Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 298–304
Marseille, 20-25 June 2022

Evaluating the Effects of Embedding with Speaker Identity Information in Dialogue Summarization

Yuji Naraki, Tetsuya Sakai, Yoshihiko Hayashi
Faculty of Science and Engineering, Waseda University
Tokyo, 169-8555 JAPAN
yuji.1277@akane.waseda.jp, tetsuyasakai@acm.org, yshk.hayashi@aoni.waseda.jp

Abstract

Automatic dialogue summarization is a task used to succinctly summarize a dialogue transcript while correctly linking the speakers and their speech, which distinguishes this task from a conventional document summarization. To address this issue and reduce the "who said what"-related errors in a summary, we propose embedding the speaker identity information in the input embedding into the dialogue transcript encoder. Unlike the speaker embedding proposed by Gu et al. (2020), our proposal takes into account the informativeness of position embedding. By experimentally comparing several embedding methods, we confirmed the ROUGE and human evaluation scores of the generated summaries were substantially increased by embedding speaker information at the less informative part of the fixed position embedding with sinusoidal functions.

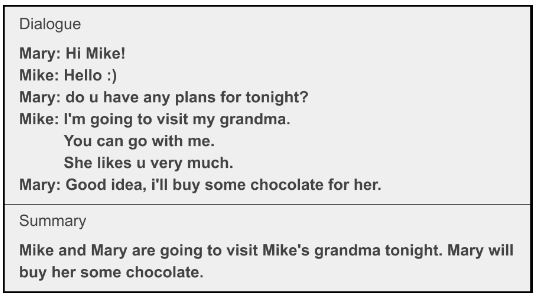
Keywords: summarization, dialogue, embedding

1. Introduction

There has been an increasing demand for automatic dialogue summarization in real-world applications, for example, in summarizing interactions in customer service centers and hospitals. An example of data for a dialogue summarization task is shown in Figure 1. However, for this task, there has been little research conducted on a dialogue summarization owing to a lack of high-quality datasets. Specifically, until SAMSum (Gliwa et al., 2019) has been available, there were few studies on dialogue-specific deep learning methods because there are no suitable or sufficiently large public datasets for the training of deep neural network (DNN) models. As the major difference between document and dialogue summarizations, the connection between speakers and their speech must be correctly captured by the model. Therefore, we focus on speaker identity information contained in dialogues and propose an embedding of speaker identity information as one of the input embeddings of Transformer-based models (Vaswani et al., 2017). More concretely, our proposed embedding is added only to the less informative parts of the position embedding. Experimental results demonstrate that the proposed method improves the convergence of the model in training and increases the average ROUGE scores of the generated summaries in comparison to existing methods of document and dialogue summarization.

2. Related Work

Abstractive Summarization. Liu and Lapata (2019) proposed BERTSumAbs, which apply BERT to an abstractive document summarization. However, the pre-training in BERTSumAbs is only applied to the encoder. Zhang et al. (2020) proposed PEGASUS, an abstractive summarization model that uses a pre-training approach called a gap sentence generation (GSG). GSG enables the decoder to be pre-trained, making it possible to specialize in sentence generation. We use PEGASUS as the base model of our proposed methods.



Dialogue

Mary: Hi Mike!
Mike: Hello :)
Mary: do u have any plans for tonight?
Mike: I'm going to visit my grandma.
You can go with me.
She likes u very much.
Mary: Good idea, i'll buy some chocolate for her.

Summary

Mike and Mary are going to visit Mike's grandma tonight. Mary will buy her some chocolate.

Figure 1: Example of dialogue and summary of SAMSum dataset (Gliwa et al., 2019) . A dialogue is a list of name-speech pairs, and a summary contains the names of the speakers in the dialogue.

Embeddings for Summarization Model. The input for Transformer-based models, such as BERTSumAbs and PEGASUS, is the sum of several types of embeddings generated from input sentences. Three types of embeddings are commonly used. Token embedding represents each token of the input sentences, segment embedding represents the two types of segments of the input sentences, and position embedding represents the positions of the input sentences. In addition to the conventional embeddings, Gu et al. (2020) add speaker embedding to the input so that they could improve the performance of the dialogue response selection task. Their speaker embedding has the same structure as segment embedding, and alternates two vectors at every speaker change. We improve Gu et al.'s speaker embedding to support dialogues with more than two speakers. In addition, we propose additive methods that take into account the informativeness of position embedding.

