q(y_t) = \sum\nolimits_{i:w_i=w} \alpha_t^{q_i}, \quad (28)$$

$$P^d(y_t) = \sum\nolimits_{i:w_i=w} \hat{\alpha}_t^{d_i}, \quad (29)$$

$$P^{all}(y_t) = [P^v(y_t), P^q(y_t), P^d(y_t)], \quad (30)$$

$$\rho = \text{softmax}(W_\rho[s_t : c_t^q : c_t^d] + b_\rho), \quad (31)$$

$$P(y_t) = \rho \cdot P^{all}(y_t), \quad (32)$$

where W_ρ and b_ρ are parameters to be learned, ρ is the multi-view pointer scalar to determine the weight of each view of the probability distribution.

3.4 End-to-end Training

Multi-view Coverage Loss. The original *coverage mechanism* (See et al., 2017) could only prevent repeated attention from one certain source text. However, the repetition problem becomes more severe, as we leverage both the question and document as the source text. Besides, similar to multi-view pointer network, coverage losses of different sources are supposed to be weighted by their contribution. Therefore, we design a multi-view coverage mechanism to address this issue as well as balance the generating and copying processes.

In each decoder timestep t, the coverage vector $c_t = \sum_{t'=0}^{t-1} a_{t'}$ is used to represent the degree of coverage so far. The coverage vector c_t will be applied to compute the attention weight α_t in Equations 19 and 21. The coverage loss is trained to penalize the repetition in updated attention weight α^t from all views. The re-normalized pointer weights $\hat{\rho} = \rho^c / \sum_{c \in \{q,d\}} \rho^c$ are employed to balance the coverage loss of different views:

$$L_{cov} = \sum \hat{\rho} \frac{1}{T} \sum_{t=1}^{T} \sum_i \min(\alpha_t^i, c_t^i). \quad (33)$$

Overall Loss Function. The overall model is trained to minimize the negative log likelihood and the multi-view coverage loss:

$$L = -\frac{1}{T} \sum_{t=0}^{T} \log P(w_t^*) + \lambda L_{cov}, \quad (34)$$

where λ is a hyper-parameter to balance losses.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate on a large-scale summarization dataset with non-factoid questions, WikiHow (Koupaee and Wang, 2018), and a non-factoid QA dataset with abstractive answers, PubMedQA (Jin et al., 2019). WikiHow is an abstractive summarization dataset collected from a community-based QA website, *WikiHow*[1], in which each sample consists of a non-factoid question, a long article, and the corresponding summary as the answer to the given question. PubMedQA is a conclusion-based biomedical QA dataset collected from *PubMed*[2] abstracts, in which each instance is composed of a question, a context, and an abstractive answer which is the summarized conclusion of the context corresponding to the question. The statistics of the WikiHow and PubMedQA datasets are shown in Table 1 [3]. We adopt ROUGE F1 (R1, R2, RL) for automatically evaluating the summarized answers. Besides, human evaluation and Distinct scores are adopted for analysis.

Dataset (train/dev/test)	WikiHow	PubMedQA
#Samples	168K / 6K / 6K	169K / 21K / 21K
Avg QLen	7.00 / 7.02 / 7.01	16.3 / 16.4 / 16.3
Avg DLen	582 / 580 / 584	238 / 238 / 239
Avg ALen	62.2 / 62.2 / 62.2	41.0 / 41.0 / 40.9
Avg #Sents/Doc	20.7 / 20.7 / 20.6	9.32 / 9.31 / 9.33

Table 1: Statictis of Dataset

4.2 Baseline Methods and Implementations

To evaluate the proposed method, we compare with several baselines and state-of-the-art methods on query-based abstractive summarization and generative QA. We first employ four widely-adopted summarization baseline methods, including two unsupervised extractive methods, **LEAD3** and **MMR**, and two abstractive methods, **S2SA** (Bahdanau et al., 2015), and **PGN** (See et al., 2017).

Then two popular query-based abstractive summarization methods are adopted for evaluation: (1) **SD**$_2$ (Nema et al., 2017), which is a sequence-to-sequence model with a query attention, and (2) **QS** (Hasselqvist et al., 2017), which incorporates question information into the pointer-generator network with the vanilla attention mechanism.

