

Attend to the beginning: A study on using bidirectional attention for extractive summarization

Ahmed Magooda^{1*}, Cezary Marcjan²

¹ University of Pittsburgh, Pittsburgh, PA, USA

² Microsoft Research FUSE lab, CCP, Bellevue, WA, USA
aem132@pitt.edu, cezarym@microsoft.com

Abstract

Forum discussion data differ in both structure and properties from generic form of textual data such as news. Henceforth, summarization techniques should, in turn, make use of such differences, and craft models that can benefit from the structural nature of discussion data. In this work, we propose attending to the beginning of a document, to improve the performance of extractive summarization models when applied to forum discussion data. Evaluations demonstrated that with the help of bidirectional attention mechanism, attending to the beginning of a document (initial comment/post) in a discussion thread, can introduce a consistent boost in ROUGE scores, as well as introducing a new State Of The Art (SOTA) ROUGE scores on the forum discussions dataset. Additionally, we explored whether this hypothesis is extendable to other generic forms of textual data. We make use of the tendency of introducing important information early in the text, by attending to the first few sentences in generic textual data. Evaluations demonstrated that attending to introductory sentences using bidirectional attention, improves the performance of extractive summarization models when even applied to more generic form of textual data.

Introduction

Recently, automatic text summarization models either extractive or abstractive witnessed fast performance strides due to the emergence of deep learning models specially seq2seq models. A large number of recent neural abstractive summarization models employ the encoder-decoder structure (See, Liu, and Manning 2017; Gehrmann, Deng, and Rush 2018; Paulus, Xiong, and Socher 2018) to convert the input sequence into a relatively shorter sequence. Most of the recent extractive models, on the other hand, employ only an encoder part to convert the input sequence into a fixed feature vector, followed by a classification part (Nallapati, Zhai, and Zhou 2017; Liu and Lapata 2019). Text summarization has been applied to different natural language domains; news, academic papers, emails, meeting notes, forum discussions, etc.. While some models can be transferable from one domain to the other, it might be more beneficial to craft ad-

ditional modifications in those models to account for differences between domains. Forum discussion data (Tarnpradab, Liu, and Hua 2017), for example, is different in both structure and properties when compared to generic textual data such as news. Discussion threads usually start with an initial post/comment (i.e. seeking knowledge, or help, etc..). The following comments tend to target the initial post/comment, providing additional information or opinions. With that said, a question arises. Can we enhance existing summarization models to benefit from such properties ?

Inspired by (Seo et al. 2017) we propose integrating bidirectional attention mechanism in extractive summarization models, to help to attend to early pieces of text (initial comment). The main objective is to benefit from the dependency between the initial comment and the following comments and try to distinguish between important, and irrelevant or superficial replies. Moreover, recent research by (Jung et al. 2019) showed that in some domains, humans tend to introduce relatively important information early at the beginning of textual articles. Unlike discussion threads, We explore the benefit of attending to the beginning in a more generic textual setting. Simply by integrating bidirectional attention mechanism and attending to the first few sentences in a document. We conducted some experiments to evaluate this hypothesis using a dataset of generic (non-discussion based) documents. Thus our contributions in this work are three-fold. First, we introduce integrating bidirectional attention mechanism into extractive summarization models, to help to attend to earlier pieces of text. Second, we achieved a new SOTA on the forum discussion dataset through the proposed attending to the beginning mechanism. Third, to further verify the transferability of our hypothesis (i.e. attending to the beginning), we perform evaluations to show that attending to earlier sentence in a more generic text, can also benefit summarization models. on different domains other than discussions.

Related Work

Automatic text summarization has seen increasing interest and improved performance due to the emergence of seq2seq models (Sutskever, Vinyals, and Le 2014) and attention mechanisms (Bahdanau, Cho, and Bengio 2014).

