Team Zoom @ AutoMin 2021: Cross-domain Pretraining for Automatic Minuting

Felix Schneider, Sebastian Stüker, Vijay Parthasarathy

Zoom Video Communications, Inc.

firstname.lastname@zoom.us

Abstract

This Paper describes Zoom's submission to the First Shared Task on Automatic Minuting at Interspeech 2021. We participated in Task A: generating abstractive summaries of meetings. For this task, we use a transformer-based summarization model which is first trained on data from a similar domain and then finetuned for domain transfer. In this configuration, our model does not yet produce usable summaries. We theorize that in the choice of pretraining corpus, the target side is more important than the source.

Index Terms: abstractive summarization, dialog summarization, automatic minuting

1. Introduction

Reasearchers from Zoom participated in the 2021 workshop on automatic minuting at Interspeech [1]. Specifically, we participated in the English side of Task A: Given a full meeting transcript, generate bulleted minutes for this meeting. This task is more challenging than the more typically studied news summarization task, for several reasons: First, the transcript is much longer compared a typical news article—the average transcript in the provided training data is over 8000 words long, whereas a news article from the CNN/DailyMail corpus [2] is only 680 words. The minutes are also much longer—on average 410 words in AutoMin vs. 48 in CNN/DailyMail. Secondly, news articles have a single topic and a predictable structure, with the important information often concentrated at the beginning of the article. Because of this structure, a simple lead sentence extractor often performs well for these articles [3]. On the other hand, meetings cover many different topics, not necessarily one after the other and information is very sparsely distributed. Finally, the style of minutes differs between annotators so the target for the task is not clearly defined.

2. Data

For our training, we decided to use the MediaSum corpus [4] for pretraining, because its source domain is dialogs. Specifically, the corpus consists of interviews from American television and radio broadcasts and usually single-sentence summaries. We hoped that this would be a closer match to the target domain than news summaries and would therefore make finetuning easier. It is also a large-scale corpus, of comparable size to CNN/DM (see the table below). However, the very short summaries and resulting high ratio of source to target length should prove to be a challenge for our model, which we discuss below. Using this corpus puts our submission in the unconstrained data condition. We used no other pretraining data.

For fine-tuning, we use the concatenation of the AMI [5], ICSI [6] and the provided AutoMin data. The suggested pre-extracted versions of AMI and ICSI from the AutoMin organizers did not suit our needs, as it is completely separated by speaker and the turn order between speakers is lost, so we made our own extraction.

Table 1: *The data that we used. CNN/Dailymail is shown for comparison, we did not use it*

Dataset	Samples	Average Source words	Average Target words
CNN/Dailymail	311k	680	48
MediaSum	463k	1180	18
AMI	142	4900	300
ICSI	75	11600	443
AutoMin	95	8600	410

In order to bring the other data closer to the final task, we applied a rudimentary anonymization to AMI, ICSI and MediaSum. We used the SpaCy[1] named entity recognizer and replaced all entities tagged as *GPE*, *LOC*, *ORG*, *PERSON* and *WORK_OF_ART* by pseudonyms similar to the AutoMin data.

3. Model

We use the SEAL model [7], which uses a hierarchical transformer-based encoder and decoder. For the encoder, we concatenate turns into snippets of at most 128 tokens. Each turn is encoded separately, then an attention pooling layer produces an encoding for the entire snippet. The decoder produces output in segments of 96 tokens. For self-attention, the segments are considered as one sequence, for encoder-decoder attention, a scorer rates each input snippet with respect to relevance to each output snippet. Encoder-decoder attention is performed on the top 8 snippets for each segment.

In order to train the scorer, an auxiliary loss is added which takes as labels the ROUGE-2 scores between each segment and snippet, which are calculated in advance. We chose the Mean Square Error criterion for this auxiliary loss.

To initialize the model, we used the parameters of a pretrained BART model [8]. As a result, our architecture follows that of BART-base.

3.1. Training details

For pretraining, we trained our model with the Adam optimizer [9] with $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$ and gradient clipping at 0.1. We increased the learning rate linearly to 0.0006 for 4000 warmup steps, then used inverse square root decay. We used an effective batch size of 3000 target tokens.

For fine-tuning, we selected the best checkpoint by perplexity on the pretrain data, which was after 7 epochs or about

[1] https://spacy.io/

Table 2: *Automatic scoring of our models.*

Model	MediaSum-test			AutoMin-dev			AutoMin-test		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Pretrain	18.99	4.72	17.02	7.76	0.51	5.19	—	—	—
Finetune	—	—	—	12.29	1.79	7.93	0.04	0.00	0.03
Finetune-force	—	—	—	22.64	4.28	13.02	—	—	—

20 000 updates. We used the same optimizer as before but started the learning rate at 0.0001 with no warmup and used an effective batch size of 20 meetings. This model started overfitting very quickly so the final checkpoint that we used for our submission was after 600 updates, which represents 40 epochs of the fine-tuning data.

Unfortunately, we did not have the time to perform a more thorough exploration of the hyperparameter space, as there are almost certainly better parameter sets.

4. Results

On the test part of MediaSum, our model achieved a ROUGE-2 score of 5.08. A cursory inspection shows that outputs were coherent and usually relevant to the source content. The anonymization makes it more difficult to judge factual correctness, but it seems the model did often confuse the relationships between the mentioned parties.

