WSL-DS: Weakly Supervised Learning with Distant Supervision for Query Focused Multi-Document Abstractive Summarization

Md Tahmid Rahman Laskar[1, 3], Enamul Hoque[2], Jimmy Xiangji Huang[2, 3]
[1] Department of Electrical Engineering and Computer Science, York University
[2] School of Information Technology, York University
[3] Information Retrieval and Knowledge Management Research Lab, York University
Toronto, Ontario, Canada
tahmedge@cse.yorku.ca, enamulh@yorku.ca, jhuang@yorku.ca

Abstract

In the Query Focused Multi-Document Summarization (QF-MDS) task, a set of documents and a query are given where the goal is to generate a summary from these documents based on the given query. However, one major challenge for this task is the lack of availability of labeled training datasets. To overcome this issue, in this paper, we propose a novel weakly supervised learning approach via utilizing distant supervision. In particular, we use datasets similar to the target dataset as the training data where we leverage pre-trained sentence similarity models to generate the weak reference summary of each individual document in a document set from the multi-document gold reference summaries. Then, we iteratively train our summarization model on each single-document to alleviate the computational complexity issue that occurs while training neural summarization models in multiple documents (i.e., long sequences) at once. Experimental results in Document Understanding Conferences[1] (DUC) datasets show that our proposed approach sets a new state-of-the-art result in terms of various evaluation metrics.

1 Introduction

With the rapid growth of textual documents on the internet, accessing information from the web has become a challenging issue (Yao et al., 2017). Often users want the summary of a topic from various sources to fulfill their information needs (Feigenblat et al., 2017). The QF-MDS task deals with such problems where the goal is to summarize a set of documents to answer a given query.

In the QF-MDS task, the summaries generated by the summarizer can be either extractive or abstractive (Yao et al., 2017; Kulkarni et al., 2020). An extractive summarizer extracts relevant text spans from the source document(s), whereas an abstractive summarizer generates a summary in natural language that may contain some words which did not appear in the source document(s) (Rush et al., 2015; Nallapati et al., 2016; Nema et al., 2017). With the rising popularity of virtual assistants in recent years, there is a growing interest to integrate abstractive summarization capabilities in these systems for natural response generation (Nishida et al., 2019).

One major challenge for the QF-MDS task is that the datasets (e.g., DUC 2005, 2006, 2007) used for such tasks do not contain any labeled training data. Therefore, neural summarization models that leverage supervised training cannot be used in these datasets. Note that for other related tasks (Allan et al., 2003; Liu et al., 2008; Miao et al., 2012), how to reduce the demands for labeling the data and how to leverage unlabeled data were also identified as a major challenge. While using datasets similar to the target dataset as the training data for the QF-MDS task, we find that these datasets only contain multi-document gold summaries. However, the state-of-the-art transformer-based (Vaswani et al., 2017) summarization models (Liu and Lapata, 2019; Laskar et al., 2020a) cannot be used in long documents due to computational complexities (Beltagy et al., 2020; Zaheer et al., 2020). To tackle these issues, we propose a novel weakly supervised approach by utilizing distant supervision to generate weak reference summary of each single-document from multi-document gold reference summaries. We train our model on each document with weak supervision and find that our proposed approach that generates abstractive summaries is very effective for the QF-MDS task. More concretely, we make the following contributions:

[1] https://duc.nist.gov/

- First, to address the issue of unlabeled individual documents in a training document set, we utilize pre-trained sentence similarity models (Liu et al., 2019; Laskar et al., 2020b) to generate the weak reference summary of each individual document from multi-document gold reference summaries.

- Second, to address the computational issue to train neural models in long documents (Zaheer et al., 2020; Beltagy et al., 2020), we propose an iterative approach that adopts a pre-trained single-document generic summarization model to leverage the effectiveness of fine-tuning such models for query focused abstractive summarization (Laskar et al., 2020a) and extends it for the QF-MDS task.