Datasets for Dialogue Summarization. Gliwa et al. (2019) released SAMSum to solve the problem of no publicly available dialogue summary datasets that have a sufficient number of high-quality data to train a DNN model. SAMSum includes various everyday conversations, including small talks and meeting arrangements, created by linguists. Zhu et al. (2021) released MediaSum, a large-scale media interview dataset consisting of interview transcripts with abstractive summaries. Unlike SAMSum, MediaSum is a much larger dataset that collects real interviews. We use both SAMSum and MediaSum in our experiments.

Recent Research on Dialogue Summarization. There are some research on dialogue summarization. Zhao et al. (2020) proposed TGDGA, which generates summaries from graph structure of input dialogues. Chen and Yang (2020) proposed a multi-view sequence-to-sequence model by first extracting conversational structure of dialogue from different views and generating summaries from the different views. Khalifa et al. (2021) experimentally demonstrated the effectiveness of several techniques for dialogue summarization tasks. Although there has been an increase in research on dialogue summarization, the present research is the first attempt to focus on speaker information in the dialogue summarization task. Although there has been an increase in research on dialogue summarization, To the best of our knowledge, ours is the first attempt to focus on speaker information in the dialogue summarization task.

3. Our Proposed Methods

3.1. Embedding of Speaker Identity Information

In order to train summarization models specific to the dialogue domain, the proposed method effectively feeds the models with speaker information. PEGASUS used in our experiment is a Transformer-based model, whose input is the sum of token embedding and position embedding obtained from the input sentences. In our proposed method, we use these embeddings plus an additional embedding containing speaker identity information, called *speaker embedding (SE)*, as shown in Figure 2. As the mechanism of this approach, IDs are assigned to the speakers of a dialogue, and the vectors corresponding to the IDs are given to the tokens in between a speaker name and the end of his or her speech (see Figure 1). Our proposed *SE* represents the speaker identity when dealing with dialogues between three or more people, not just two. This solves the problem of Gu et al.'s speaker embeddings that only represent the turns of the speakers.

3.2. Additive Methods Based on Position Embedding

As shown in Figure 2, *SE* is added to the input embedding, and we have devised an additive method to maximize the effect of *SE*. Before introducing the additive method, we describe the position embedding (PE) used in PEGASUS, called sinusoidal positional embedding. Sinusoidal positional embedding (Vaswani et al., 2017) is a type of PE with fixed parameters used in a Transformer, PEGASUS, etc. This is expressed by the following equations in our model. In addition, *pos* represents the position in the sequence, *i* represents the dimension, and *dim* represents the number of dimensions of the embedding input into the model.

$$PE_{(pos,i)} = \sin\left(pos/10000^{2i/dim}\right) \quad (1)$$

$$PE_{(pos,i+dim/2)} = \cos\left(pos/10000^{2i/dim}\right) \quad (2)$$

Figure 3 presents a heatmap of the parameters of the sinusoidal positional embedding used in our experiments. The vertical axis represents the position within the input sequence, and the horizontal axis represents the dimensions. We can see that the amount of information varies depending on the embedding dimension. Because the latter half of the dimension is less informative than the former half, as our hypothesis indicates, adding *SE* only to the latter half of the dimension would have a greater effect. To confirm our hypothesis, we compared a whole dimension addition method and several partial dimension addition methods of *SE*, such as adding it to each half or quarter of the dimension. Because *SE* only indicates the speaker identity information, reducing the number of dimensions does not prevent the embedding from being less expressive.

Figure 4 shows that, whereas the number of input dimensions is 1024, our partial embeddings cover one quarter ($128 + 128 = 256$) of this. The types of SEs are referred to by their parts where they are added. For example, we refer to *SE* added to the whole, and only to the latter half and fourth quarter of the dimension, as *whole SE*, *latter half SE*, and *fourth quarter SE*, respectively.

4. Dataset and Experimental Setup

4.1. Dataset

We used the SAMSum dataset and a part of the MediaSum dataset to train our models. In our experimental setup, the maximum number of input tokens was 512. Because approximately 5% of all dialogue data in MediaSum is made up of less than 512 tokens, the data used in our experiment are filtered under two conditions: the number of tokens of the dialogue text is 512 or less, and each dialogue involves two or more people [1].

