Automated Multi-document Text Summarization from Heterogeneous Data Sources

Mahsa Abazari Kia(✉)

School of Computer Science and Electronic Engineering, University of Essex, Essex, UK
ma19194@essex.ac.uk

We are currently witnessing an exponential increase of data which emanate from varied sources such as different types of records in companies, online social networks, videos, unstructured text in web pages, and others. It is very challenging to process this sparse, noisy and domain specific data. For instance, BT, a technology company in the UK and the PhD project sponsor, has a significant workforce of field engineers, desk-based agents, and customer support services who generate, collect and manage large volumes of temporally organized unstructured and semi-structured information every day. The problem that they face is effectively and efficiently answering client questions regarding order status live. The main challenge in my PhD is to propose computational models that could effectively and efficiently distil relevant information for the user who could answer the client's questions from various technical order record documents which are very noisy and follow no structural pattern. To this end, what would be useful is to automatically summarize and derive meaningful information in the form of short answers to queries from this vast source of distributed occurring data. The solution that my thesis proposes is based on a computational model that jointly learns extractive and abstractive summarization techniques with a temporal structure. Besides, the model interplays with a question-answering framework to help answer questions.

The information about orders at BT are in structured and free text format which is captured by and stored in different internal systems. These are being used by teams handling the orders. This volume of text about orders is an invaluable source of information which needs to be effectively and efficiently summarized in a way containing the information most relevant to:

- *'When' is the next action on the order.*
- *'What' is the current stage of the order.*
- *'Why' is it in this stage.*

It will help the desk agents to have a clear picture of the latest status of the order journey at the point in time t, instead of checking the order information from several places for the time that a customer calls in to find out about the progress of their order. In other words, we would like to generate summaries at the point in time t which can help humans comprehend the text content effectively and efficiently.

BT orders are structured in such a way that there is an update at regular intervals of time. These updates are required to be input to the summarizer as data over time. It means that summaries at time t are required to contain an update to the previous summary at

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 667–671, 2021.
https://doi.org/10.1007/978-3-030-72240-1_78

time $t - 1$. The summary S will contain the latest important information about an order including order progress information and answers for WH questions that are mentioned earlier. For this reason, we consider the setup mentioned above as a temporal question based multi-document summarization where input is a sequence $S = <d_1, ..., d_n>$ of time-stamped documents d_t covering some information for a specific order at BT.

In general, there are two different approaches for automatic text summarization: extractive and abstraction. Extractive summarization methods work by identifying important sections of the text. In contrast, abstractive summarization methods aim at producing important material in a new way.

The fundamental question surrounding my thesis revolves around proposing novel computational multi-document text summarization models that, 1) consider timestamps in noisy and sparse orders when generating summaries, 2) answer questions about the order progress including the three WH questions using an integrated abstractive and extractive approach. Since the BT orders information are written by different people with different roles, some noise and inconsistencies are introduced. Therefore, regenerating sentences with an abstractive technique would enhance the quality of the summary. Using a joint abstractive and extractive approach takes advantages of both approaches can produce high-quality summaries.

The summarization setting in our problem scenario is different from traditional summarization in such a way that it aims to capture only the latest relevant information about the orders so that desk agents could easily follow the latest order status. My other goal is to also generalize my novel approaches to other domains such as news, TREC datasets, scientific datasets, and others which largely fall under single document text summarization. This would help make my framework useful to others applying both single and multi-document summarization who wish to use other kinds of datasets. This is another fundamental challenge in my thesis.

Figure 1 shows the high-level model architecture for generating summaries and in the following sections I present details on these components.

1 Temporal Summarization

Three different approaches are proposed for generating temporal summaries.

Text Summarization Using Topic Modeling Over Time: Text summarization using topic modeling over time will generate the summary at time t regarding the evolution of topics and we could have a summary that contains the latest important sentences about important topics in order notes. So, applying topic modeling over time to text summarization, considering sentence diversity using semantic spaces such as sentence embeddings from BERT and maximum separation, as in maximum-margin setup, between topics could lead us to generate qualitative temporal summaries.

Unsupervised Temporal Abstractive Summarization: MeanSum [1] proposed an end-to-end, neural model architecture to perform unsupervised abstractive summarization, which could be used as a base method for our temporal summarization.

Adding temporal component, improving the sentence representation and making structural changes in MeanSum model which could process time-stamped documents.