- Experimental results on DUC 2005-07 datasets show that our proposed approach sets a new state-of-the-art result in terms of various ROUGE scores. As a secondary contribution, we will make our source codes publicly available here: `https://github.com/tahmedge/WSL-DS-COLING-2020`.

2 Related Work

Early work on multi-document summarization was mostly focused on generic summarization (Nayeem et al., 2018), whereas the amount of work for QF-MDS had been very limited (Yao et al., 2017). Due to the lack of training data for the QF-MDS task, most previous works were based on various unsupervised approaches that could only generate extractive summaries (Wang et al., 2008; Haghighi and Vanderwende, 2009; Wan and Xiao, 2009; Yao et al., 2015; Zhong et al., 2015; Wan and Zhang, 2014; Ma et al., 2016; Feigenblat et al., 2017).

To generate the abstractive summaries for the QF-MDS task, (Baumel et al., 2018) proposed a transfer learning technique to tackle the issue of no training data. They adopted the Pointer Generation Network (PGN) (See et al., 2017) pre-trained for the generic abstractive summarization task in a large dataset to predict the query focused summaries in the target dataset via modifying the attention mechanism of the PGN model. However, their model failed to outperform different extractive approaches in terms of various ROUGE scores (Feigenblat et al., 2017; Roitman et al., 2020).

Identifying sentences which are relevant to the query is an important step for the QF-MDS task. For this purpose, various approaches were utilized such as counting word overlaps (Baumel et al., 2018) or the Cross-Entropy Method (Feigenblat et al., 2017). Though neural models based on supervised training have significantly outperformed various non-neural models for the answer selection task in recent years (Laskar et al., 2019; Laskar et al., 2020b), such neural models have not been effectively used for the QF-MDS task yet due to the absence of labeled data for the relevant sentences in the QF-MDS datasets.

Recently, (Garg et al., 2019) showed that neural models pre-trained in a large Question Answering (QA) dataset could effectively select answers in other QA datasets. More recently, such pre-trained answer selection models for the QF-MDS task were used by (Xu and Lapata, 2020). In their work, they utilized distant supervision from various QA datasets using the fine-tuned BERT (Devlin et al., 2019) model to filter out the irrelevant sentences from the documents. However, (Baumel et al., 2018) showed that filtering sentences as an early step could lead to performance deterioration for the QF-MDS task. Thus, instead of applying distant supervision to filter out some sentences from the document, we apply it to generate the weak reference summary of each unlabeled document in our training datasets. Our proposed weakly supervised learning approach not only allows us to leverage the advantage of fine-tuning pre-trained generic summarization models (Laskar et al., 2020a), but also allows us to overcome the limitation of training neural models in long documents (Beltagy et al., 2020; Zaheer et al., 2020).

3 Our Proposed Approach

Suppose, we have a query $Q = q_1, q_2, ..., q_k$ containing k words and a set of N documents $D = d_1, d_2, ..., d_N$. For the QF-MDS task, the goal is to generate a summary $S = s_1, s_2, ...s_n$ containing n words from the document set D for the given query Q.

In figure 1, we show the overall architecture of our proposed approach. Since there is no training data available for the QF-MDS task, we provide supervised training to our target dataset by using other QF-MDS datasets as the training data. However, the available QF-MDS datasets (Feigenblat et al.,