We inserted two special tokens to clarify the borders between speaker names and speech as a preprocessing of the dialogue text. The [SEP] token represents

[1] There are some data with one speaker in MediaSum.

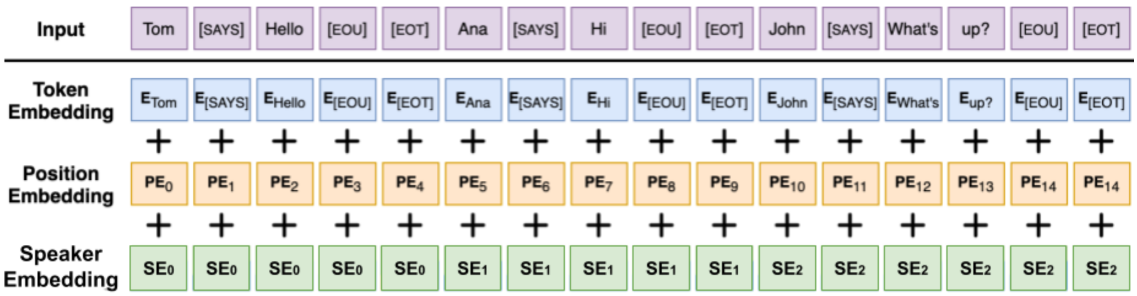

Input
Tom [SAYS] Hello [EOU] [EOT] Ana [SAYS] Hi [EOU] [EOT] John [SAYS] What's up? [EOU] [EOT]
Token Embedding
E_{Tom} $E_{[SAYS]}$ E_{Hello} $E_{[EOU]}$ $E_{[EOT]}$ E_{Ana} $E_{[SAYS]}$ E_{Hi} $E_{[EOU]}$ $E_{[EOT]}$ E_{John} $E_{[SAYS]}$ $E_{What's}$ $E_{up?}$ $E_{[EOU]}$ $E_{[EOT]}$
Position Embedding
PE_0 PE_1 PE_2 PE_3 PE_4 PE_5 PE_6 PE_7 PE_8 PE_9 PE_{10} PE_{11} PE_{12} PE_{13} PE_{14} PE_{14}
Speaker Embedding
SE_0 SE_0 SE_0 SE_0 SE_0 SE_1 SE_1 SE_1 SE_1 SE_1 SE_2 SE_2 SE_2 SE_2 SE_2 SE_2

Figure 2: Architecture of input representation at a dialogue sequence level. Input representation is composed of the union of two traditional embedding (token embedding and position embedding) and a speaker embedding.

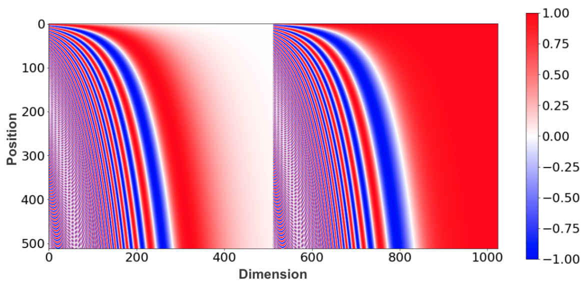


Figure 3: Heatmap of parameters of sinusoidal positional embedding.

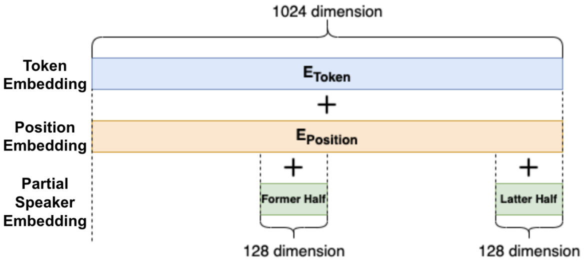
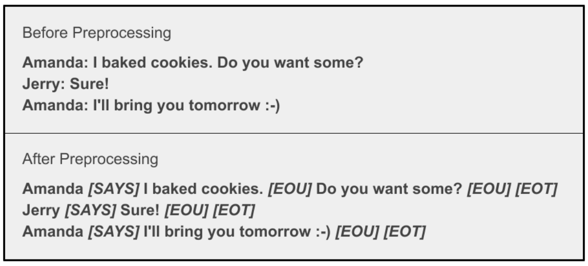


Figure 4: Architecture of *fourth quarter Speaker Embedding* at a token level. Partial speaker embedding is limitied to fractions of the dimension of the input representation.

speaker changes, and the [SAYS] token is placed between speaker names and speech. An example of a dialogue text before and after this preprocessing is shown in Figure 5.