Finally, we implement two latest generative QA models for comparisons: (1) **S2S-MT** (Fan et al., 2019), which uses a multi-task Seq2Seq model with the concatenation of question and support document, and (2) **QPGN** (Deng et al., 2020a), which is a question-driven pointer-generator network with co-attention between the question and document.

We train all the models with pre-trained GloVE embeddings[4] of 300 dimensions and set the vocabulary size to 50k. During training and testing procedure, we restrict the length of generated summaries within 50 words. As for the proposed method, we train with a learning rate of 0.15 and an initial accumulator value of 0.1. The dropout rate is set to 0.5. The hidden unit sizes of the BiLSTM encoder and the LSTM decoder are all set to 256. We train our models with the batch size of 32. All other parameters are randomly initialized from [-0.05, 0.05]. Similar to the original coverage loss (See

[1] https://www.wikihow.com
[2] https://www.ncbi.nlm.nih.gov/pubmed/
[3] https://github.com/dengyang17/msg
[4] http://nlp.stanford.edu/data/glove.42B.zip

Model	WikiHow			PubMedQA		
	R1	R2	RL	R1	R2	RL
LEAD3	26.0*	7.2*	24.3*	30.9	9.8	21.2
MMR	26.8	6.1	23.6	30.1	9.0	24.4
S2SA	22.0*	6.3*	20.9*	32.4	11.0	27.3
PGN	28.5*	9.2*	26.5*	32.9	11.5	28.1
SD_2	27.7	7.9	25.8	32.3	10.5	26.0
QS	28.8	9.9	27.6	32.6	11.1	26.7
S2S-MT	28.6	9.6	27.5	33.2	12.2	27.8
QPGN	28.8	9.7	27.7	34.2	12.8	28.7
MSG (1-Hop)	30.0	10.2	29.0	36.5	14.4	30.0
MSG (2-Hop)	30.2	10.3	29.1	37.0	14.7	**30.4**
MSG (3-Hop)	**30.5**	**10.5**	**29.3**	**37.2**	**14.8**	30.2

Table 2: Results on WikiHow and PubMedQA. * represents results reported from Koupaee and Wang (2018).

Model	Info	Conc	Read	Corr
SD_2	3.48	3.34	3.30	3.04
QS	3.62	3.30	3.48	3.24
QPGN	3.58	3.52	3.68	3.32
MSG	**4.14**	**3.88**	**3.82**	**3.78**

Table 3: Human Evaluation Results

et al., 2017), we first train the model without multi-view coverage loss for 20 epochs, and then train with it for another 5 epochs with λ as 0.1.

4.3 Performance Comparison

Table 2 summarizes the experimental results on both datasets. As for WikiHow, which is an abstractive summarization dataset with non-factoid questions, current query-based summarization (SD_2, QS) and generative QA approaches (S2S-MT, QPGN) barely improve the performance from traditional summarization approaches. It indicates that the question information is not fully exploited for summarization, while MSG outperforms all these methods with a noticeable margin, about 2%.

Besides, since PubMedQA is a QA dataset with abstractive answers, we can observe that QPGN, which employs special design for modeling the interaction between the question and document, achieves relatively better performance than other summarization methods. Favorably MSG raises the state-of-the-art result by about 3%. Furthermore, MSG achieves promising improvements via the multi-hop inference on these two datasets.

We conduct human evaluation to evaluate the generated answer from four aspects: (1) Informativity: how rich is the generated answer in information? (2) Conciseness: how concise is the summary? (3) Readability: how fluent and coherent is the summary? (4) Correctness: how well does the generated answer respond to the given question? We randomly sample 50 questions from two datasets and generate their answers with three query-based summarization methods, including SD_2, QS, QPGN and the proposed MSG. Three annotators are asked to score each generated answer with 1 to 5 (higher the better). Results are presented in Table 3. We observe that MSG consistently and substantially outperforms existing query-based summarization methods in all aspects, especially for the informativeness and correctness. The results show that MSG effectively generates concise but also informative answers, since MSG not only considers question-related information, but also captures logically necessary content for answering the given question via multi-hop reasoning. Consequently, it leads to a more precise answer.

