A Bag of Tricks for Dialogue Summarization

Muhammad Khalifa[2*], **Miguel Ballesteros**[1], **Kathleen McKeown**[1,3]
[1]Amazon AI, USA
[2]Cairo University, Cairo, Egypt
[3]Department of Computer Science, Columbia University, NY, USA
muhammad.e.khalifa@gmail.com
{ballemig, mckeownk}@amazon.com

Abstract

Dialogue summarization comes with its own peculiar challenges as opposed to news or scientific articles summarization. In this work, we explore four different challenges of the task: handling and differentiating parts of the dialogue belonging to multiple speakers, negation understanding, reasoning about the situation, and informal language understanding. Using a pretrained sequence-to-sequence language model, we explore speaker name substitution, negation scope highlighting, multi-task learning with relevant tasks, and pretraining on in-domain data. Our experiments show that our proposed techniques indeed improve summarization performance, outperforming strong baselines.

1 Introduction

The nature of dialogue poses additional challenges to summarizers beyond what is required when processing structured, single-speaker documents (Zhu and Penn, 2006). Given that dialogues typically represent an interaction between many speakers, a summarizer model must keep track of the different lines of thoughts of individual speakers, distinguish salient from non-salient utterances, and finally produce a coherent, monologue summary of the dialogue.

Dialogues usually include unfinished sentences where speakers were interrupted or repetitions, where a speaker expresses their thoughts more than once and possibly in different styles. Moreover, a single dialogue could touch on many topics without a clear boundary between the different topics. All the aforementioned phenomena certainly add to the difficulty of the task (Zechner and Waibel, 2000; Zechner, 2002; Chen and Yang, 2020).

Our work focuses on SAMSum (Gliwa et al., 2019), which is a dialogue summarization dataset comprised of ~16K everyday dialogues with their human-written summaries. As our backbone model, we use BART (Lewis et al., 2020), a state-of-the-art pretrained encoder-decoder language model that is suitable for sequence-to-sequence tasks. Table 1 shows an example of a summary generated using BART (Lewis et al., 2020), fine-tuned on SAMSum. Clearly, a level of reasoning is required to make sense of the conversation, which BART fails to do and therefore produces an incorrect summary.

We propose a combination of techniques to tackle a set of dialogue summarization challenges. The first challenge is having ***multiple speakers*** (generally, more than 2), where it becomes harder for the model to keep track of different utterances and determine their saliency. The second challenge is ***multiple negations***, which is thought by Chen and Yang (2020) to pose some difficulty to dialogue understanding. The third of these challenges is ***reasoning***, where the model is required to reason about the dialogue context, and infer information that is not explicitly expressed. The last challenge is ***informal language***. Since we focus on random, everyday conversations, these are usually filled with non-standard language (abbreviations, social media terms, etc.).

The contributions in this work are:

- We propose a set of novel techniques to address four dialogue summarization challenges: multiple speakers, negation, reasoning and informal language. Our techniques include name substitution, negation scope highlighting, multi-task learning with relevant tasks, and pretraining on in-domain corpora.

- We show impressive improvements on the summarization performance using three of these, outperforming very strong baselines.

[*] Work done during an internship at Amazon.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8014–8022
November 7–11, 2021.

Dialogue:
Orion: I miss him :(
Cordelia: Need i remind you that he cheated on you? You deserve alot better than that
Orion: ...what? oh, right noo - im talking about my rat ... he died
...
Vanilla BART Output:
Orion's rat died. He cheated on her.
MTL BART output:
Orion's rat died and he misses him.
Reference:
Orion is grieving after the death of her rat.

Table 1: Example from SAMSum (Gliwa et al., 2019) of a dialogue and its generated summaries using two BART models: vanilla and multi-tasked. The summary generated by the vanilla model indicates that the rat is the cheater, pointing to a lack of commonsense reasoning on the model side. The output of our multi-tasked model (section 3.5) clearly shows better understanding of the dialogue.