*This work was done during internship
Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

This is true for both automatically generating coherent summary (abstractive summarization), and extracting salient pieces of text (extractive summarization). The majority of recent research has been directed towards the news domain (See, Liu, and Manning 2017; Paulus, Xiong, and Socher 2018) (i.e. due to the existence of huge annotated datasets CNN/DailMail, Gigawords, Newyork Times). Unlike news, other domains such as (emails, discussions, meeting notes, students feedback, and opinions) can still be considered underexplored.

Recent efforts to tackle such domains started to emerge, (Luo, Liu, and Litman 2016) targeted student feedback summarization by extracting a set of representative phrases. (Li et al. 2019) proposed doing abstractive summarization for meeting notes by employing textual and visual information into a multi-model setting. (Li, Li, and Zong 2019; Yang et al. 2018) tackled the problem of opinion and review summarization. A work targeting similar domain as ours is done by (Tarnpradab, Liu, and Hua 2017). In which they proposed doing hierarchical attention to perform extractive summarization over a dataset of forum discussions collected from trip advisor. Another work which shares a similar design concept as ours was done by (Wang, Quan, and Wang 2019). They also integrated bidirectional attention mechanism in their model, however, their model and ours are different in both intuition and application. The major motive to integrate bidirectional attention in their design is to attend to an external template during the summarization process, while in ours we propose attending to early pieces of input text using the bidirectional attention mechanism. Another major difference is the intended application. They developed their model for the task of abstractive summarization over news data, while ours is intended for extractive summarization task.

Dataset

In this work, we employ two extractive summarization datasets. First, we used the discussion dataset proposed by (Tarnpradab, Liu, and Hua 2017)¹. The discussion dataset is extracted from trip advisor forum discussions. The data consists of 700 threads. In their work, (Tarnpradab, Liu, and Hua 2017) used 600 threads for training and 100 for validation. We didn't use the same data distribution reported by the authors, however, we kept the same testing data size for comparability reasons. We used our own split to verify the utility of our proposed techniques. We used 500 threads for training, 100 for validation, and 100 for testing. Moreover, we conducted additional experiments using (MSW) dataset². MSW dataset is a generic textual dataset that is not publicly available and is used to verify the transferability of our hypothesis to more generic textual domains. We verify whether we can benefit from documents' structure, and human's tendency to present important information earlier, by attending to early sentences. MSW dataset consists of a collection of 532 generic documents of different domains. We split the

¹Data can be downloaded from <https://www.dropbox.com/s/heevii01b1l6s0a/threadDataSet.zip?dl=0>

²This dataset is not publicly available

| Dataset | Part | # Documents | Total # Sentences |
|--------------|-------|-------------|-------------------|
| Trip Advisor | Train | 500 | 29671 |
| | Val | 100 | 6251 |
| | Test | 100 | 4280 |
| MSW | Train | 266 | 19748 |
| | Val | 138 | 11488 |
| | Test | 128 | 9898 |

Table 1: Model Datasets

data into training, validation, and testing of 266, 138, and 128 documents respectively. Table 1 summarizes the distribution of datasets used.

Baselines

In order to validate our hypothesis and show the utility of our proposed enhancements, we implemented 4 baselines. The following sections provide additional details regarding each of the baselines implemented.

Sumy³

We used the python off the shelf package for text summarization Sumy. Summarization is done hierarchically, where each comment from the discussion thread is passed separately to Sumy. The resultant summaries are then combined and passed to Sumy as one final document to get the thread summary.

LSA + clustering

We implemented a simple baseline for extractive summarization. The baseline uses Latent Semantic Analysis (LSA) to embed sentences into vector space. Sentences are then clustered using the K-means clustering algorithm. We use a number of clusters = \sqrt{n} where n is the number of sentences in the input document. Lastly, for each cluster, a cluster head is picked. The cluster head is the sentence closest to the mean point of the cluster.

SummaRuNNer

SummaRuNNer is an auto-regressive extractive summarization model proposed by (Nallapati, Zhai, and Zhou 2017)⁴.

SiATL

(SiATL) is sentence classification model developed by (Chronopoulou, Baziotis, and Potamianos 2019). The model employs multi-task learning by integrating language modeling auxiliary loss during the training process. SiATL model was developed originally as a sentence classification model. However, we decided to deal with extractive summarization as a pure sentence classification problem, and use SiATL model as extractive summarizer.