4.2. Training details

We used PyTorch (Paszke et al., 2019) and HuggingFace transformers (Wolf et al., 2020). As the base model, we used PEGASUS, an encoder-decoder model of the document summarization, and employ the weights pre-trained by the XSum dataset (Narayan et al., 2018) as its initial parameters. SAMSum or MediaSum was used as the dataset for fine-tuning a model in the corresponding evaluation. After tuning the hyperparameters, we decided on a multiplier of 10 for our proposed additional embedding.


Before Preprocessing

Amanda: I baked cookies. Do you want some?
Jerry: Sure!
Amanda: I'll bring you tomorrow :-)

After Preprocessing

Amanda *[SAYS]* I baked cookies. *[EOU]* Do you want some? *[EOU] [EOT]*
Jerry *[SAYS]* Sure! *[EOU] [EOT]*
Amanda *[SAYS]* I'll bring you tomorrow :-) *[EOU] [EOT]*

Figure 5: Example of dialogue text before and after preprocessing.

5. Results and Discussions

5.1. Model Convergence

We analyzed the convergence of the model using the validation loss. Figure 6 shows the validation loss changes when fine-tuning using SAMSum. It can be seen that, whereas PEGASUS with the *former half*, *first quarter*, and *whole SE* failed to reduce the loss to the same level as pure PEGASUS, PEGASUS with *latter half*, and *fourth quarter SE* successfully reduced the loss to the same level as pure PEGASUS *and* converged more quickly. A similar result was observed when fine-tuning with MediaSum. These results show that providing speaker identity information to the model has a positive effect, whereas adding additional embedding into the informative part of the PE, particularly in the first quarter of the dimension, has a negative effect of not being sufficiently trained.

5.2. ROUGE Scores

We evaluated the generated summaries using ROUGE (Lin, 2004). Table 1 shows the average scores of ROUGE-1, ROUGE-2, and ROUGE-L on the test data for the baselines and our proposed methods. We note that the scores from the three baseline methods (i.e., Longest-3 [2], Transformer, and TGDGA) are quoted values from previous studies. It can be observed that, whereas PEGASUS with *whole* and *former half SE* underperform pure PEGASUS, PEGASUS with *latter*

[2] This model is commonly used in news summarization tasks, which treats the three longest utterances in order of length as a summary.

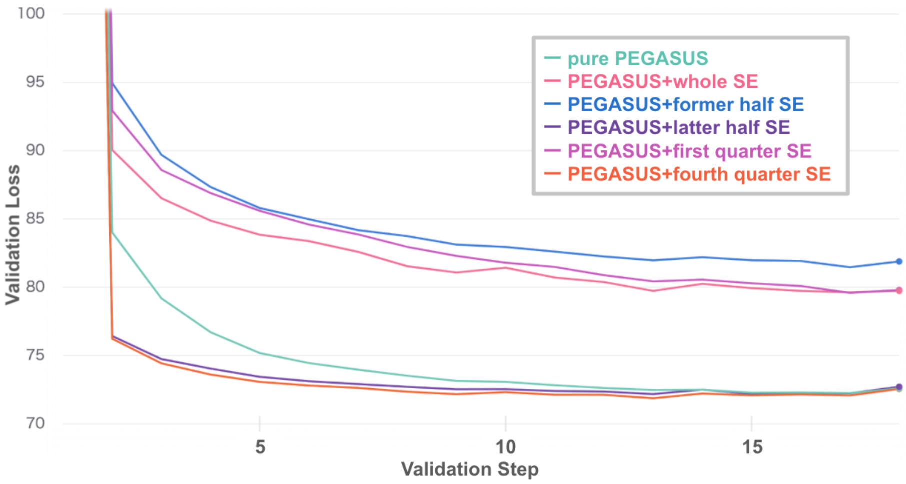


Figure 6: Validation loss transitions.

