The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Attend to the Beginning: A Study on Bidirectional Attention for Extractive Summarization

Ahmed Magooda,[1*] **Cezary Marcjan**[2]
[1] Computer Science Department, University of Pittsburgh, Pitssburgh, PA, USA
[2] Microsoft Research FUSE lab, CCP, Bellevue, WA, USA
amagooda@cs.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 extractive and abstractive witnessed fast performance strides due to the emergence of seq2seq models. Most of the recent extractive models, employ an encoder to convert the input sequence into a fixed feature vector, followed by a classifier (Nallapati, Zhai, and Zhou 2017; Liu and Lapata 2019). Text summarization has been applied to different natural language domains; news, academic papers, forum discussions, etc.. While some models are transferable from one domain to the other, it might be more beneficial to craft additional 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. Inspired by (Seo et al. 2017) we propose integrating bidirectional attention 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 articles. Unlike discussion threads, We explore the benefit of attending to the beginning in a more generic textual setting. Simply by attending to the first few sentences in a document. We conducted some experiments to evaluate this hypothesis using a dataset of generic 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. 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.

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 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). 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. (Li, Li, and Zong 2019) tackled the problem of opinion and review summarization. A work targeting same domain as ours is done by (Tarnpradab, Liu, and Hua 2017). In which they proposed doing hierarchical attention to perform extractive summa-

[*] This work was done during internship

	Trip Advisor			MSW		
	Train	Val	Test	Train	Val	Test
Docs	500	100	100	266	138	128
Sentences	29671	6251	4280	19748	11488	9898

Table 1: Model Datasets

rization 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).

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. Moreover, we conducted additional experiments using Microsoft Word (MSW) dataset[2]. MSW dataset was 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. In which for each document, important sentences are selected by human annotators to represent a document summary. Table 1 summarizes the distribution of datasets used.

Baselines

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

LSA + clustering: This is a simple baseline that uses Latent Semantic Analysis (LSA) to embed sentences into vector space. Sentences are then clustered using the K-means. Number of clusters = $\sqrt{n}$ where n is the number of sentences in the input document. Lastly, the sentence closest to the cluster mean is picked.

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

SiATL: A sentence classification model developed by (Chronopoulou, Baziotis, and Potamianos 2019). The model integrates language modeling auxiliary loss during the training process. SiATL model was developed originally as a sentence classification model. Unlike SummaRuNNer which is an auto-regressive model (i.e. previous decisions made by the model, affect future decisions), SiATL performs classification independently for each sentence. Thus, we decided to compare the performance of auto-regressive and non-auto-regressive models within the extent of this 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. The underlying mechanism to integrate bidirectional attention in (SiATL, and SummaRuNNer)[3] 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). 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.

Additional Proposed Modifications

BERT Embedding: Recently released BERT(Devlin et al. 2018) embeddings outperforms simply using shallow word embeddings. In this work, we initializing the embedding layer with BERT embeddings and freeze it during training.

Keyword Extraction: Another modification we introduce in this work is feeding the model with an extra signal (keywords). New sentence embeddings are formed by concatenating the original document aware sentence representation and the keywords representation.

Experiments

To verify our hypotheses and validate the utility of our proposed modifications, we conducted a number of experiments. Our experimental designs address these 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: For the LSA baseline, a vector space of 200 dimensions trained on the forum discussion dataset

[1] www.dropbox.com/s/heevii01b1l6s0a/threadDataSet.zip?dl=0

[2] This dataset is not publicly available

[3] github.com/amagooda/SummaRuNNer_coattention

using ScikitLearn python package.

SummaRuNNer: We implemented SummaRuNNer with input embeddings of size 64. The hidden state size of the LSTM is 128. Input is truncated to 75 tokens.

SiATL (H2): We used the implementation of SiATL released by the authors. 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.

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 (initial post). On the other hand,for the MSW dataset, the beginning part is the first 3 sentences in each document.

SummaRuNNer + BERT Embedding (H3): To initialize SummaRuNNer with BERT word embeddings, BERT base uncased large embeddings were used. Each word is represented by the concatenation of BERT's last two layers, which leads to a word representation of size = 2x768.

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 are concatenated and then passed to SummaRuNNer as separate inputs.

SiATL + Bidirectional Att. (H1, H2, and H4): Unlike SummaRuNNer, SiATL operates on the level of sentences. Thus, the bidirectional attention mechanism operates on the level of words. For forum discussion data, the beginning part is all the words from the initial comment in the thread. On the other hand, for the MSW dataset, the beginning part is all the words from the first 3 sentences in each document.