Unlike SummaRuNNer which is an auto-regressive model (i.e. previous decisions made by the model, affect its future decisions), SiATL performs classification independently for

³<https://pypi.org/project/sumy/>

⁴Refer to paper for original model design

each sentence. While it may always seem that autoregressive models would perform better, (Xiao and Carenini 2019) showed that non-auto-regressive models can sometimes be more efficient. Thus, we decided to use the SiATL model as a baseline, and compare the performance of auto-regressive and non-auto-regressive models within the extent of our study.

Attend to the beginning

Throughout this work, we hypothesize that attending to the initial part of a text during extractive summarization would help in selecting more salient sentences. The intuition is that in some situations (e.g. discussion threads), the initial part of a text holds important topical information. Henceforth it renders an important factor in selecting salient sentences for summarization objective. Thus we validate this hypothesis by calculating the importance of a sentence with respect to the initial part of the text, in the form of attention. Influenced by (Seo et al. 2017; Wang, Quan, and Wang 2019), the same interaction approach is employed here to produce beginning-aware sentence representations, for each sentence in the document. First, a sentence representation is produced for each sentence of the document as well as each sentence of the beginning part of the document (i.e. initial post/comment in the case of discussion dataset). Similarity matrix $S \in R^{m \times n}$ is then computed for each pair of document and beginning part sentences, $s_{ij} = W_0[h_i^d; h_j^b; h_i^d \otimes h_j^b]$. Where n is the number of sentences in the beginning part, m is the number of sentences in the input document, $;$ is the concatenation operator, h_i^d is the sentence representation of the i 's sentence in the document, and h_j^b is sentence representation of the j 's sentence in the beginning part. Each row and column of S is then normalized by softmax, which produces two new matrices \bar{S} and \bar{S}^T . Bidirectional attention is then calculated as $A = \bar{S} \cdot h^b$, $B = \bar{S}^T \cdot h^d$, where A represents document-to-beginning attention and B represents beginning-to-document attention. Finally, we obtain the beginning-aware sentence representations for each sentence in the document: $G_i^d \forall i \in m$, where $G_i^d = [h_i^d; A_i; h_i^d \otimes A_i; h_i^d \otimes B_i]$

The underlying mechanism to integrate bidirectional attention in (SiATL, and SummaRuNNer)⁵ is very much the same, except for the level of granularity in which attention operates on. SummaRuNNer operates on the level of document, so the bidirectional attention mechanism is calculated on the level of sentences between (all document sentences, and the beginning sentences) (i.e. h_i^d is the sentence representation of the i 's sentence in the document, and h_j^b is sentence representation of the j 's sentence in the beginning part). On the other hand, SiATL operates on the level of sentence, thus bidirectional attention is calculated between words of the input sentence, and words of the beginning part of the document (i.e. h_i^d is the word representation of the i 's word in the input sentence, and h_j^b is word representation of the j 's word in the beginning part).

⁵Code is available at github.com/amagooda/SummaRuNNer_coattention operate on forum discussion data, comments are split into

Additional Proposed Modifications

BERT Embedding

Recently released BERT(Devlin et al. 2018) embeddings showed the ability to outperform simply using shallow word embeddings. Moreover, It helped pushing the state of the art for numerous tasks within the NLP community. In this work, we integrate BERT embeddings within the SummaRuNNer model. Instead of initializing word embeddings randomly, BERT embeddings are used. The model embedding layer is initialized with Bert embedding and froze during the training phase.

