Mind the gap! Further Investigation on Commonsense Effectiveness

1st Alessandro De Marco
dept. Computer Engineering
Politecnico di Torino
Turin, Italy
s317626@studenti.polito.it

2nd Sergiu Abed
dept. Computer Engineering
Politecnico di Torino
Turin, Italy
s295149@studenti.polito.it

3rd Riccardo Musumarra
dept. Computer Engineering
Politecnico di Torino
Turin, Italy
s295103@studenti.polito.it

4th Tito Ndjekoua Nzeumi
dept. Data Science
Politecnico di Torino
Turin, Italy
s297926@studenti.polito.it

***Abstract*—In this paper, we explore the impact of commonsense knowledge on abstractive dialogue summarization, building upon the "Mind the gap!" framework that integrates commonsense information with dialogue utterances using the COMET system for enhanced summary generation. Our study further explores the SICK model and its advanced version, SICK++. This dual-decoder architecture aims to ensure the effective use of commonsense knowledge in the summarization process. We conduct two sets of experiments to evaluate the utility of commonsense integration. The first experiment compares the performance of different models (T5-small, T5-base, BART-base, and PEGASUS-large) trained on the DialogSum and SamSum datasets, both with and without commonsense augmentation. The second experiment extends our investigation to the TweetSumm dataset, focusing on Twitter customer care conversations, to assess the models' effectiveness in a distinct, real-world context. This involves training the SICK and SICK++ models on TweetSumm with varying degrees of commonsense integration. Our research aims to ascertain whether the inclusion of commonsense knowledge significantly enhances the quality and relevance of abstractive summaries. Our findings reveal that while the integration of commonsense knowledge does not uniformly improve summarization performance across all tested models, the advanced SICK++ model demonstrates a notable improvement in summary quality on the TweetSumm dataset, underscoring the potential benefits of structured commonsense integration in specific contexts**

I. Introduction

Abstractive dialogue summarization is a challenging task within the field of natural language processing (NLP), aiming to generate concise summaries of dialogues while preserving essential and underlying information. The generation of summaries has garnered significant attention and achieved notable advancements in NLP. Unlike traditional summarization, dialogue summarization demands the extraction of not explicitly stated intentions or implied information among dialogue participants, adding layers of complexity to the task.

Numerous studies have explored dialogue summarization using models like GPT, BART, and PEGASUS, trained on datasets such as SamSum or DialogueSum, dedicated to dialogue summarization. While these studies have reported satisfactory results, there remains room for improvement, especially in capturing the nuanced and hidden aspects of dialogues.

A particularly notable study [1] highlights the critical role of hidden information in dialogues and how commonsense knowledge can enhance the accuracy of abstractive summaries. The approach involves generating commonsense knowledge for each utterance in a dialogue using models like COMET [2], designed to infer commonsense knowledge from text. This knowledge can encompass social relations (e.g., XINTENT, XWANT) or be event-based (e.g., HINDEREDBY, XREASON, XNEED), effectively "filling in the gaps" in the dialogue.

The study introduces the SICK model and its extended version, SICK++, both based on an encoder-decoder structure utilizing BART-large as the foundational model. SICK processes dialogue utterances, each followed by its associated commonsense knowledge, to produce abstractive summaries. SICK++, with an additional decoder, generates commonsense for the dialogue summary itself, ensuring the model considers commonsense in the summarization process. This innovative approach has shown promising results in enhancing the quality of abstractive summaries.

Our research[1] aims to validate the utility and effectiveness of incorporating commonsense knowledge into the summarization process. We conducted two main experiments: the first involved using various models such as the T5 family, BART-base, and PEGASUS as the base for the SICK model. The second experiment utilized the TweetSumm dataset, which comprises customer support dialogues from Twitter, to test the impact of commonsense knowledge on summarization quality.

II. Methodology

As mentioned earlier, our work builds on top of "Mind the gap!" [1] work, where commonsense knowledge derived from the utterances of the dialogue dataset is injected together with the input dialogues to improve abstractive summarization results.

[1]Github folder: https://github.com/sergiuabed/SICK_Summarization/

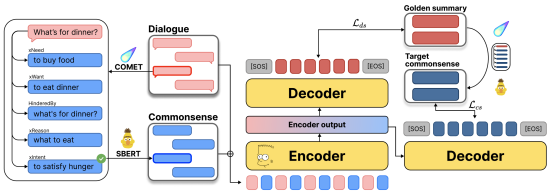

Fig. 1. The overall framework of SICK and SICK++. The decoder generating target commonsense is used for SICK++

A. Pipeline

Before discussing the extensions, we would like to recall the general pipeline

1) Compute commonsense on each utterance of a dialogue:

$$f_{COMET}(u_i) = c_i, \forall u_i \in D$$

where $D = \{u_1, u_2, ..., u_n\}$ is a set containing the utterances in a dialogue.