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). Performance is evaluated using ROUGE(1, 2, L) (Lin 2004) on F1. 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, R-2 and R-L scores. Additionally, SiATL introduced a new SOTA, with a huge improvement in ROUGE scores compared to the previous work using hierarchical attention (rows 6, 14 and 1). We also see the same benefits of attending to the beginning for SiATL. Compared to using only the self-attention, using bidirectional attention 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 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, or 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). Similarly, adding BERT contextual embedding introduces an improvement over ROUGE scores for both variants of SummaRuNNer (rows 5, 7 and 8, 12). Surprisingly, adding both features (BERT, and keywords), tends to be harmful to the model (rows 8, 12 and 11, 13). Further analysis for this behavior is needed to reach a conclusion.

Summarization Model	**R-1**	**R-2**	**R-L**	
Tarnpradab (Best)	37.6	14.4	33.8	1
Sumy (pypi.org/project/sumy)	38	15.06	21.95	2
LSA + kmeans	35.94	19.05	23.03	3
SummaRuNNer (Basic)	36.97	15.84	24.5	5
SiATL (Self Att.)	45.15	26.12	43.3	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.	39.36	17.71	26.78	12
+ BERT + Bidir. Att. + KWs	38.43	16.74	25.65	13
SiATL				
Bidir. Att.	**46.5**	28.53	**44.65**	14
Self Att.+ Bidir. Att.	46.32	**28.69**	44.41	15

Table 2: ROUGE results. **Boldface** indicates best result.

Results on MSW dataset

Table 3 presents summarization performance results for Lead3 baseline, for 3 different recent extractive baselines, for the original and proposed variants of SummaRuNNer, for the best-performing variant of SiATL, and finally for the Oracle (i.e. human to human) performance. 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 6, and 7). Additionally, we can see that combining bidirectional attention with BERT embeddings further improves the performance to outperform the human-to-human performance (rows 6, 9, and 1).

Model	R-1	R-2	R-L	
Oracle3	65.01	59.44	64.03	1
Lead3	41.62	29.55	39.74	2
BertSum + Transformer	43.49	32.14	41.85	3
BertSum + Classifier	58.63	47.75	56.95	4
(Cheng and Lapata 2016)	60.21	49.81	58.62	5
SummaRuNNer	63.48	54.51	61.66	6
+ *Bidir. Att.*	*64.23*	*55.23*	*62.21*	7
+ *BERT*	*65.81*	*57.99*	*63.9*	8
+ Bidir. Att. + BERT	***66.12***	***58.56***	***64.48***	9
SiATL (Self Att. + Bidir. Att.)	44.81	27.02	42.79	10

Table 3: ROUGE results over MSW dataset.

	Human		SummaRuNNer		SiATL	
	Avg	STD	Avg	STD	Avg	STD
Forum	13.38	8.16	8.2	3.52	16	6.48
MSW	7.15	7.58	6.4	3.6	21.8	12.7

Table 4: Avg Summary size and STD for different models

Discussion & Analysis

Unlike its promising performance on discussions dataset (table 2), SiATL performed poorly on MSW dataset (table 3). Through analyzing different criteria of the generated output for SummaRuNNer and SiATL, over the two datasets. We observed that SiATL tends to generate longer summaries compared to SummaRuNNer, and this most likely due to its non-auto-regressive nature. 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, the expected summary length is $\sim$14, $\sim$16 for SiATL, and $\sim$8 for SummaRuNNer. 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 is $\sim$8, $\sim$6.5 for SummaRuNNer, and $\sim$22 for SiATL. It is clear that the huge difference in the length between the human and SiATL generated causes SiATL to underperform on MSW. A potential solution for the SiATL model would be by adding a final post-processing step to control the summary size.

Conclusion & Future work

We explored improving neural extractive summarizers when applied to discussion threads by attending to the beginning of the text (i.e. initial post) through bidirectional attention. We showed that attending to the beginning of the text, improved ROUGE scores for different models. We also used a sentence classification model (SiATL) for extractive summarization, and introduced a new SOTA ROUGE score on the forum discussion dataset. Additionally, we showed that attending to the beginning of the text is both helpful and not limited to discussion threads. We showed that it is transferable to more generic forms of text, in which we can attend to the text's first N sentences. Future plans include verifying the utility of attending to the beginning over more datasets and different values for N. 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.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cheng, J., and Lapata, M. 2016. Neural summarization by extracting sentences and words. In *Proc. of ACL*.

Chronopoulou, A.; Baziotis, C.; and Potamianos, A. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proc. of NAACL-HLT*.

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*.

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 ACL*.

Li, J.; Li, H.; and Zong, C. 2019. Towards personalized review summarization via user-aware sequence network. In *Proc. of AAAI*, 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 (EMNLP-IJCNLP)*.

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*.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

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 ACL*.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. In *Proc. of ICLR*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Tarnpradab, S.; Liu, F.; and Hua, K. A. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The 13th inter. Flairs Conf.*

Wang, K.; Quan, X.; and Wang, R. 2019. Biset: Bi-directional selective encoding with template for abstractive summarization. *arXiv preprint arXiv:1906.05012*.