Keyword Extraction

Attention mechanisms aim to weight tokens differently based on their importance. Another modification we introduce in this work is directed towards feeding the model with an extra signal. The extra signal, in this case, is keywords. The intuition behind feeding the model with keywords is pushing the model to give more attention to some specific words. The way we integrate keywords in SummaRuNNer model is by extracting keywords from each sentence S_i . Then separately encode the keywords into hidden states using BiLSTM ($h_j^{kwi} \forall j \in N_{kwi}$, where N_{kwi} is the number of extracted keywords from sentence S_i). The last hidden state is then used to represent all the keywords (h_i^{kw}). A new sentence embedding (h_i^{dkw}) is then formed by directly concatenating the original document aware sentence representation and the keywords representation.

$$h_i^{dkw} = [h_i^d; h_i^{kw}]$$

Experiments

To verify our hypotheses and validate the utility of our proposed modifications, we conducted a number of experiments. Our experimental designs address the following hypotheses:

Hypothesis 1 (H1) : Attending to the beginning of a discussion thread, would help extractive summarization models to select more salient sentences.

Hypothesis 2 (H2) : Non-auto-regressive models such as SiATL might be more suitable for thread discussion summarization, compared to auto-regressive models such as SummaRuNNer.

Hypothesis 3 (H3) : Adding additional features, such as contextual embeddings (e.g. BERT) and keywords can give summarization models a boost in performance.

Hypothesis 4 (H4) : Attend to the beginning is transferable to different forms of text other than discussion threads.

LSA + Kmeans. As part of the LSA baseline, two LSA vector spaces were used; First a vector space trained on a part of Wikipedia. Second, a vector space trained using the forum discussion dataset. ScikitLearn python package was used to produce LSA vector spaces of 200 dimensions.

SummaRuNNer. We implemented SummaRuNNer model following (Nallapati, Zhai, and Zhou 2017). To

sentences using Stanford sentence parser. All sentences are then concatenated into a single document.

$D = \{S_i \text{ for } S_i \text{ in } C_1\}; \dots; \{S_i \text{ for } S_i \text{ in } C_j\}$ for $j \in [1..n]$. Where S_i is the i 's sentence, ',' is the concatenation operator, C_j is the j 's comment, and n is the number of comments. Moreover, to operate on MSW dataset, each document is also split into sentences using Stanford sentence parser. SummaRuNNer used randomly initialized embeddings of size 64. The hidden state size of the LSTM is 128. Input sentences are truncated to 75 tokens, while shorter sentences are padded. The model is trained with batch size = 32 for 100 epoch. We calculate ROUGE score over the development set on each epoch. Later on, the checkpoint with maximum ROUGE is used for testing.

SiATL (H2). We used the implementation of SiATL released by the authors⁶. The model used embeddings of size 400 dimensions. The hidden state size of the shared LSTM is 1000. The task LSTM is of size 100. Input sentences are truncated to 80 tokens, while shorter sentences are padded. The model is trained with batch size = 32 for 100 epoch. Similarly, we calculate the ROUGE score over the development set. Later on, the checkpoint with the maximum ROUGE score is used for testing.

SummaRuNNer + Bidirectional Att. (H1, and H4). The bidirectional attention mechanism integrated in SummaRuNNer operates on the level of document. To conduct experiments on forum discussion data, the beginning part is the first comment which is the initial post in the thread. On the other hand, during experiments on the MSW dataset, the beginning part is the first N sentences in each document. In this work, we used $N = 3$.

SummaRuNNer + BERT Embedding (H3). To initialize SummaRuNNer with BERT word embeddings, BERT base uncased embeddings were used⁷. Each word is represented by the concatenation of BERT's last two layers, which leads to a word representation of size = $2 \times 768 = 1536$. We tried combining different number of layers (1, 2, and 3) for each word representation. We found that combining 2, or 3 layers performs better than using only the last layer. We decided to use only 2 layers to reduce the number of model parameters.

SummaRuNNer + Keyword extraction (H3). To extract keywords, we use Rapid Automatic Keyword Extraction (RAKE) (Rose et al. 2010) to identify keywords. For each sentence in the document, Keywords extracted and concatenated. Each pair of sentence and corresponding concatenated keywords are then passed to SummaRuNNer as separate inputs.