On the fine-tuning development data, we achieve a ROUGE-2 score of 1.79. The model still produces summaries of a similar length as in the pretraining. The limited fine-tuning is not enough to make any significant change to the genre of the output summary. Such a short summary must necessarily have poor recall with regard to information in the source. However, we also observe that the accuracy of the summaries is usually poor—the pseudonyms of persons and organizations are used interchangeably with no regard to their actual relation. As a result, the output would not be useful as a summary of the meeting, because it incorrectly represents who decided what or what organization someone belongs to.

Outputs on the AutoMin test set are similar to the dev set. The scores are extremely low, which may be due to preprocessing. For example, we removed the participant list, date/time and annotator attribution from the summaries and never generated these.

By forcing the output to a minimum length during decoding, we can score significantly better on automatic metrics, but the output is neither sensible nor fluent. We include this result only as a showcase of the shortcomings of automatic metrics.

We also analyzed the scores predicted by the scorer module and found that it assigns practically uniform scores to all input snippets, making the selection of the top 8 snippets essentially random. We believe that to be a consequence of the MSE loss and will try other options in the future.

5. Conclusions

We presented our system for the automatic minuting task. Although we did not produce usable minutes, we nevertheless leave the task with relevant findings, which we summarize as follows:

- Our training regimen was not able to make significant use of the available finetuning data, highlighting the need for further hyperparameter search or better finetuning methods
- Adapting the length and style of the summary in fine-tuning proved very challenging. Other approaches [10] pretrain on typical news summarization modified into a very artificial dialogue, but providing a closer match to the target summary length. We hypothesize that the target side of the pretraining corpus is more important than the source, whereas we chose MediaSum for its similarity in the source domain
- The style of anonymization made it very easy for the model to fantasize incorrect person names and more difficult to verify factual accuracy of the summaries

We would like to thank the organizers for hosting this shared task and providing the data and we look forward to the next installment of this workshop.

6. Samples

Source: AutoMin dev set meeting 1.
Reference:

```
[PROJECT6] meeting, 2020-07-13
Participants of the meeting:  [PERSON9],
[PERSON19], [PERSON6], [PERSON21],
[PERSON2]
It was discussed at the meeting:
```

- `[PERSON9] finished second article in the journal`
- `[PERSON19] still need to write introduction and morfology and some other small things`
- `mostly [PERSON19]´s part is ready and everyone can read it and comment on it`
- `in the book it´s better to say we and add the reference`
- `section about neural language models should be merged to the previous chapture or deleted`
- `[PERSON6] needs to write the probing and other unsupervised methods`
- `[PERSON2] explained [PERSON6] what he wanted to say in his comments`
- `the next meeting is planned on 21th July`
- `it´s necessary to start thinking about revieweres`

`The minutes was written by: [ANNOTATOR1]`

Pretrain:

`Commentator [PERSON1] says that when [PERSON9] and [PERSON6] were in high-tech fashion, they were not just the ones who needed to make the difference.`

Finetune:

`[PERSON6] and [PERSON9] have been working on a new book, [PERSON9], which is set to be released this week.`

Finetune-Force:

`[PERSON6] and [PERSON9] have been working on a new book, [ORGANIZATION1], which lists the most important elements of their work. They discuss how to create a user-friendly system, and how they are able to use the same technology as other products. [PATION1] is considering the book as well as the project, which will be developed in the next few weeks, and will be used as a model for using the next year's work on the project. The team will work with the team will use the team to create the same model for the team, and the team members will work together to work on a system, as the next meeting with the next generation of the team's own, and their own, as they will be working on the next step in the project and the next stage, as a team will have to work with [PGANIZERSON6], and the project will work on their own and the same way to make the project to work in the final work on how to make it more difficult to make their own to make a case for the project in the remote control of the next case for [PIZATION2] to the next month. [ORERSON6]. [P [P] will be tested with the project for the group will be made of the group, and whether to be tested for the first time, and what they would be done in the team should be done with the group should be made in the first meeting with [LOCATION1]. [ORIZATION3] will have been done.`

7. References

[1] T. Ghosal, O. Bojar, M. Singh, and A. Nedoluzhko, "Overview of the first shared task on automatic minuting (automin) at interspeech 2021," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–25. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-1

[2] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *Advances in neural information processing systems*, vol. 28, pp. 1693–1701, 2015.

[3] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *arXiv preprint arXiv:1808.08745*, 2018.

[4] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "Mediasum: A large-scale media interview dataset for dialogue summarization," *arXiv preprint arXiv:2103.06410*, 2021.

[5] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th international conference on methods and techniques in behavioral research*, vol. 88. Citeseer, 2005, p. 100.

[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[7] Y. Zhao, M. Saleh, and P. J. Liu, "Seal: Segment-wise extractive-abstractive long-form text summarization," *arXiv preprint arXiv:2006.10213*, 2020.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[10] C. Zhu, R. Xu, M. Zeng, and X. Huang, "A hierarchical network for abstractive meeting summarization with cross-domain pretraining," *arXiv preprint arXiv:2004.02016*, 2020.