2 Related Work

Early work on dialogue summarization focused more on extractive than abstractive techniques for summarization of meetings (Murray et al., 2005; Riedhammer et al., 2008) or random conversations (Murray and Renals, 2007). In the context of meeting summarization, Shang et al. (2018) proposed an unsupervised graph-based sentence compression approach for meeting summarization on the AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003) benchmarks. Goo and Chen (2018) leveraged hidden representations from a dialogue act classifier through a gated attention mechanism to guide the summary decoder.

More recently, Gliwa et al. (2019) proposed SAMSum, a benchmark for abstractive everyday dialogue summarization. Zhao et al. (2020) modeled dialogues using a graph structure of words and utterances and summaries are generated using a graph-to-sequence architecture. Chen and Yang (2020) proposed a multi-view summarization model, where views can include topic or stage. They also pointed out to seven different challenges to dialogue summarization and analysed the effect each challenge can have on summarization performance using examples from SAMSum.

3 Challenges

We now present our four techniques for dialogue summarization: name substitution (section 3.3), negation scope highlighting (section 3.4), multi-task learning on common sense tasks (section 3.5), and pretraining on an in-domain dialogue corpus (section 3.6).

3.1 Experimental Setup

For all our experiments, we use BART large architecture (Lewis et al., 2020).[1] All our experiments are run using fairseq (Ott et al., 2019).

3.2 Baselines

We compare our techniques to two summarization baselines:

- **Vanilla BART**: Fine-tuning the original BART large checkpoint model on SAMSum.
- **Multi-view Seq2Seq** (Chen and Yang, 2020) : This is based on BART, as well, but during the summarization, the model considers multiple views, each of which defines a certain structure for the dialogue. We compare to their best model which combines topic and stage views.

3.3 Multiple Speakers

We hypothesize that uncommon (less frequent in the original pretraining data) or new names could be an issue to a pretrained model, especially if such names were seen very few times, or not at all, during pretraining. Such issues could specifically show up in multi-participant conversations, and could introduce co-reference issues when generating the summary. As a simple technique to alleviate this, we preprocess SAMSum by replacing speaker names with more common, frequent names, ones that the model is more likely to have seen during pretraining. Since we are dealing with English dialogue summarization, we use a list[2] of common English names and replace each speaker name with a randomly sampled same-gender name from this list. Since the name list is divided by gender (male or female), we use `gender guesser`[3] to replace with a same-gender name. To avoid modifying the ground truth summaries and to ensure a fair comparison with other models, the original name is replaced back into the generated summary before evaluation.

[1] For fine-tuning, we use ADAM optimizer with a learning rate of 0.00002 and label smoothing with $\alpha = 0.1$.

[2] https://www.ssa.gov/oact/babynames/decades/century.html

[3] https://github.com/lead-ratings/gender-guesser

Table 2 compares the performance of this technique to fine-tuning BART on the original SAMSum data. We observe ROUGE improvements on both validation and test sets of SAMSum. In addition, we perform an analysis of the performance with respect to the number of participants per dialogue. Figure 1 plots the summarization performance against the number of speakers. We can see that conversation with more participants (7, 8, 12) exhibit higher ROUGE boost than conversations with fewer speakers (2, 3, 4). In other words, we observe that the more participants in the summary, the more effect this technique has. Notably, the average number of speakers per dialogue in SAMSum is only ~2.4. and we expect name substitution to work even better with datasets that have many more speakers per dialogue.

3.4 Negation Understanding

Chen and Yang (2020) argue that negations represent a challenge for dialogues. We experiment with marking negation scopes in the input dialogues before feeding them to BART. To do that, we fine-tune a RoBERTa base model on the CD-SCO dataset from SEM Shared Task 2012 for negation scope prediction (Morante and Blanco, 2012). Then, we mark negation scope using two designated special tokens to mark the start and the end of the negation scope. For example, the sentence *"I don't know what to do"* becomes *"I don't* `<NEG>` *know what to do* `<\NEG>`*"* after negation scope highlighting. We initialize the embeddings of the special tokens `<NEG>` and `<\NEG>` randomly.