SiATL + Bidirectional Att. (H1, H2, and H4). Unlike SummaRuNNer, SiATL operates on the level of

individual sentences. Thus, the bidirectional attention mechanism integrated operates on the level of words. To conduct experiments on forum discussion data, the beginning part is all the words from the initial comment in the thread. On the other hand, during experiments on the MSW dataset, the beginning part is all the words from the first N sentences in each document. In this work, we used $N = 3$.

Results on forum dataset

Table 2 presents summarization performance results for the 2 non-neural extractive baselines, for the original and proposed variants of the two summarization models SummaRuNNer and SiATL, and finally for the highest score reported by (Tarnpradab, Liu, and Hua 2017). Following (Tarnpradab, Liu, and Hua 2017) and other recent work, performance is evaluated using ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) (Lin 2004) on F1.

| Summarization Model | R-1 | R-2 | R-L | |
|----------------------------|--------------|--------------|--------------|----|
| Baselines | | | | |
| Tarnpradab (Best) | 37.6 | 14.4 | 33.8 | 1 |
| Sumy | 38 | 15.06 | 21.95 | 2 |
| LSA + kmeans (Discussions) | 35.94 | 19.05 | 23.03 | 3 |
| LSA + kmeans (Wikipedia) | 35.4 | 18.49 | 22.57 | 4 |
| SummaRuNNer (Basic) | 36.97 | 15.84 | 24.5 | 5 |
| SiATL (Self Attention) | <u>45.15</u> | <u>26.12</u> | <u>43.3</u> | 6 |
| SummaRuNNer | | | | |
| + Bidir. Att. | 37.46 | 16.17 | 24.5 | 7 |
| + BERT | 38.48 | 16.88 | 25.63 | 8 |
| + Keywords (KWs) | 37.3 | 15.85 | 24.98 | 9 |
| + Bidir. Att. + KWs | 37.79 | 16.25 | 24.76 | 10 |
| + BERT + KWs | 37.97 | 16.75 | 25.85 | 11 |
| + BERT + Bidir. Att. | <u>39.36</u> | <u>17.71</u> | <u>26.78</u> | 12 |
| + BERT + Bidir. Att. + KWs | 38.43 | 16.74 | 25.65 | 13 |
| SiATL | | | | |
| <i>Bidir. Att.</i> | 46.5 | 28.53 | 44.65 | 14 |
| Self Att.+ Bidir. Att. | 46.32 | 28.69 | 44.41 | 15 |

Table 2: ROUGE results. *Italics* indicates outperforms all baselines. **Boldface** indicates best result over all models. Underlining indicates best result within model group

The motivation for using bidirectional attention mechanism is our hypothesis (**H1**). Table 2 supports this hypothesis. All ROUGE scores for SummaRuNNer and SiATL, that involves attending to the beginning by using bidirectional attention mechanism (rows 7, 10, 12, and 14), Outperform their corresponding counterpart, without using bidirectional attention (rows 5, 9, 8, and 6) respectively. Our second hypothesis (**H2**) is non-auto-regressive models might be more suitable than auto-regressive ones, for discussion summarization. Table 2 shows that using non-auto-regressive model (SiATL) indeed improve ROUGE scores compared to the auto-regressive model (SummaRuNNer). In rows 6 and 5, we see that SiATL improved R-1 scores from 36.97 to 45.15. Similarly, R-2 and R-L are also improved from 15.84 to 26.12 and from 24.5 to 43.3 respectively. Additionally, SiATL introduced a new SOTA, with a huge improvement in

⁶<https://github.com/alexandra-chron/siatl>

⁷https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

ROUGE scores compared to the previous work using hierarchical attention (rows 6, 14 and 1), in which R-1 improved by 23.6%. R-2 improved by 98.12%, and finally, R-L improved by 32.1%. We also see the same benefits of attending to the beginning for SiATL model: compared to using only the self-attention mechanism (i.e. original model), using only bidirectional attention or combining both attention mechanisms (self and bidirectional) boost ROUGE scores (rows 6, 14, and 15)