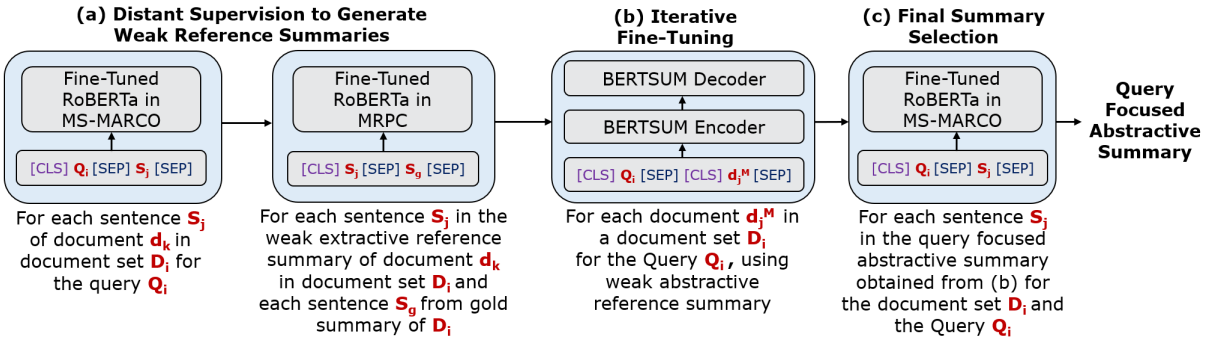

Figure 1: An overview of our model that generates (a) weak reference summary using **RoBERTa** for (b) iterative fine-tuning using **BERTSUM** to generate query focused abstractive summaries which are then ranked by (c) **RoBERTa**. [CLS] and [SEP] are the special tokens used with inputs (Devlin et al., 2019).

2017) only contain the gold summaries generated by human experts from multiple documents and do not contain the gold summary of each individual document. Due to the limitations of using neural models in long documents (Beltagy et al., 2020; Zaheer et al., 2020), we propose an iterative approach to train our model on each document of a document set. For this purpose, we generate the weak reference summary of each document from the multi-document gold summaries using distant supervision to train our model for the QF-MDS task. Finally, we rank the generated query focused summaries via an answer selection model (Laskar et al., 2020b). In the following, we give a detailed description of our proposed approach.

3.1 Weakly Supervised Learning with Distant Supervision

To generate the weak reference summaries using distant supervision, we utilize the pre-trained RoBERTa model (Liu et al., 2019) in two steps (see Figure 1a). At first, we generate the weak extractive reference summary of each individual document using a RoBERTa sentence similarity model fine-tuned for the *answer selection* task. Then, we measure the similarity score between each sentence in the human written (abstractive) multi-document gold summaries with each sentence in the weak extractive reference summary using a RoBERTa sentence similarity model fine-tuned for the *paraphrase identification* task. Based on the similarity score, we select the most relevant sentences from the gold reference summaries as the weak abstractive reference summary for each document. Below we describe these steps in detail.

RoBERTa Answer Selection Model: In this step, we first generate the initial weak extractive reference summary of each individual document d_k by measuring the relevance between the query Q_i and each sentence S_j in d_k. For this purpose, we adopt the RoBERTa sentence similarity model from (Laskar et al., 2020b) for its impressive performance in the answer sentence selection task and fine-tune it in the QA-ALL dataset of MS-MARCO (Bajaj et al., 2016). The fine-tuned $\text{RoBERTa}_{\text{MS-MARCO}}$ model was then utilized in our training dataset to measure the similarity score between each sentence in the document and the query. Based on the similarity score, we select the top $K = 3$ most relevant sentences as the weak extractive reference summary. Note that we use the value of $K = 3$ because extracting only 3 sentences was found effective in different extractive summarizers such as the LEAD-3 baseline (See et al., 2017; Liu and Lapata, 2019), as well as the $\text{BERTSUM}_{\text{EXT}}$ model (Liu and Lapata, 2019).

RoBERTa Paraphrase Identification Model: We provide distant supervision to generate the weak abstractive reference summary by replacing each sentence in the weak extractive reference summary (generated in the previous step) with the most similar sentence found in the multi-document gold summaries. For this purpose, we fine-tune the RoBERTa model for the paraphrase identification task in the MRPC dataset (Liu et al., 2019). Then for each document d_k in a document set D_i, we utilize the fine-tuned $\text{RoBERTa}_{\text{MRPC}}$ paraphrase identification model to replace each sentence S_j in the weak extractive reference summary of d_k with the most similar sentence S_g found in the gold summaries (among the sentences that are not already selected for the same document) of D_i.