Results are shown in Table 3. While we expected to see a performance boost due to negation scope highlighting, we actually saw a performance drop except on ROUGE-L on the test set. To understand why, we investigate the negation challenge dialogues put together in (Chen and Yang, 2020). We found that in all examples, negation did not seem to be a problem, and that BART was able to handle multiple negations very well. Therefore marking negation scopes could have introduced unneeded noise into the model, causing the observed performance drop.

3.5 Reasoning

Reasoning is often necessary for dialogue summarization (Chen and Yang, 2020), especially in cases where there is missing information or implicit assumptions regarding the situation. Unfortunately, it is difficult for the model to learn to conduct such reasoning by relying only on the reference summaries (this difficulty is exacerbated by the fact that SAMSum is of a relatively small size). Multi-task learning (MTL) enables knowledge transfer across relevant tasks. For instance Li et al. (2019) improved their summarization performance by jointly learning summarization and topic segmentation. Also, Konar et al. (2020) improved commonsense reasoning through multi-task learning on relevant datasets. Similarly, we propose to simultaneously learn summarization and other reasoning-based tasks.

More specifically, we jointly fine-tune BART on the following tasks :

- **Short Story Ending Prediction**: this task could be helpful as predicting story ending requires intuitive understanding of the events. Also, conversation endings could be essential to understand the point of the dialogue (See examples 1 and 2 in Table 7 in the Appendix A). We use the ROC stories dataset (Mostafazadeh et al., 2016).
- **Commonsense Generation**: Generative commonsense reasoning (Lin et al., 2020) is a task involving generating an everyday scenario description given basic concepts. We assume such task could help the model reason more about conversations, which is certainly needed in many dialogues (see example 3 in Table 7 in Appendix A).
- **Commonsense Knowledge Base Construction**: The task here is to generate relation triplets similar to (Bosselut et al., 2019). More specifically, we train our model to predict relation objects given both relation and subject. We use ConceptNet (Liu and Singh, 2004).

Table 4 shows the summarization performance after multi-task fine-tuning of BART. We also show the results of combining ROC and CommonGen with SAMSum. It is clear that MTL gives a performance boost in almost all cases, outperforming the vanilla BART and the Multi-view SS baseline on both the development and test sets. It is worth noting that due to the small size of both validation and test splits (~800 dialogues), it is difficult to test the statistical significance of these results.

3.6 Informal Language

We hypothesize that pretrained language models, BART in our case, are not well-adapted to the dia-

Data	Val			Test		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
SAMSum	49.22	26.47	47.80	48.65	25.20	47.08
SAMSum + name substitution	**49.98**	**26.50**	**48.48**	**49.09**	**25.91**	**47.87**

Table 2: ROUGE-1, ROUGE-2, and ROUGE-L on SAMSum with and without names substitution. Results are shown on the validation and test splits from (Gliwa et al., 2019).

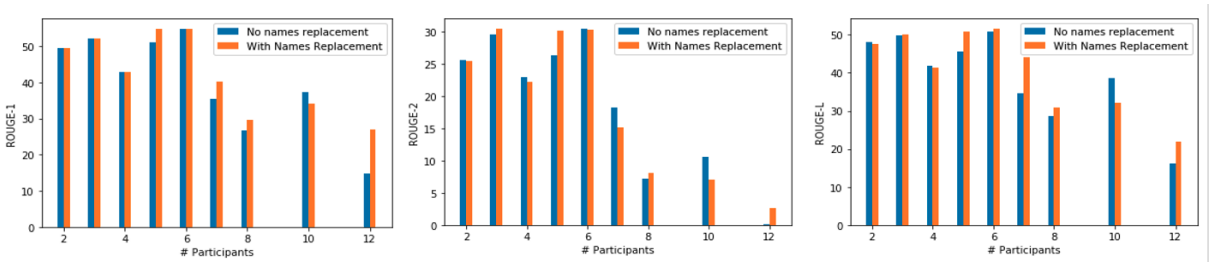

Figure 1: ROUGE values against the number of participants per dialogue on the development set of SAMSum. Performance boost is more clear in dialogues with more participants