Our next hypothesis (**H3**) is that enriching models with additional features such as (Contextual embeddings, or keywords) would boost the performance. For these experiments, we only used SummaRuNNer model, since it still has a room for improvement to catch up with the SiATL model. Table 2 shows that our third hypothesis is a valid one, but not for all cases. It shows that while adding Contextual embeddings by itself, or adding keywords by itself helps the model. Combining contextual embeddings with keywords tends to harm the model. We can see that Adding keywords to both variants of SummaRuNNer (original, and with bidirectional attention) introduces a slight improvement over ROUGE scores (rows 5, 7 and 9, 10). Where R-1 improved from (36.97, and 37.46) to (37.3, and 37.79) respectively. R-2 improved from (15.84, and 16.17) to (15.85, and 16.25), and R-L improved from (24.5, and 24.5) to (24.98, and 24.76). Similarly, adding BERT contextual embedding introduces a good improvement over ROUGE scores for both variants of SummaRuNNer (rows 5, 7 and 8, 12). Where R-1 improved from (36.97, and 37.46) to (38.48, and 39.36) respectively. R-2 improved from (15.84, and 16.17) to (16.88, and 17.71), and R-L improved from (24.5, and 24.5) to (26.99, and 27.95). Surprisingly, adding both features (BERT, and keywords), tends to be harmful to the model (rows 8, 12 and 11, 13). Further analysis is still needed to reach a solid conclusion for such behavior.

Results on MSW dataset

Table 3 presents summarization performance results for the original and proposed variants of SummaRuNNer, for the best-performing variant of SiATL. The motivation behind conducting experiments on the MSW dataset is to validate **our last hypothesis (H4)**. We can see that table 3 clearly shows that our hypothesis is a valid one. It shows that attending to the beginning of a document helps selecting more salient sentences, not just for discussion threads, but even for generic textual documents. Similar to the results on the discussions dataset, we can see that attending to the beginning through a bidirectional attention mechanism boosts ROUGE scores (rows 1, and 2). Additionally, we can see that combining bidirectional attention with BERT embeddings further improves the performance (rows 1, and 4).

Discussion & Analysis

Unlike its promising performance on discussions dataset (table 2), SiATL performed poorly on MSW dataset (table 3). Surprisingly, it was only able to outperform the lead3 baseline. Through analyzing different criteria of the generated output for SummaRuNNer and SiATL, over the

| Model | R-1 | R-2 | R-L | |
|---------------------------------|-------|-------|-------|---|
| SummaRuNNer | 63.48 | 54.51 | 61.66 | 1 |
| + <i>Bidir. Att.</i> | 64.23 | 55.23 | 62.21 | 2 |
| + <i>BERT</i> | 65.81 | 57.99 | 63.9 | 3 |
| + <i>Bidir. Att. + BERT</i> | 66.12 | 58.56 | 64.48 | 4 |
| SiATL (Self Att. + Bidir. Att.) | 44.81 | 27.02 | 42.79 | 5 |

Table 3: ROUGE results over MSW dataset.

two used datasets. We observed that SiATL tends to generate longer summaries compared to SummaRuNNer, and this most likely due to its non-auto-regressive nature. SummaRuNNer, on the other hand, tends to generate shorter summaries. Table 4 shows the average and standard deviation of the number of sentences generated using SummaRuNNer and SiATL model, compared to the human annotation. It shows that for the forum discussion dataset, the expected summary length is ~ 14 sentences. For the same dataset, the SiATL model generates summaries of length ~ 16 sentences, while SummaRuNNer generates summaries of length ~ 8 sentences. This can justify the superior performance of SiATL compared to SummaRuNNer on the forum discussion dataset. On the other hand, we can notice that the expected summary length for the MSW dataset is ~ 8 sentences. For the same dataset, SummaRuNNer consistently generates shorter summaries compared to SiATL of lengths ~ 6.5 compared to 22 respectively. It is clear that the huge difference in the length of the summary between the human and SiATL generated is the reason SiATL underperforms on the MSW dataset. A potential solution for the SiATL model would be by adding a final post-processing step (e.g. clustering, redundancy reduction, etc.). The rule of the post-processing step would be slightly filtering the generated summary, and help to pick a number of sentences close to the average number humans select.