(a) F1 Score:	DUC 2005			DUC 2006			DUC 2007		
Model	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
(Feigenblat et al., 2017) *	37.78	7.45	13.02	40.47	9.13	14.73	42.86	11.34	16.53
(Xu and Lapata, 2020) *	-	-	-	41.6	9.5	15.3	43.3	11.6	16.8
(Roitman et al., 2020) *	38.08	7.54	13.17	41.23	9.47	14.97	43.24	11.78	16.83
PQSUM$_{EXT}$ *	37.52	7.84	13.29	40.68	9.29	14.66	42.57	11.20	15.98
PQSUM$_{ABS}$	38.35	7.94	13.44	40.87	9.43	14.83	42.17	10.82	15.98
PQSUM$_{WSL\text{-}DS}$	**40.32**	**9.17**	**14.73**	**43.49**	**10.78**	**16.45**	**44.72**	**12.44**	**17.72**

(b) Recall:	DUC 2005			DUC 2006			DUC 2007		
Model	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
(Feigenblat et al., 2017) *	40.35	7.94	13.91	43.01	9.69	15.65	45.45	12.02	17.54
(Baumel et al., 2018)	39.82	6.98	**15.73**	42.89	8.73	**17.75**	43.92	10.13	**18.54**
(Roitman et al., 2020) *	**40.82**	8.07	14.13	**43.94**	10.09	15.96	**46.02**	**12.53**	17.91
PQSUM$_{EXT}$ *	37.55	7.84	13.31	40.41	9.22	14.56	42.41	11.08	15.92
PQSUM$_{ABS}$	38.36	7.92	13.43	40.59	9.39	14.73	42.05	10.79	15.91
PQSUM$_{WSL\text{-}DS}$	40.36	**9.17**	14.74	43.22	**10.70**	16.35	44.61	12.40	17.66

Table 1: Performance comparisons in terms of (a) **F1** and (b) **Recall**. '*' denotes extractive model.

3.2 Iterative Fine-Tuning for Multi-Document Summarization

For the QF-MDS task, we adopt the transformer-based (Vaswani et al., 2017) BERTSUM model pre-trained for generic abstractive summarization on the CNN/DailyMail dataset (Liu and Lapata, 2019) to leverage the advantages of fine-tuning it for the query focused abstractive summarization task (Laskar et al., 2020a). However, BERTSUM was trained for the single-document summarization task by considering at most 512 tokens (Liu and Lapata, 2019; Beltagy et al., 2020; Zaheer et al., 2020). To address this issue for the multi-document scenario, we take an iterative approach (see Figure 1b). At first, we incorporate query relevance via concatenating the query with each document, similar to the work of (Laskar et al., 2020a). Then, we fine-tune BERTSUM using the weak abstractive reference summary to generate the query focused abstractive summary of each document in a document set. The sentences in the generated query focused summaries of each document set are then ranked using the fine-tuned RoBERTa$_{MS\text{-}MARCO}$ answer selection model to select the sentences that are most relevant to the query (see Figure 1c).

4 Experimental Setup

We now describe the datasets used in this paper, followed by the details of our implementation.

Datasets: We use the DUC 2005, 2006, and 2007 datasets for the QF-MDS task. The number of document sets were 50, 50, and 45 where the number of documents in each document set were 32, 25, and 25 in DUC 2005, 2006 and 2007 datasets respectively (Feigenblat et al., 2017). Each document set is associated with a query and the objective is to generate a summary containing at most 250 words from the document set based on the given query. Given the absence of the training data, to evaluate our model in each year's dataset we use the datasets from the other two years for training. From each year's training data, we randomly selected 20% of the document sets for validation while we used the rest for training.

Implementation: For the RoBERTa model, we used its Large version (Liu et al., 2019; Laskar et al., 2020b) and implemented using HuggingFace's Transformer (Wolf et al., 2019). For fine-tuning the summarization model, we used the BERTSUM$_{EXT\text{-}ABS}$[2] model pre-trained on the CNN/DailyMail dataset (Liu and Lapata, 2019). While selecting the most relevant sentences as the final query focused summary, we used the Trigram Blocking to reduce redundancy (Paulus et al., 2018). To fine-tune the BERTSUM model, we kept most parameters similar to the original work (Liu and Lapata, 2019) and ran 50 steps with batch size equal to 200. Among these 50 steps, we selected the step for evaluation that performed the best on the validation set. All of our models were run in multi-GPU settings using 4 NVIDIA V100 GPUs. We report the results based on both Recall and F1 scores in terms of ROUGE-1, ROUGE-2, and ROUGE-SU4 metrics (Lin, 2004). From now on, we denote ROUGE as **R**.