logue domain. Therefore, we adapt BART to dialogue inputs by further pretraining of BART on a dialogue corpus and with dialogue-specific objectives. [4]

3.6.1 Pretraining Corpora

We consider the following 2 corpora for further pretraining of BART: PersonaChat (140K utterances) (Zhang et al., 2018), and a collection of 12M Reddit comments. We experiment with both whole word masking and span masking (masking random contiguous tokens). Our experimental setup is described in the Appendix in section B.1.

Table 5 shows the results of fine-tuning BART pretrained on dialogue corpora.[5] The best model (PersonaChat, word masking) outperforms the vanilla BART on all metrics and the Multiview SS baseline on test set ROUGE-2 and ROUGE-L. We can see that in general, BART pretrained on PersonaChat is better than pretraining on both PersonaChat and Reddit, which is surprising since more pretraining data usually means better performance. This could be explained by the dissimilarity between Reddit comments and the dialogues in SAMSum. We can also see that whole word masking performs slightly better than span masking. Based on these results, it is obvious that further pretraining on in-domain corpora can be helpful when dealing with inputs of special nature such as dialogues.

Also, we can see that pretraining using dialogue-specific objectives is performing well (on either PersonaChat only or with Reddit), and even outperforming random span masking on the validation set. This certainly shows that task-specific pretraining could be beneficial.

At last, we combine pretraining with MTL by fine-tuning a pretrained model in a multi-task learning fashion. Table 6 compares this to separate pretraining and MTL. We can see that pretraining on PersonaChat and fine-tuning on both SAMSum and ROC gives the best performance over the validation set, outperforming all other settings. On the test set, it is performing very well but slightly outperformed by multi-tasking with ROC in both ROUGE-2 and ROUGE-L. Lastly, we combine named substitution with the best model here and the results are also shown in Table 6. We observe that name substitution does not give a performance boost when used in combination with pretraining and MTL.

4 Conclusion

In this paper, we explored different techniques to improve dialogue summarization performance by addressing different challenges to the task individually. The proposed techniques included name substitution, negation scope highlighting, multi-task learning with relevant tasks, and pretraining on in-domain corpora. On one hand, our experiments on three challenges showed the effectiveness of our proposed techniques which outperformed strong baselines on the task. On the other hand, our proposed technique to handle multiple negations performed poorly and by analyzing the outputs on

[4] Our proposed dialogue-specific pretraining objectives are explained in Appendix B.2.

[5] We experimented with pretraining only on Reddit, but found it to perform worse.

Data	Val			Test		
	R-1	R-2	R-L	R-1	R-2	R-L
Original SAMSum	**49.22**	**26.47**	**47.80**	**48.65**	**25.20**	47.08
SAMSum + negation scope marked	48.61	25.45	47.82	48.59	24.96	**47.32**

Table 3: Summarization performance on SAMSum when highlighting negation scope.

Tasks	Val			Test		
	R-1	R-2	R-L	R-1	R-2	R-L
SAMSum	49.22	26.47	47.80	48.65	25.20	47.08
Multi-view SS (Chen and Yang, 2020)	-	-	-	49.30	25.60	47.70
SAMSum + ROC	**50.44**	26.63	**48.78**	49.31	26.18	**48.18**
SAMSum + CommonGen	50.09	**26.86**	48.73	49.12	25.76	47.71
SAMSum + ConceptNet	49.70	26.65	48.26	49.03	25.71	47.92
SAMSum + ROC + CommonGen	49.22	26.47	47.80	**49.45**	**26.20**	47.93

Table 4: Summarization performance on SAMSum when fine-tuning BART with multi-task learning of Commonsense generation (CommonGen), Knowledge Base Construction (ConceptNet), and Story Ending completion (ROC).

