Abstractive Dialogue Summarization

Ali Magzari
EECE Department
The University of British Columbia
Vancouver, BC, Canada
amagzari@student.ubc.ca

Abstract

The information technology and communications revolution has led to an explosion of information that has never been witnessed before. At the beginning of 2020, data quantity was estimated to be 44 zettabytes across the world. This number is expected to reach 463 exabytes by 2025. This substantial increase has not spared text data, which motivates the crucial need to explore effective text summarization techniques. Extractive summarization has first approached the issue before scientific findings have led to the development of abstractive text summarization methods. These methods generate a summary based on paraphrasing and using vocabulary that is unseen in the source document, thus generating human-level summaries. This paper reviews, implements and compares the performance of various versions of three major transformer-based encoder-decoder models: T5, BART, and PEGASUS. More specifically, the scope of this project is to fine-tune and evaluate the results of the aforementioned models in the context of dialogue summarization.

1 Introduction

In light of the information age, the amount of generated data is exponentially increasing. This leads to the issue of information overload that is responsible for killing productivity and dampening creativity [1]. This phenomenon is prevalent across data of different nature, text being one of them. As a matter of fact, in recent years, the amount of text data has witnessed an explosion from various sources [2]. The overwhelming amount of generated text documents has motivated important research in the area of automatic text summarization [2].

The rise of video conferencing in the realm of remote team collaboration and job interviews, coupled with the recent advances in speech-to-text transcription stimulates an opportunity to develop and improve text-based dialogue summarization for higher efficiency. While speech-to-text transcription is an essential part of the process, the project focuses solely on the summarization of text-based dialogue. An automatically generated conversation summary could improve various internal processes such as the rendition of meeting minutes after a team call, or the key elements of a video interview for faster hiring decisions.

Extractive summarization consists of generating a shorter version of a document by extracting relevant sentences from the original document to preserve the most salient information [3]. On the other hand, while it captures the salient idea of a passage, abstractive summarization paraphrases the main content of the document by potentially using vocabulary unseen in the original document [4]. Extractive summarization approaches the problem by choosing the most significant sentences from the source document. This is achieved by scoring sentences based on their features, after which the sentences with the highest scores are selected to appear in the summary [5]. A number of methods have been used to this end. Zhang et Li [6] propose a cluster based approach developed in three stages: sentence clustering based on the semantic distance among the document sentences,

Preprint. Under review.

calculation of accumulation sentence similarity, and selecting topic sentences via extraction rules. Neural networks have also been utilized to summarize new articles. Kaikhah [7] trains a neural network to learn relevant sentence features before modifying it to act as a filter that summarizes news articles. Michalcea [8] presents a graph based method where every sentence is a vertex whereas the edge represents a common semantic relation with a specific weight. Vertices with high cardinality characterize important sentences that are then included in the summary. For the purpose of generating human-level text, abstractive summarization utilizes advanced deep learning methods. In 2019 Raffel et al at Google [9] introduce a Text-to-Text Transfer Transformer, also known as T5, which is a model founded on a unified framework that converts all text-based language problems into a text-to-text format. The model was designed to execute a number of tasks including translation, question answering, and classification. In the same month and year, Lewis et al at Facebook AI Language [10] present BART (Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension), which is an avant-garde NLP (Natural Language Processing) model that excelled at a variety of abstractive tasks. In December 2019, Zhang et al at Google AI Language [11] propose PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization), which achieves state-of-the-art performance on 12 downstream datasets.

2 Related Works

Designed by Google scientists, T5 is inspired by the concept of transfer learning where a model is pre-trained on a large corpus in hope of developing general abilities to be transferred to downstream tasks [9]. The possibility to pre-training NLP models on unlabeled data in an unsupervised manner proves to be very advantageous due to the large availability of text data on the internet, such as C4 [12], which stands for Common Crawl's web crawl corpus. T5 follows the standard encoder-decoder transformer architecture [13] barring the removal of the layer norm bias, the placement of the layer normalization outside the residual path, and the use of relative position embeddings instead of fixed embedding [9]. T5 is pre-trained on C4 after being cleansed through a set of actions such as only retaining lines ending in a terminal punctuation, and discarding pages with less than five sentences. The model was tested on various language abilities such as machine translation and question answering. The baseline version of the T5 model generated the following results in terms of summarization after being fine-tuned on the CNN/DM dataset, a famous article-summary records used to fine-tune and evaluate models on summarization tasks [14]: **42.05, 20.34 and 39.40 for R-1, R-2 and R-L respectively** [9]. This metric is discussed in the Performance Metrics subsection of this report. On the other hand, the T5-11B version was able to generate scores (**R-1: 43.52, R-2: 21.55, R-L: 40.69**) that surpassed the performance of UNILM [15], that reported the previous best scores at the time (**R-1: 43.47, R-2: 20.30, R-L: 40.63**).

Designed by Facebook AI, BART is formally defined as a denoising autoencoder for pretraining sequence-to-sequence models [10]. It consists of a bidirectional encoder and left-to-right decoder. The model relies on self-supervised training methods as they have showcased exceptional results in the realm of NLP tasks. The first step of pretraining involves corrupting text by randomly shuffling the order of the original sentences and using an in-filling scheme where text spans of arbitrary lengths are replaced with a single mask token, in hope of generalizing original word masking and next sentence prediction objectives. The second step is the learning of a sequence-to-sequence model to rebuild the original text. After being finetuned on the CNN/DM datset, the model generates the follwoing results: **R-1: 44.16, R-2: 21.28, R-L: 40.90** [10], which are more satisfactory than baseline T5 scores and T5-11B (except R-2).