[2]https://github.com/nlpyang/PreSumm

Model	Recall	F1	Statistically Significant
PQSUM$_{\text{WSL-DS}}$	**42.73**	**42.84**	
without Distant Supervision	41.77 (- 2.25%)	41.88 (- 2.24%)	**No**, based on paired t-test ($p \leq .05$)
without Trigram Blocking	40.92 (- 4.24%)	41.01 (- 4.27%)	**No**, based on paired t-test ($p \leq .05$)
without Weakly Supervised Learning	40.01 (- 6.37%)	40.12 (- 6.35%)	**Yes**, based on paired t-test ($p \leq .05$)

Table 2: Ablation test result based on the average **R-1**. '-' denotes 'deterioration from original model'.

5 Results and Discussions

We now analyze the performance of our proposed model by comparing with other models (see Table 1). We also perform ablation test to further investigate its effectiveness (see Table 2). We denote our approach of using the **P**re-trained models (RoBERTa and BERTSUM) for **Q**uery focused **SUM**mary generation via utilizing Weakly Supervised Learning with Distant Supervision (WSL-DS) as **PQSUM$_{\text{WSL-DS}}$**. For performance comparisons, we use two baselines that do not utilize weak supervision and fine-tuning. Note that both of these baselines use the BERTSUM (Liu and Lapata, 2019) model pre-trained on the CNN/DailyMail dataset. One of them is pre-trained for extractive summarization: **PQSUM$_{\text{EXT}}$**; while the other is pre-trained for abstractive summarization: **PQSUM$_{\text{ABS}}$**. These baselines generate the summaries of all documents in a document set which are then ranked using RoBERTa$_{\text{MS-MARCO}}$. Moreover, we compare our model with four recent works: i) CES-50 (Feigenblat et al., 2017), ii) RSA (Baumel et al., 2018), iii) QUERYSUM (Xu and Lapata, 2020), and iv) DUAL-CES (Roitman et al., 2020).

Performance Comparisons: From Table 1(a), we find that our **PQSUM$_{\text{WSL-DS}}$** model sets a new state-of-the-art in all datasets based on the F1 metric for all ROUGE scores. Specifically, based on **R-1**, it improves by 5.88% in DUC 2005 from (Roitman et al., 2020) along with 4.54% and 3.28% from (Xu and Lapata, 2020) in DUC 2006 and 2007 respectively. From Table 1(b), we find that our model also sets a new state-of-the-art in terms of **R-2** Recall in DUC 2005 and 2006, but fails to outperform the DUAL-CES (Roitman et al., 2020) in DUC 2007. In comparison to the abstractive RSA model (Baumel et al., 2018), we find that our model outperforms them in all datasets in terms of both **R-1** and **R-2** Recall, but fails to outperform them in **R-SU4** scores. Moreover, we find based on paired t-test ($p \leq .05$) that the weakly supervised learning **significantly** outperforms the baselines in terms of both Recall and F1.

Ablation Test: The result of our ablation test based on the average of **R-1** scores across all datasets is shown in Table 2. We find that the performance is deteriorated if we exclude *Distant Supervision* via removing the RoBERTa$_{\text{MRPC}}$ model, as well as if the *Trigram Blocking* is not used. Moreover, the performance is **significantly** degraded if the summary is generated by only ranking the sentences in the documents using the fine-tuned RoBERTa$_{\text{MS-MARCO}}$ without utilizing *Weakly Supervised Learning*.