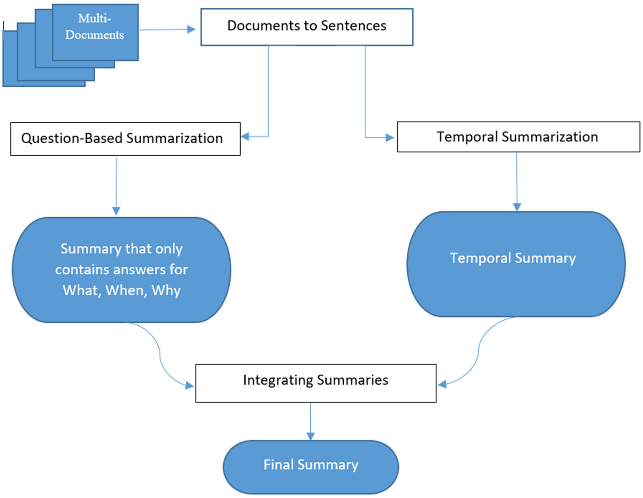

Fig. 1. A high-level architecture of my model.

Appling the updates to the summary (from Time t − 1 to t), would be ideal for our problem scenario due to the lack of training data.

Unsupervised Temporal Extractive Summarization: Tuning an unsupervised summarizer based on modern language models (which are recent trends in text summarization) for generating temporal summaries is another probable solution for this problem scenario. A BERT extractive summarizer is one of the unsupervised summarizers that could be utilized as a base for generating temporal extractive summaries.

2 Question-Based Summarization

Most of the techniques for answer extraction are supervised and due to the lack of training data in BT data, unsupervised approaches are more suitable. Giveme5W1H [2] is an open-source system that extracts phrases answering the journalistic 5W1H questions describing a news article's main event, i.e., who did what, when, where, why, and how? we need the sentences containing answers. Ranking and selecting the best answers (sentences) for When, What, and Why questions based on their date-time.

3 Merging Summaries

After obtaining temporal summary and question-based extractive summary our next goal is merging and aggregating these two summaries considering coherency, removing redundancy and temporal order. To this end, the main challenge is to find the correct position for placing question-based summary sentences considering their date-time. Another challenge lies in tuning the appropriate length of the summary using the development set.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) measures summary quality by counting overlapping units such as the n-gram, word sequences, and word pairs between the system summary and the reference summary. There are automatic evaluation methods ROUGE-N which is an n-gram recall between a candidate summary and a set of reference summaries [3]. ROUGE-N is computed as follows:

$$ROUGE - N = \frac{\sum\limits_{S \in REF} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in REF} \sum\limits_{gram_n \in S} Count(gram_n)}$$

Since no benchmark is available, we can use KL-Divergence (Kullback–Leibler) Optimization [4] for evaluating the results for BT data in two steps:

- Discovering topics and their distribution for main document and summary document and computing KL-Divergence between them
- Doing summarization and KL-Divergence optimization at the same time

There are substantial challenges to the problem mentioned above because documents related to each order are stored in different internal systems, i.e., they are distributed in nature and they were produced by different teams and engineers which incorporates a considerable amount of incoherence and different distributions of the vocabulary use. In other words, we will have a heterogeneous set of documents where the main topic for the documents is unrelated, but they contain some information that is related to the order progress. Furthermore, there is no labeled data in this use case. We must consider temporal information and summarize order documents in a way that it contains information relevant to three WH questions.

Given the recent progress in automated text summarization, extractive techniques are still attractive as they are less complex, less expensive, and generate grammatically and semantically correct summaries [5]. However, these methods suffer from the inherent drawbacks of discourse incoherence and long, redundant sentences [6]. On the other hand, due to the difficulty of natural language understanding and generation most previous research on document summarization is more focused on extractive methods and most recent abstractive work has focused on headline generation tasks. Regarding the challenges for this use case and drawbacks for both extractive and abstractive methods, I would like to combine abstractive and extractive techniques to generate temporal summaries that contain answers for WH questions. Proposing a hybrid multi-document text summarization technique with contextualized language representations could generate accurate and coherent summaries since it can overcome natural language understanding challenges. What makes my work more unique is that it is going to be used by the company to solve real-life applied problem towards the end of my PhD which potentially could generate economic impact and help its users.

References

1. Chu, E., Liu, P.J.: MeanSum: a neural model for unsupervised multi-document abstractive summarization. arXiv preprint arXiv:1810.05739 (2018)

2. Hamborg, F., Breitinger, C., Schubotz, M., Lachnit, S., Gipp, B.: Extraction of main event descriptors from news articles by answering the journalistic five W and one H questions. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 339–340 (2018)
3. Song, S., Huang, H., Ruan, T.: Abstractive text summarization using LSTM-CNN based deep learning. Multimed. Tools Appl. **78**(1), 857–875 (2018). https://doi.org/10.1007/s11042-018-5749-3
4. Hu, Z., Hong, L.J.: Kullback-Leibler divergence constrained distributionally robust optimization. Available at Optimization Online (2013)
5. Nallapati, R., Zhai, F., Zhou, B.: Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
6. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks, Vancouver, Canada, pp. 1073–1083 (2017)