Method	SAMSum ($n = 819$)			filtered MediaSum ($n = 456$)		
	R-1	R-2	R-L	R-1	R-2	R-L
Longest-3	32.46	10.27	29.92	-	-	-
Transformer (Vaswani et al., 2017)	36.62	11.18	33.06	-	-	-
TGDGA (Zhao et al., 2020)	43.11	19.15	40.49	-	-	-
PEGASUS	50.82	26.62	42.65	38.65	20.54	34.79
PEGASUS + *whole SE*	46.40	21.92	38.52	34.84	15.77	30.53
PEGASUS + *former half SE*	44.4	19.48	36.20	33.62	15.24	29.80
PEGASUS + *latter half SE*	51.39	27.03	43.39	37.80	20.76	34.39
PEGASUS + *first quarter SE*	46.19	20.99	38.04	34.36	15.15	30.11
PEGASUS + *second quarter SE*	48.56	24.81	40.80	38.29	20.65	34.21
PEGASUS + *third quarter SE*	51.17	27.37	43.51	38.85	21.01	35.11
PEGASUS + *fourth quarter SE*	**51.39**	**27.58**	**43.61**	**40.05**	**21.90**	**36.20**

Table 1: ROUGE score of test data. n represents the number of test data.

half and *fourth quarter SE* outperform pure PEGASUS.

5.3. Example of Generated Summaries

Figure 7 shows two example dialogues from the SAMSum test data and corresponding reference and summaries by pure PEGASUS and PEGASUS + *fourth quarter SE*. In Figure 7-(a), the summary of our proposed method is almost a synonymous sentence with the reference. On the other hand, in Figure 7-(b), the summary of our method is incorrect because the speakers will not be discussing IMF lecture tomorrow evening.

6. Human Evaluation

ROUGE is an evaluation index frequently used in text summarization. It only focuses on the overlaps of words or N-grams in the reference and a generated summary; it does not consider their semantic matching.

Therefore, we conducted a human preference evaluation for pure PEGASUS and PEGASUS with *fourth quarter SE* on the SAMSum dataset.