2) Cross-concatenate the utterances with their respective commonsense knowledge:

$$x_i = u_i + < I > c_i < /I >$$

where $< I >$ and $< /I >$ are special tokens used for encapsulating the commonsense before appending it to the utterance.

3) Put the results obtained at the previous step together to rebuild the dialogue, which now contains the commonsense knowledge:

$$D' = \{x_1, x_2, ..., x_n\}$$

4) Feed D' to the model to obtain the summary and, in the case of SICK++, generate also commonsense knowledge providing additional information to *fill in the gap* in the summary.

B. COMET: Commonsense Transformer

Developed by integrating transformer-based models with a structured knowledge base, COMET is designed to infer and generate commonsense knowledge entries that are not explicitly stated in the text, thereby enabling AI models to perform more human-like reasoning and interpretation of natural language contexts.

With a total of 23 distinct relationship types, COMET offers a comprehensive toolkit for enriching natural language understanding tasks with nuanced and context-specific commonsense inferences.

For the purposes of our study, we meticulously selected five relationship types that are particularly adept at encapsulating the context of utterances within dialogues. These are:

- **HinderedBy**: Identifies obstacles or challenges that may impede actions or outcomes.
- **xWant**: Captures desires or wants of the subject, often implicit in the dialogue.

TABLE I
EXAMPLE OF COMMONSENSE KNOWLEDGE GENERATED BY COMET GIVEN AN UTTERANCE FROM TWEETSUMM

Utterance	Customer: @VirginTrains what's off peak times between WGN AND EUS on 27/12/17
HinderedBy	The service is not working.
xWant	to fix the problem
xIntent	to be helpful
xNeed	to call the service
xReason	PersonY asks for help

- **xIntent**: Reflects the intentions or goals underlying the subject's actions.
- **xNeed**: Pinpoints necessities or requirements for actions to take place.
- **xReason**: Elucidates reasons or motivations behind actions or events.

These selected relationships are instrumental in generating commonsense knowledge that deeply contextualizes the utterances, providing our models with a richer understanding of the dialogues(e.g. Table I).

C. Metrics

To assess the performance of the various models, we used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores metrics. More specifically, we used the following:

- ROUGE-N ($N = 1, 2, 3$), which measures N-gram co-occurrences between reference and candidate summaries;
- ROUGE-L, measuring the longest common subsequence between the generated and target summaries;
- ROUGE-Lsum, which is similar to ROUGE-L, except that it is computed over the entire summary (ROUGE-L is an average over individual sentences)

D. Extension 1: Injecting commonsense on several other models

Given the effectiveness of using commonsense knowledge on BART (i.e. the backbone of SICK), we asked ourselves how beneficial would it be to use commonsense knowledge when training other well-known transformer-based models. To answer this question, we chose the following models: T5-small, T5-base, PEGASUS, BART-base

T5 is a family of sequence-to-sequence models consisting of 5 different variations: T5-small, T5-base, T5-large, T5-3b and T5-11b. We chose to test the first two, since they are the smallest in terms of number of parameters and we have limited computational power for training large language models.

PEGASUS [8] is designed specifically for the task of abstractive text summarization. It introduces a novel pre-training technique where sentences deemed to be the most informative are removed and then the model is tasked with generating these "gap sentences" from the remaining text, closely mirroring the summarization process.

BART-base is a smaller version of the BART model used in the implementation of SICK. We wanted to see what kind of results we would get on a smaller model.

E. Extension 2

The second extension includes the TweetSumm dataset, which comprises Twitter-based customer service dialogues, aiming to evaluate the performance of our models in a unique, practical scenario. This phase involves adapting the SICK and SICK++ models for use with TweetSumm, incorporating varying levels of commonsense knowledge. The objective of our research is to determine the extent to which integrating commonsense knowledge can substantially improve the quality and pertinence of the generated abstractive summaries.

III. Experiments

A. Experimental design

- Hardware:
  - First extension: NVIDIA Tesla T4 GPU on Google Colab
  - Second extension: NVIDIA A100 GPU on Google Colab Pro
- Software/Libraries: PyTorch and Huggingface Transformers
- Validation method: Hold-out (Train and Validation sets splits)
- Performance metrics: ROUGE scores

B. First experiment

1) Datasets: For the first experiment we utilizes two prominent datasets designed for dialogue summarization and topic generation: SamSum [3] and DialogSum [4].