6 Conclusions and Future Work

In this paper, we propose a novel weakly supervised approach for the Query Focused Multi-Document Abstractive Summarization task to tackle the issue of no available labeled training data for such tasks. We also propose an iterative approach to address the computational problem that occurs while training neural models in long documents (Kitaev et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020). Experimental results in three datasets show that our proposed approach sets a new state-of-the-art result in various evaluation metrics. In the future, we will apply our models on more tasks, such as information retrieval applications (Huang and Hu, 2009; Huang et al., 2003; Yin et al., 2013; Huang et al., 2005), sentiment analysis (Liu et al., 2007; Yu et al., 2012), learning from imbalanced or unlabeled datasets (Liu et al., 2006; Bari et al., 2019; Bari et al., 2020), and automatic chart question answering (Kim et al., 2020).

Acknowledgements

We gratefully appreciate the area chair(s) and the reviewers for their excellent review comments. This research is supported by the Natural Sciences & Engineering Research Council (NSERC) of Canada, the York Research Chairs (YRC) program and an ORF-RE (Ontario Research Fund-Research Excellence) award in BRAIN Alliance. We acknowledge Compute Canada for providing us the computing resources.

References

James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J Harper, et al. 2003. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. In *ACM SIGIR Forum*, volume 37, pages 31–47.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2019. Zero-Resource Cross-Lingual Named Entity Recognition. *arXiv preprint arXiv:1911.09812*.

M Saiful Bari, Muhammad Tasnim Mohiuddin, and Shafiq Joty. 2020. Multimix: A robust data augmentation strategy for cross-lingual nlp. *arXiv preprint arXiv:2004.13240*.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 961–964.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2019. TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection. *arXiv preprint arXiv:1911.04118*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Xiangji Huang and Qinmin Hu. 2009. A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval,*, pages 307–314.

Xiangji Huang, Fuchun Peng, Dale Schuurmans, Nick Cercone, and Stephen E. Robertson. 2003. Applying machine learning to text segmentation for information retrieval. *Information Retrieval*, 6(3-4):333–362.

Xiangji Huang, Ming Zhong, and Luo Si. 2005. York University at TREC 2005: Genomics track. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC*.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Huang. 2019. Utilizing bidirectional encoder representations from transformers for answer selection task. In *Proceedings of the V AMMCS International Conference: Extended Abstract*, page 221.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Huang. 2020a. Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In *Canadian Conference on Artificial Intelligence*, pages 342–348. Springer.

Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. 2020b. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5505–5514.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3721–3731.

Yang Liu, Aijun An, and Xiangji Huang. 2006. Boosting prediction accuracy on imbalanced datasets with SVM ensembles. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD*, pages 107–118.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2007. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 607–614.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 443–452.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. 2016. An unsupervised multi-document summarization framework based on neural document model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1514–1523.

Jun Miao, Jimmy Xiangji Huang, and Zheng Ye. 2012. Proximity-based rocchio's model for pseudo relevance. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 535–544.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204.

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*, pages 1063–1072.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2273–2284.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Haggai Roitman, Guy Feigenblat, Doron Cohen, Odellia Boni, and David Konopnicki. 2020. Unsupervised dual-cascade learning with pseudo-feedback distillation for query-focused extractive summarization. In *Proceedings of The Web Conference 2020*, pages 2577–2584.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

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*, pages 1073–1083.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xiaojun Wan and Jianguo Xiao. 2009. Graph-based multi-modality learning for topic-focused multi-document summarization. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1586–1591.

Xiaojun Wan and Jianmin Zhang. 2014. Ctsum: extracting more certain summaries for news articles. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 787–796.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2008. Integrating clustering and multi-document summarization to improve document understanding. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 1435–1436.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *arXiv preprint arXiv:2004.03027*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Compressive document summarization via sparse optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1376–1382.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.

Xiaoshi Yin, Jimmy Xiangji Huang, Zhoujun Li, and Xiaofeng Zhou. 2013. A survival modeling approach to biomedical search result diversification using wikipedia. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1201–1212.

Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):720–734.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*.

Sheng-hua Zhong, Yan Liu, Bin Li, and Jing Long. 2015. Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications*, 42(21):8146–8155.