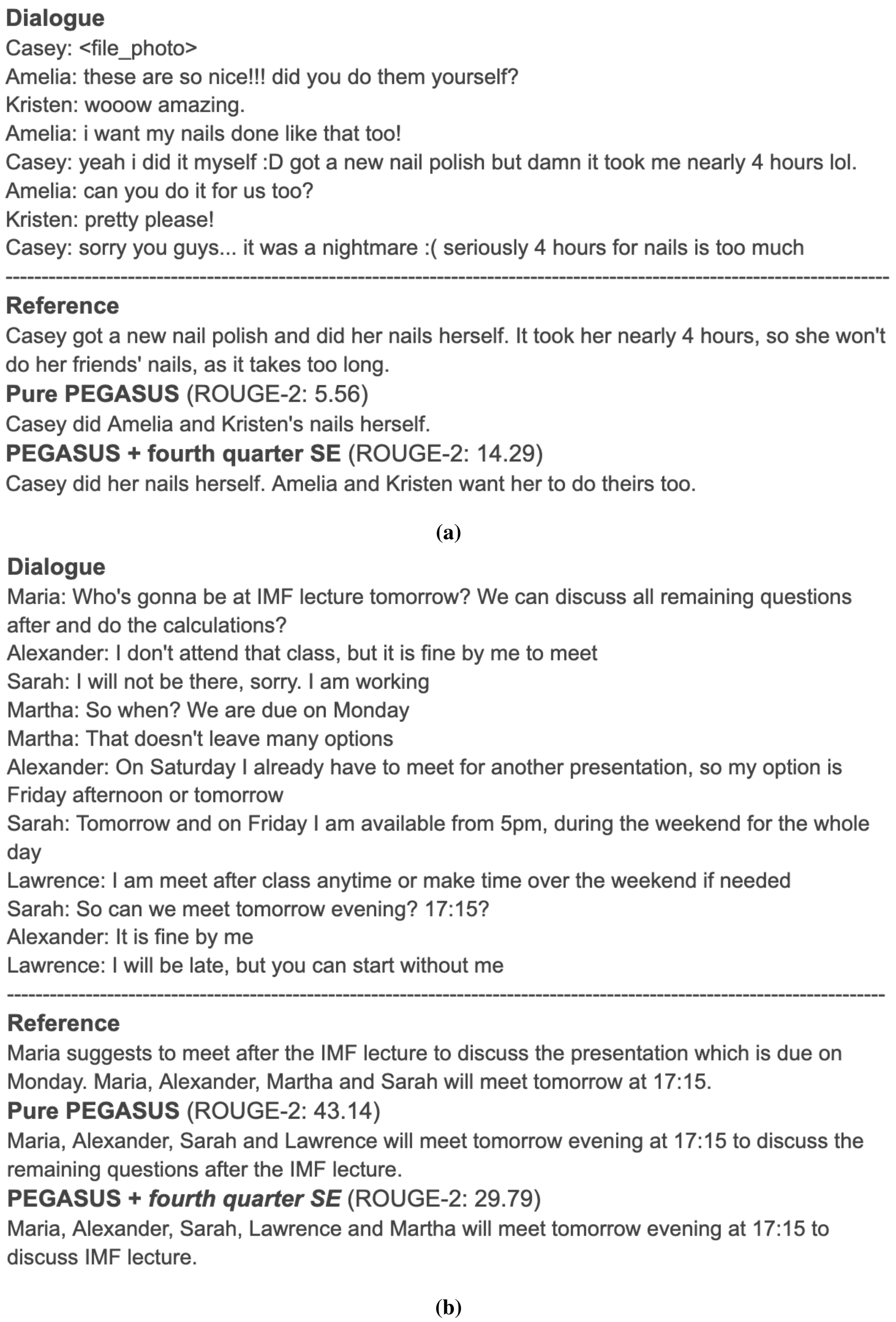
6.1. Setting for Human Evaluation

We use Amazon Mechanical Turk for the human preference evaluation. This section describes the specific settings for crowdsourcing.

First, We conducted selection of assessors to prevent poor-quality evaluations. We implemented three measures: First, the assessors must have scored well on the prepared pre-task; second, the user interface had to prevent them from answering too quickly; and third, the assessors had to provide quality-assured responses up to that point.

For the data used in the evaluation, we randomly selected 100 dialogues from the test data under two conditions: the number of input tokens is 512 or less, and the Jaccard coefficient of the word sets of the summaries generated by both methods is 0.8 or less. These conditions ensure that the model sees the entire dialogue and that the selected data are sufficiently different for a preference evaluation.

We conducted crowdsourcing each summary pair to be evaluated by 5 assessors. Thus, we obtained 500 human preferences (pure PEGASUS $>$ PEGASUS + *fourth quarter SE* or pure PEGASUS $<$ PEGASUS +



Dialogue
Casey: <file_photo>
Amelia: these are so nice!!! did you do them yourself?
Kristen: wooow amazing.
Amelia: i want my nails done like that too!
Casey: yeah i did it myself :D got a new nail polish but damn it took me nearly 4 hours lol.
Amelia: can you do it for us too?
Kristen: pretty please!
Casey: sorry you guys... it was a nightmare :(seriously 4 hours for nails is too much

----

Reference
Casey got a new nail polish and did her nails herself. It took her nearly 4 hours, so she won't do her friends' nails, as it takes too long.
Pure PEGASUS (ROUGE-2: 5.56)
Casey did Amelia and Kristen's nails herself.
PEGASUS + fourth quarter SE (ROUGE-2: 14.29)
Casey did her nails herself. Amelia and Kristen want her to do theirs too.

(a)

Dialogue
Maria: Who's gonna be at IMF lecture tomorrow? We can discuss all remaining questions after and do the calculations?
Alexander: I don't attend that class, but it is fine by me to meet
Sarah: I will not be there, sorry. I am working
Martha: So when? We are due on Monday
Martha: That doesn't leave many options
Alexander: On Saturday I already have to meet for another presentation, so my option is Friday afternoon or tomorrow
Sarah: Tomorrow and on Friday I am available from 5pm, during the weekend for the whole day
Lawrence: I am meet after class anytime or make time over the weekend if needed
Sarah: So can we meet tomorrow evening? 17:15?
Alexander: It is fine by me
Lawrence: I will be late, but you can start without me

Reference
Maria suggests to meet after the IMF lecture to discuss the presentation which is due on Monday. Maria, Alexander, Martha and Sarah will meet tomorrow at 17:15.
Pure PEGASUS (ROUGE-2: 43.14)
Maria, Alexander, Sarah and Lawrence will meet tomorrow evening at 17:15 to discuss the remaining questions after the IMF lecture.
PEGASUS + *fourth quarter SE* (ROUGE-2: 29.79)
Maria, Alexander, Sarah, Lawrence and Martha will meet tomorrow evening at 17:15 to discuss IMF lecture.