For commonsense knowledge, we used the ones precomputed by the authors of *Mind the gap!* on SAMSum and DialogSum. They generated commonsense using both COMET and PARA-COMET. Unlike COMET, PARA-COMET generates commonsense knowledge by considering also previous utterances, not just the current one.

2) Results: Table 2 reports the results obtained during this experiment. BART-base gained an increase in ROUGE scores when using commonsense generated by COMET for SAMSum, whereas T5-base saw better results on DialogSum when using COMET. T5-small and PEGASUS did not see any improvement.

Overall, commonsense did not prove to be very useful, since in most cases it resulted in slightly lower scores and in the cases where it gave better results, the gain was not substantial.

One important detail regarding the first experiment is that we used the free version of Google Colab and this proved to be not suited for this kind of work. Due to limited availability of the GPU, we encountered timeout prompts during multiple runs, which led to incomplete experiments. For most of the models (i.e. T5-small, T5-base and BART-base), we managed to fine-tune them for 1-3 epochs, whereas for PEGASUS we couldn't complete one epoch before the process being stopped.

PEGASUS is the largest model among the ones we tested and because of this we had to reduce the size of the training set. When training on SAMSum, we reduced the dataset to 15%, whereas on DialogSum we were only able to train the model without adding commonsense knowledge. When adding commonsense knowledge, the GPU would run out of memory and the only way to test its effectiveness would have been to reduce DialogSum under 10%, which we think it is not very fruitful.

For the second experiment, we decided to purchase a subscription for Google Colab Pro.

C. Second experiment

For our second experiment, we chose to explore the capabilities of our models on a dataset distinct from those typically encountered in dialogue summarization research: Tweetsumm [5].

1) TweetSumm: The TweetSumm dataset is derived from the "Kaggle Customer Support on Twitter" dataset [6], which is a vast repository of conversations between consumers and customer support agents on twitter.com, covering a wide range of sectors from airlines to video games. The original Kaggle dataset consists of 49,155 unique dialogues. A preliminary filtration process was conducted to remove dialogues that were either too short or involved more than two parties, resulting in a refined collection of 32,081 dialogues. From this collection, the creators of TweetSumm randomly selected 1,100 unique dialogues for inclusion in the dataset.

For each of these dialogues, up to three extractive summaries and three abstractive summaries were generated, leading to a total of 6,500 summaries. The summary generation employed models such as BART-large, followed by a rigorous human evaluation to ascertain the summaries' quality and relevance. The TweetSumm dataset is organized into a training set of 880 dialogues, with 110 dialogues allocated for both the validation and test sets.

2) Commonsense Generation Process: To infuse our dialogue summarization models with nuanced understanding, we meticulously selected five key relations from the twenty-three available within the COMET framework: *HinderedBy, xWant, xIntent, xNeed,* and *xReason*. For each of these relations, we generated five distinct commonsense assertions.

We employed Sentence-BERT (SBERT), a modification of the pre-trained BERT network that excels in capturing semantic similarity between pieces of text. By computing similarity scores between each utterance (or summary) and the corresponding set of generated commonsense assertions, we were able to identify and select the assertion with the highest relevance as measured by its similarity score. This ensures that the chosen commonsense knowledge closely aligns with the context and content of the utterance or summary, thereby enriching the model's output with meaningful and contextually appropriate insights.

TABLE II
EXTENSION 1 ROUGE SCORES OF SEVERAL MODELS WITH AND WITHOUT COMMONSENSE

		Samsum				Dialogsum			
Model	Commonsense	R1	R2	RL	RLsum	R1	R2	RL	RLSum
T5-Small	None	**46.63**	**22.87**	**38.56**	**43.03**	**41.73**	**16.35**	**33.39**	**37.30**
	COMET	46.01	22.31	38.01	42.35	41.52	16.02	33.19	37.02
	PARACOMET	45.11	21.58	37.40	41.70	41.55	16.05	33.17	37.07
T5-Base	None	**48.88**	**24.34**	**40.44**	**45.16**	44.73	19.13	35.87	39.74
	COMET	48.09	24.09	39.94	44.67	**45.37**	**19.82**	36.69	**40.46**
	PARACOMET	48.10	24.12	39.99	44.64	45.34	19.79	**36.73**	**40.46**
BART-Base	None	48.95	25.46	41.11	45.14	**47.76**	**23.40**	**40.08**	**43.04**
	COMET	**49.75**	**25.50**	**41.42**	**45.44**	47.32	22.82	39.39	42.47
	PARACOMET	49.04	25.12	41.07	45.13	47.68	23.11	40.01	42.89
PEGASUS	None	**42.58**	**18.86**	**33.62**	**38.44**	-	-	-	-
	COMET	40.36	18.12	33.34	35.93	-	-	-	-
	PARACOMET	23.77	7.63	19.73	21.16	-	-	-	-

3) Results: This section presents the outcomes of our second experiment, which involved training the SICK and SICK++ models on the TweetSumm dataset across 30 epochs. The primary objective was to assess the impact of commonsense knowledge and commonsense supervision on the quality of dialogue summarization, as measured by ROUGE scores.