Pretraining Corpus	Val			Test		
	R-1	R-2	R-L	R-1	R-2	R-L
Original BART	49.22	26.47	47.80	48.65	25.20	47.08
Multi-view SS (Chen and Yang, 2020)	-	-	-	**49.30**	25.60	47.70
PersonaChat (entities, pronouns, tfidf)	50.07	26.81	48.68	48.66	25.26	47.39
PersonaChat (span masking)	49.59	26.11	47.97	48.88	25.52	47.63
PersonaChat (word masking)	**50.17**	**26.99**	**48.95**	49.22	25.64	**47.90**
PersonaChat + Reddit (entities, pronouns, tfidf)	49.64	26.31	48.38	48.43	25.09	47.23
PersonaChat + Reddit (span masking)	49.43	25.92	48.00	49.20	**25.87**	47.74
PersonaChat + Reddit (word masking)	49.12	26.03	47.84	48.99	25.52	47.63

Table 5: Summarization performance on SAMSum when BART is pretrained on an in-domain corpus. We also include results when using additional dialogue-specific pretraining objectives (See Appendix B.2).

Tasks	Val			Test		
	R-1	R-2	R-L	R-1	R-2	R-L
Original BART	49.22	26.47	47.80	48.65	25.20	47.08
Multi-view SS (Chen and Yang, 2020)	-	-	-	49.30	25.60	47.70
SAMSum + ROC	50.44	26.63	48.78	49.31	**26.18**	**48.18**
Pretraining + MTL(SAMSum, ROC)	**50.48**	**27.25**	**48.90**	**49.34**	25.54	47.88
Pretraining + MTL(SAMSum, ROC, CommonGen)	50.29	27.21	49.05	**49.34**	25.81	47.85
Pretraining + MTL(SAMSum, ROC) + name substitution	49.97	26.94	48.88	48.87	25.70	47.72

Table 6: Summarization performance on SAMSum when BART is pretrained on an in-domain corpus and then fine-tuned in a multi-task fashion.

negation-intensive dialogues, we found that multiple negations do not represent a challenge for dialogue summarization systems.

5 Ethics Discussion

We refer to Section 3.3, where we explain how we aid the model with name substitution using more common names (common here means more frequent in the pre/training data, and not by any preconception to us or any other entity). As explained above, we are using a list of the most common names in American English, which is divided in feminine and masculine names. We therefore use `gender guesser` to ensure that the pronouns in the dialogue co-refer correctly with the replaced names. It is however worth mentioning that even if the character in the dialogue is non binary and/or the pronouns used in the dialogue are *they/them*, our approach would work given that the replaced name would still co-refer with those pronouns and the name that is being replaced. We however hope to work in the future with datasets and list of names that contain non-binary gender.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association*

for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 4762–4779. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 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 2020, Online, November 16-20, 2020*, pages 4106–4118. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Anandh Konar, Chenyang Huang, Amine Trabelsi, and Osmar R Zaiane. 2020. Ana at semeval-2020 task 4: Multi-task learning for commonsense reasoning (union). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 367–373.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Bill Yuchen Lin, Ming Shen, Wangchunshu Zhou, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

Roser Morante and Eduardo Blanco. 2012. * sem 2012 shared task: Resolving the scope and focus of negation. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Gabriel Murray and Steve Renals. 2007. Term-weighting for summarization of multi-party spoken dialogues. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 156–167. Springer.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 593–596. ISCA.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations, pages 48–53. Association for Computational Linguistics.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2008. A keyphrase based approach to interactive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 153–156. IEEE.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Klaus Zechner and Alex Waibel. 2000. Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449.