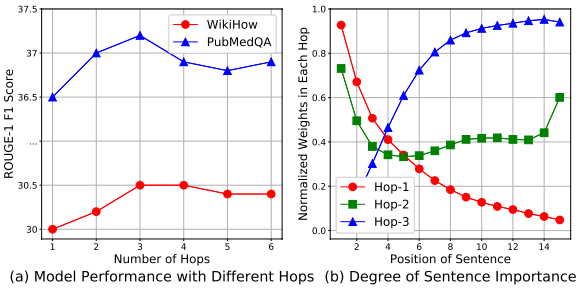
Model	WikiHow		PubMedQA	
	R1	RL	R1	RL
MSG (3-Hop)	30.5	29.3	37.2	30.2
- multi-hop inference	29.5	28.4	35.7	29.2
- hops aggregation[1]	30.1	29.0	37.0	30.1
- hops attention	30.3	29.2	37.0	30.1
- MAR unit[2]	30.0	29.1	36.8	30.0
- co-attention	30.2	29.0	37.0	30.1
- gated attention[3]	30.2	28.9	36.6	29.8
- question pointer	30.3	29.1	35.5	29.1
- MVC loss	29.6	28.5	35.9	29.3

Table 4: Ablation Study on Model Components. [1]Use the sentence representation learned from the last hop, instead of merging all the hops. [2]Replace all the MAR Unit with Attentive Unit. [3]Replace the *sigmoid* function with *softmax* function.

5 Discussions

5.1 Ablation Study

We conduct ablation study to validate the effectiveness of different components in MSG as well as the detailed design for the multi-hop inference module. The upper part in Table 4 presents the ablation study on multi-hop inference module. First of all, the model performance suffers a great decrease from discarding the multi-hop inference module on two datasets, showing the necessity of incorporating the multi-hop reasoning into the question-driven summarization. In specific, the fusion of the selective sentence representations from all hops brings performance improvement, including aggre-



(a) Model Performance with Different Hops (b) Degree of Sentence Importance

Figure 4: Analysis of Multi-hop Reasoning

gating all the hops as well as applying attention to weight the importance of each hop. Besides, it also achieves better performance to apply the proposed MAR Unit as the multi-hop unit, instead of repeatedly using Attentive Unit, indicating that it is not enough to only consider the question-related information, while the interrelationship among different sentences also attaches great importance.

The second part in Table 4 presents the ablation study in terms of discarding other model components in MSG. In general, all the components contribute to the final performance to a certain extent. In detail, there are several notable observations: (1) Some existing works (Hsu et al., 2018; Nishida et al., 2019b) apply softmax function to normalize the weights of different sentences in the decoding phase, which falls short of differentiating the importance degree of each sentence. The result shows that MSG achieves better performance by employing gated attention to distinguish salient justification sentences for generating the summaries. (2) Discarding the question pointer casts a noticeably greater decrease on PubMedQA than WikiHow. We conjecture that those questions from PubMedQA contain more words available to be copied for generating precise summaries, as the statistic of the question length shown in Table 1. These results also validate the importance of multi-view PGN on question-driven abstractive summarization, which is underutilized in current methods. (3) Multi-view coverage (MVC) loss makes a great contribution to the performance by alleviating the severe repetition problem along with the multi-view PGN.

5.2 Analysis of Multi-hop Reasoning

As the results presented in Section 4.3, MSG (3-Hop) outperforms MSG (1-Hop) by 0.5% and 0.7% on WikiHow and PubMedQA, respectively, indicating the effectiveness of incorporating multi-hop reasoning in question-driven summarization. Figure 4(a) presents the model performance in terms of using different hops of reasoning. We can see that, as expected, the performance of the model begins with growth when increasing the number of hops for reasoning. However, the performance becomes generally unchanged (e.g., WikiHow) or even slightly decreases (e.g., PubMedQA) when we further increase the number of hops. In practice, it is actually unnecessary to reason for too many hops, which may cause over-fitting. And adopting 3-hops in the implementation can be regarded as a hyper-parameter that is tuned on the datasets.

In addition, we extract and normalize the sentence weights from Eq. 7&9 to analyze some characteristics of the justification sentences in multi-hop inference. Figure 4(b) summarizes the statistic result of the sentence importance degree in each hop. We observe that the most important sentences in the 1st-hop of reasoning are likely to appear at the beginning of the document, while those in the 3rd-hop are concentrated in the latter part of the document. Comparatively, the important sentences in the 2nd-hop appear equally in all positions of the document. The results show that the proposed multi-hop inference procedure of justification sentences is generally in accordance with human-like reading habits.