| Model | Forum Discussions | | MSW | |
|-------------|-------------------|------|-------|-------|
| | Avg | Std | Avg | Std |
| Human | 13.38 | 8.16 | 7.15 | 7.58 |
| SummaRuNNer | 8.2 | 3.52 | 6.4 | 3.6 |
| SiATL | 16 | 6.48 | 21.85 | 12.69 |

Table 4: Average and standard deviation of the number of sentences generated from each model, and the human selected sentences

Conclusion & Future work

We explored improving the performance of neural extractive summarizers when applied to discussion threads by attending to the beginning of the text (i.e. initial comment/post) through a bidirectional attention mechanism. We showed that attending to the beginning of the text, improved ROUGE scores of different models and different variants of these models. We also showed the applicability of using a recent sentence classification model (SiATL) for extractive summarization and introduced a new SOTA ROUGE score on the trip advisor forum discussion dataset. Additionally, we

showed that attending to the beginning of the text is not limited to datasets in the form of discussion threads. We showed that it is transferable to more generic forms of text, in which we can attend to the first N sentences of the text, similar to attending to the initial post/comment in discussion threads. Lastly, we showed that the utility of attending to the beginning is constant, regardless of the used model or dataset. Integrating bidirectional attention always introduces an improvement in ROUGE scores. Future plans include trying more generic datasets such as news, to further verify the utility of attending to the beginning. Further experimenting with the SiATL model on other datasets, as it showed promising results when used as extractive summarizer. We also plan to try extending the SiATL model with a post-processing step to enforce more control over the output length. We also plan to try different values for N , the number of sentences as initial part from generic documents.

References

- Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Chronopoulou, A.; Baziotis, C.; and Potamianos, A. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proc. of the 2019 Conference of NAACL-HLT, Volume 1 (Long and Short Papers)*, 2089–2095.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Gehrmann, S.; Deng, Y.; and Rush, A. 2018. Bottom-up abstractive summarization. In *Proc. of the 2018 Conference on EMNLP*, 4098–4109.
- Jung, T.; Kang, D.; Mentch, L.; and Hovy, E. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proc. of the 2019 Conference on EMNLP and the 9th IJCNLP*, 3315–3326.
- Li, M.; Zhang, L.; Ji, H.; and Radke, R. J. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proc. of the 57th Annual Meeting of ACL*, 2190–2196. Florence, Italy: ACL.
- Li, J.; Li, H.; and Zong, C. 2019. Towards personalized review summarization via user-aware sequence network. In *Proc. of the AAAI Conference on Artificial Intelligence*, volume 33, 6690–6697.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Liu, Y., and Lapata, M. 2019. Text summarization with pretrained encoders. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3721–3731.
- Luo, W.; Liu, F.; and Litman, D. 2016. An improved phrase-based approach to annotating and summarizing student course responses. In *Proc. of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 53–63.
- Nallapati, R.; Zhai, F.; and Zhou, B. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Paulus, R.; Xiong, C.; and Socher, R. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.
- Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory* 1–20.
- See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of the 55th Annual Meeting of the ACL (Volume 1: Long Papers)*, volume 1, 1073–1083.
- Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. In *Proc. of the 5th International Conference on Learning representations (ICLR)*.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.
- Tarnpradab, S.; Liu, F.; and Hua, K. A. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The Thirtieth International Flairs Conference*.
- Wang, K.; Quan, X.; and Wang, R. 2019. Biset: Bidirectional selective encoding with template for abstractive summarization. *arXiv preprint arXiv:1906.05012*.
- Xiao, W., and Carenini, G. 2019. Extractive summarization of long documents by combining global and local context. *arXiv preprint arXiv:1909.08089*.
- Yang, M.; Qu, Q.; Shen, Y.; Liu, Q.; Zhao, W.; and Zhu, J. 2018. Aspect and sentiment aware abstractive review summarization. In *Proc. of the 27th international conference on computational linguistics*, 1110–1120.