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.

A Reasoning

Here we show examples from SAMSum validation set where both story end understanding and reasoning about the situation are essential for correct summarization. Rows (1) and (2) in Table 7 are examples of a dialogue where the main point of the conversation is only known in the last utterance. Consequently, we hypothesize that learning to predict story endings could teach the model to focus more on dialogue endings.

Row (3) is an example of a situation that requires high-level commonsense reasoning. Given the information in the dialogue, it is *very difficult* for the model to infer that the conversation is about a marriage proposal. Through our error analysis, we find that incorrect or incomplete reasoning is a major source of error in summarization. For example the output of vanilla BART on this dialogue is: "Colin congratulates Patrick on his girlfriend", which shows that the model clearly misses the point. Our best MTL model, on the other hand, produces "Patrick is over the moon because she said yes.", which is certainly better than vanilla BART.

B In-domain pretraining

B.1 Experimental Settings

We continued pretraining BART for 50K gradient update steps with batch size of 1024 tokens and a learning rate of 0.00001. We use $p_{mask} = 0.3$ and for span masking, we sample span lengths from a Poisson distribution with $\lambda = 3$ and replace these with a single mask token. We do not replace by a random token similar to BERT (Devlin et al., 2019) as early experiments showed it does not perform very well.

B.2 Pretraining Objectives

The original BART pretraining involved a number of de-noising tasks including span masking, token deletion, sentence permutation, and document rotation (Lewis et al., 2020). However, we argue in this work that these objectives are overly general and not specific for the dialogue domain. Here, we describe our proposed pretraining tasks:

- **Masking pronouns** Conversations are usually rife with pronouns used for co-reference. In many cases. predicting the correct pronoun would require sufficient understanding of the dialogue context. We use a separate probability $P_{mask\_pronoun} = 0.5$ of masking a specific pronoun.

- **Masking High-content tokens** While BART masking objective treats all tokens equally i.e all tokens are equally likely to be masked, we know that certain tokens are more relevant to a particular dialogue than other. Thus, here we choose to mask more salient tokens where salience is measured using TF-IDF weights. We start by computing TF-IDF weights over the whole dataset. Then for every input, we select the top 25% weighted tokens and mask these with probability $p_{mask\_tfidf} = 0.7$. This is somehow similar to PEGASUS (Zhang et al., 2020), but here we mask tokens not sentences.

- **Masking Entities** Guu et al. (2020) showed that masking entities and dates could be helpful for IR tasks. We hypothesize that masking entities such as persons and locations can be particularly important for dialogues. Here we mask entities with probability $P_{mask\_entity} = 0.7$. We use Spacy[6] English NER model to detect entities.

[6] https://spacy.io/

#	Dialogue	Reference
1	**Keith:** Meg, pls buy some milk and cereals, I see now we've run out of them **Megan:** hm, sure, I can do that **Megan:** but did you check in the drawer next to the fridge? **Keith:** nope, let me have a look **Keith:** ok, false alarm, we have cereal and milk :D	Megan needn't buy milk and cereals. They're in the drawer next to the fridge.
2	**Taylor:** I have a question!! **Isabel:** Yes? **Taylor:** Why haven't you introduced me even once your bf to me? **Taylor:** All of my friends' daughters bring their bfs and introduced them. **Taylor:** You know I'm such a cool mum. I won't make him stressful. **Taylor:** Just bring him. **Isabel:** Because mum...I haven't had any!	Taylor wants to meet Isabel's boyfriend but she has never had any.
3	**Colin:** DUUDE, congrats! **Patrick:** Thanks! **Patrick:** She said yes, I'm over the moon! **Colin:** Lucky guy	Patrick's girlfriend accepted his proposal.

Table 7: Sample dialogues from SAMSum (Gliwa et al., 2019) that require reasoning for correct understanding/summarization.