TABLE III
COMPARISON OF ROUGE SCORES FOR SICK AND SICK++ MODELS ON THE TWEETSUMM DATASET.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum
BART-large	32.71	12.32	27.10	30.25
Sick	40.23	17.53	33.67	36.71
Sick++	**41.60**	**18.60**	**35.39**	**38.31**

4) Analysis of the results: The results outlined in Table III illustrate a significant improvement in summarization quality with the incorporation of commonsense knowledge, as evidenced by the superior ROUGE scores of SICK and SICK++ models compared to the baseline BART-large. Notably, the SICK++ model, which integrates commonsense supervision, achieves the highest scores across all ROUGE metrics, underscoring the value of embedding commonsense knowledge in the summarization process.

Our initial experiments conducted using the free tier of Google Colab were limited by the constraints on computational resources, leading to suboptimal training conditions for the models. Specifically, models such as T5-small, T5-base, and BART-base were trained across the entire dataset but for only 1 to 3 epochs, whereas PEGASUS was trained on merely 15% of the dataset for 3 epochs. In contrast, for the second experiment with the TweetSumm dataset, we opted for Colab Pro, enabling us to train the models on the full dataset for 25 epochs.

The significant difference in outcomes between our two experimental setups clearly shows that proper training is crucial for realizing the advantages of incorporating commonsense knowledge into abstractive summarization tasks. The substantial performance gap between BART-large, SICK, and SICK++ in the second experiment validates our assertion that commonsense can significantly improve summary generation, provided the models undergo comprehensive training.

IV. CONCLUSION

In this study, we embarked on an ambitious exploration of integrating commonsense knowledge into abstractive dialogue summarization, leveraging the SICK and SICK++ models. Our investigation yielded remarkable outcomes, demonstrating that commonsense augmentation can indeed enhance the quality of generated summaries, as evidenced by the performance improvements observed when moving from traditional BART-large to SICK and further to SICK++ models. These findings underscore the value of embedding deeper contextual insights into summarization processes, aligning with the pioneering assertions made in the "Mind the Gap!" paper.

The primary challenge encountered was the initial training of models on the constrained environment of Google Colab's free tier, which led to an underestimation of the utility of commonsense knowledge due to limited training epochs.

This experience underlines a critical lesson: the depth of model training significantly influences the effectiveness of commonsense integration in enhancing summarization tasks.

Moving forward, we are inspired to explore alternative methods for selecting and integrating commonsense knowledge, beyond the current similarity-based approach, to further enrich the quality of abstractive summaries.

REFERENCES

[1] Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seungwon Hwang, and Jinyoung Yeo. 2022. Mind the Gap! Injecting Commonsense Knowledge for Abstractive Dialogue Summarization. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6285–6300, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

[2] Hwang, J.D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2020). COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI Conference on Artificial Intelligence.

[3] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SamSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, Hong Kong, China, Nov. 2019, pp. 70–79, doi: 10.18653/v1/D19-5409. [Online]. Available: https://www.aclweb.org/anthology/D19-5409

[4] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, “DialogSum: A Real-Life Scenario Dialogue Summarization Dataset,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 5062–5074, doi: 10.18653/v1/2021.findings-acl.449. [Online]. Available: https://aclanthology.org/2021.findings-acl.449

[5] G. Feigenblat, C. Gunasekara, B. Sznajder, S. Joshi, D. Konopnicki, and R. Aharonov, ”TWEETSUMM – A Dialog Summarization Dataset for Customer Service,” 2021, arXiv:2111.11894. [Online]. Available: https://arxiv.org/abs/2111.11894

[6] S. Axelbrooke, ”Customer Support on Twitter,” Kaggle, 2017. [Online]. Available: https://www.kaggle.com/dsv/8841, DOI: 10.34740/KAGGLE/DSV/8841.

[7] 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,” 2019, arXiv:1910.13461. [Online]. Available: https://arxiv.org/abs/1910.13461

[8] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, ”PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” 2020, arXiv:1912.08777. [Online]. Available: https://arxiv.org/abs/1912.08777

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, ”Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” 2023, arXiv:1910.10683. [Online]. Available: https://arxiv.org/abs/1910.10683

[10] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.