(b)

Figure 7: Two example sets of dialogues, references, and two summaries by pure PEGASUS and PEGASUS + *fourth quarter SE*.

fourth quarter SE). For each summary pair, the winner is determined based on a majority vote. In addition, the assessors were asked to choose the reasons for their decisions from among four options ("Not a natural sentence," "Different from the facts stated in the dialogue," "The non-essential points," and "Other reason"). Prior to the evaluation, the assessors were informed that the goal of the dialogue summarization was to take an objective view of the dialogue and condense a piece of text into a shorter version that covers the main points succinctly. Note that the two summaries and three choices other than "Other reasons" were arranged randomly to avoid a position bias.

6.2. Result of Human Evaluation

Table 2 shows the results of a human preference evaluation. "Wins" illustrates the results of the majority votes. It can be observed that PEGASUS with

Method	Wins
PEGASUS	37
PEGASUS + *fourth quarter SE*	**57**
Even	6

Table 2: Results of human preference evaluation.

the *fourth quarter SE* outperformed pure PEGASUS in scores of "Wins". These results demonstrate that the *fourth quarter SE* method, in particular, is effective in not only improving the automatic score ROUGE, but also producing more human preferable summaries. However, the detailed analysis of inter-annotator agreement presented in the Appendix shows that the overall differences in the human quality assessment were not significantly large.

7. Conclusion

We proposed *speaker embedding*s to indicate speaker identity information in dialogue summarization. Among the implemented embedding methods, *latter half*, *third quarter*, *fourth quarter SE* given to less informative parts improved the convergence and increased ROUGE scores, whereas *Whole former half*, *first quarter*, *second quarter SE* given to highly informative parts of the position embedding had negative effects. Furthermore, the results of a human preference evaluation suggested that summaries generated by PEGASUS with *fourth quarter SE* are better than the summaries generated by pure PEGASUS from a human preference perspective.

For future work, we analyze the types of errors in the generated summaries, namely "who said what"-related errors, and discuss the effect of our proposed method in more detail. We further plan to verify the effectiveness of our proposed methods in other tasks in the dialogue domain. For example, our speaker embeddings can be employed to enhance a dialogue history that is vital for a chatbot to manage multi-party dialogues.

8. Bibliographical References

Chen, J. and Yang, D. (2020). Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online, November. Association for Computational Linguistics.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November. Association for Computational Linguistics.

Gu, J.-C., Li, T., Liu, Q., Ling, Z.-H., Su, Z., Wei, S., and Zhu, X. (2020). Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*.

Khalifa, M., Ballesteros, M., and McKeown, K. (2021). A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (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 (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China, November. Association for Computational Linguistics.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November. Association for Computational Linguistics.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*.

Zhao, L., Xu, W., and Guo, J. (2020). Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

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

Appendix: Inter-Annotator Agreement between Assessors of Human Evaluation

This section discusses the degree of inter-annotator agreement in our human evaluation. As the human evaluation was conducted by crowdsourcing employing a group of unspecified workers, metrics generally used in assessing inter-annotator agreement are not applicable. We thus use the absolute difference in the number of annotators who preferred the summary generated by one method over another as the score to approximate the degree of agreement. A higher value for this score indicates a higher degree of agreement. Figure 8 shows a bar chart that represents the distribution of the scores over the test data, showing that the overall differences in the human quality assessment were not significantly large as we expected.

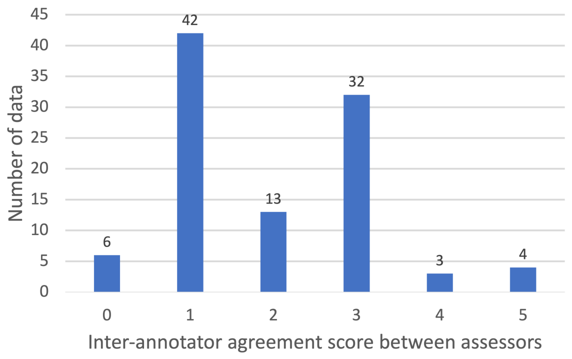


Figure 8: Inter-annotator agreement between assessors.