5.3 Case Study

We present a case study in Figure 5 with generated answers from the proposed method and some baseline methods, QPGN, QS, and SD_2, to intuitively compare these methods. With the multi-hop reasoning process in MSG, we can obtain a clear clue of how to answer the given question. As it can be observed that the reference answer is composed of the information from the ***1st-hop*** and **3rd-hop** inference sentences, it is inadequate to simply summarize the question-related content for generating the answer. For the generated summaries, there are several observations as follows: (1) MSG (3-hop) successfully summarizes the source document with all the necessary and correct information. (2) MSG (2-hop) also effectively summarizes the 1st-hop and 2nd-hop inference content in the document. However, in this case, 3-hop inference is required to answer the given question. (3) MSG (1-hop) only measures the semantic relevance to the given question, leading to an incomplete summary that is lack of some necessary content, and even introduces some general sentences due to the data-driven learning. (4) QPGN only considers the

Question: ***Does high molecular weight hyaluronan decrease oxidative DNA damage induced by EDTA in human corneal epithelial cells?***
Document: <S>***To investigate the toxic effects of Ethylenediaminetetraacetic Acid Disodium Salt (EDTA), a corneal penetration enhancer in topical ophthalmic formulations, on DNA in human corneal epithelial cells (HCEs), and to investigate whether the effect induced by EDTA can be inhibited by high molecular weight hyaluronan (HA).*** <S>Cells were exposed to EDTA in concentrations ranging from 0.00001 to 0.01% for 60 min, or 30 min high molecular weight HA pretreatment followed by EDTA treatment. <S>…<S>EDTA exhibited no adverse effect on cell viability and did not induce cell apoptosis in human corneal epithelial cells at concentrations lower than 0.01%. <S>**However, a significant increase of DNA single- and double-strand breaks was observed in a dose-dependent manner with all the concentrations of EDTA tested in HCEs.** <S> **In addition, EDTA treatment led to elevated ROS generation.** <S>**Moreover, 30 min preincubation with high molecular weight HA significantly decreased EDTA-induced ROS generation and DNA damage.** <S>
Reference Answer: ***EDTA could induce DNA damage in HCEs, probably through oxidative stress.*** **Furthermore, high molecular weight HA was an effective protective agent that had antioxidant properties and decreased DNA damage induced by EDTA.**
MSG (3-Hop): ***High molecular weight HA reduces oxidative DNA damage induced by EDTA in human corneal epithelial cells.*** **Moreover, high molecular weight HA significantly decreased EDTA-induced ROS generation and DNA damage.**
MSG (2-Hop): ***High molecular weight HA pretreatment followed by EDTA reduces oxidative DNA damage induced by EDTA in human corneal epithelial cells.*** However, a significant increase of DNA single- and double-strand breaks was observed in a dose-dependent manner with high molecular weight HA.
MSG (1-Hop): **High molecular weight HA significantly reduced EDTA-induced ROS generation and DNA damage in human corneal epithelial cells.** These results suggest that EDTA may be a potential therapeutic agent for the treatment of corneal penetration.
QPGN: ***In addition to the corneal penetration of HCEs, EDTA can induce cell apoptosis in human corneal epithelial cells, and the effect induced by EDTA in human corneal epithelial cells can be inhibited by EDTA.***
QS: EDTA induces cell apoptosis in human corneal epithelial cells, suggesting that EDTA may be a potential therapeutic agent for the treatment of corneal epithelial cells, in the prevention of DNA damage in the corneal epithelial cell population.
SD2: These results suggest that EDTA may be a potential therapeutic agent for the treatment of human corneal epithelial damage caused by EDTA in the topical ophthalmic formulation of topical ophthalmic formulations.

Figure 5: A case study with the same legend as Figure 1. The **highlighted** sentences are attended by MSG (3-hop).

semantic relevance to the given question, leading to an incomplete summary that is lack of some necessary content. (5) QS and SD_2 fail to capture the key information, resulting in generating irrelevant summaries to the given question, or producing some general sentences due to the data-driven learning. It shows the capability of MSG to implement multi-hop reasoning and provide justification sentences.

Additionally, we observe that many cases probably require more than 3-hop inference or only involve one or two hops. However, we can still evaluate how MSG works in these cases. Compared to the reference answer, MSG (3-hop) can still capture most of the useful information to generate a good summary for answering the question. Besides, MSG (2-hop) and MSG (1-hop) also manage to attend some important content in the document. In general, our model is able to only attend a single hop if one-hop is enough, while our model may regard several hops as an integral hop when more hops are required. However, the baseline methods introduce much unnecessary or even incorrect information into the summarized answers.

5.4 Duplication Analysis in Answers

We adopt Distinct scores to analyze whether the multi-view coverage mechanism can alleviate the repetition issue in the generation procedure of multi-view PGN. Figure 6 summarizes the percentage of n-grams duplication on the ground-truth answers and the generated answers with or without the original (See et al., 2017) and multi-view coverage mechanism. We observe that the original

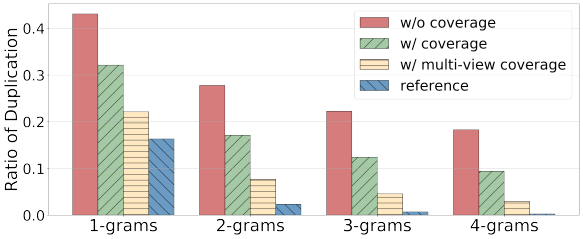


Figure 6: Duplication Analysis in Answers

coverage mechanism can still reduce word repetition in multi-view PGN. Moreover, multi-view coverage further reduces the ratio of duplication to a great extent, since multi-view coverage not only prevents repeatedly attending to the same element in both question and document, but also balances the weight of penalty between them.

6 Conclusion

We propose a novel question-driven abstractive summarization method, Multi-hop Selective Generator (MSG), to summarize concise but informative answers for non-factoid QA. We incorporate multi-hop reasoning to infer justification sentences for abstractive summarization. Experimental results show that the proposed method achieves state-of-the-art performance on two benchmark non-factoid QA datasets, namely WikiHow and PubMedQA.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd Inter-*

national Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2016. Topic concentration in query focused summarization datasets. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2573–2579.

Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. Attsum: Joint learning of focusing and summarization with neural attention. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 547–556.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336.

Wen Chan, Xiangdong Zhou, Wei Wang, and Tat-Seng Chua. 2012. Community answer summarization for multi-sentence question with group L1 regularization. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 582–591.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2020a. Joint learning of answer selection and answer summary generation in community question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7651–7658.

Yang Deng, Wenxuan Zhang, Yaliang Li, Min Yang, Wai Lam, and Ying Shen. 2020b. 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 2020, Virtual Event, China, July 25-30, 2020*, pages 1693–1696.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2296–2309.

Johan Hasselqvist, Niklas Helmertz, and Mikael Kågebäck. 2017. Query-based abstractive summarization using neural networks. *CoRR*, abs/1712.06100.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 132–141.

Ryu Iida, Canasai Kruengkrai, Ryo Ishida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Exploiting background knowledge in compact answer generation for why-questions. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 142–151.

Ryo Ishida, Kentaro Torisawa, Jong-Hoon Oh, Ryu Iida, Canasai Kruengkrai, and Julien Kloetzer. 2018. Semi-distantly supervised neural model for generating compact answers to open-domain why questions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5803–5811.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Yanran Li and Sujian Li. 2014. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1197–1207.

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 *Human Language Technologies: Con-*

ference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA, pages 305–308.

Makoto Nakatsuji and Sohei Okui. 2020. Conclusion-supplement answer generation for non-factoid questions. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8520–8527.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, James R. Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 269–281.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1063–1072.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019a. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2335–2345.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019b. Multi-style generative reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2273–2284.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Chao Shen and Tao Li. 2011. Learning to rank for query-focused multi-document summarization. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 626–634.

Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing answers in non-factoid community question-answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, pages 405–414.

Mattia Tomasoni and Minlie Huang. 2010. Metadata-aware measures for answer summarization in community question answering. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 760–769.

Lu Wang, Hema Raghavan, Claire Cardie, and Vittorio Castelli. 2014. Query-focused opinion summarization for user-generated content. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1660–1669.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A sentence compression based framework to query-focused multi-document summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1384–1394.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2578–2589.

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):15–28.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020. Answer ranking for product-related questions via multiple semantic relations modeling. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 569–578.