Google researchers designed PEGASUS with the intent to explore pre-training objectives specific to abstractive text summarization [11]. They use a novel self-supervised objective called Gap Sentences Generation (GSG) which consists of masking entire sentences (gap enctences) from the source document and generating them as output. The idea is that such objective could allow the understanding of the entire document, and generate outputs that could be considered as abstractive summaries. The large PEGASUS model pre-trained on C4 achieved important results on the CNN/DM dataset: **R-1: 43.90, R-2: 21.20, R-L: 40.76**, while the model pre-trained on HugeNews (a corpus composed of 1.5 billion articles) produced even higher results: **R-1: 44.17, R-2: 21.47, R-L: 41.11** [11].

3 Model/Method

The idea is to directlty fine-tune base transformer models while pre-train a second time large transformer models on two different datasets, separately, and compare the generated summaries, as shown in Figure 1.

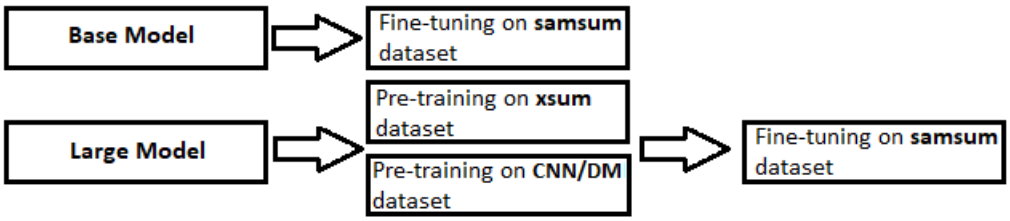

Figure 1: Model fine-tuning

The SAMSUM fine-tuning dataset [16] contains two main features: a dialogue, and its summary.

Table 1: SAMSUM Dataset

	Number of instances
Training set	14,732
Validation set	818
Test set	819

The XSUM pre-training dataset [17] contains two main features: a document, and its short summary.

Table 2: XSUM Dataset

	Number of instances
Training set	204,045
Validation set	11,332
Test set	11,334

The CNN/DM pre-training dataset [14] contains two main features: a document, and its highlights.

Table 3: CNN/DM Dataset

	Number of instances
Training set	287,113
Validation set	13,368
Test set	11,490

The following tables displays the models used in this study:

Table 4: CNN/DM Dataset

Model Name	Original Model	XSUM	CNN/DM	SAMSUM
T5-base	google/t5-v1_1-base [18]	No	No	Yes
BART-base	facebook/bart-base [19]	No	No	Yes
PEGASUS-base	google/pegasus-x-base [20]	No	No	Yes
BART-xsum	facebook/bart-large-xsum [21]	Yes	No	Yes
PEGASUS-xsum	google/pegasus-xsum [22]	Yes	No	Yes
BART-cnn/dm	facebook/bart-large-cnn [23]	No	Yes	Yes
PEGASUS-cnn/dm	google/pegasus-cnn_dailymail [24]	No	Yes	Yes

4 Experiments

This section details the following:

- The evaluation metrics used to assess the quality of the generated summary

- Quantitative results

- Qualitative results

4.0.1 Performance Metrics

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [25] is a set of metrics used to assess machine-genrated texts. ROUGE compares generated text to reference text via the equation:

$R - n = \frac{N_{gen,ref}}{N_{ref}}$

when $N_{gen,ref}$ = Number of n-grams present in both generated and reference text, and N_{ref} = Number of n-grams present in the reference text.

R-1, R-2 and R-L compute the recall of unigrams, bigrams and L-grams respectively. L-grams consist of the longest common sequence (LCS) which must be in the same order, but not necessarily continuous.

4.0.2 Quantitative Results

The table below displays the ROUGE scores for text generated by the base models. BART-base clearly leads the way in terms of R-1, R-2, and R-L, followed by T5-base, then PEGASUS-base.

Table 5: Results of base models

	T5-base	**BART-base**	**PEGASUS-base**
R-1	44.51	**47.39**	41.49
R-2	20.90	**24.07**	19.23
R-L	37.30	**40.03**	34.74

Table 6: Results of base models pretrained on the **xsum** or **cnn/dm datasets**

	BART-xsum	**PEGASUS-xsum**	**BART-cnn/dm**	**PEGASUS-cnn/dm**
R-1	**54.20**	48.68	42.27	45.30
R-2	**29.35**	24.43	21.81	21.76
R-L	**44.63**	40.74	32.52	35.86

4.0.3 Qualitative Results

This section shares the performance of the implemented models on unseen dialogues. The first test dialogue is an instance of the dialogsum dataset [26].

- The first test dialogue:

Andrew: Excuse me, do you know where the visa office is?
Mark: Yes, I do. I'll walk you there. Andrew: Thanks.
Mark: Are you applying to study or work abroad?
Andrew: I want to study abroad.
Mark: What do you want to study?
Andrew: I hope to study English Literature.
Mark: Have you got your visa yet?
Andrew: Not yet. I have an interview with a visa official today.
Mark: I see. Is it your first interview?
Andrew: No, I'Ve already been here for 3 interviews.

The generated summaries are as follow:

- T5-base:
Andrew has an interview with a visa official today. He has already been here for 3 interviews.

- BART-base:
Andrew has an interview with a visa official today. He has already been here for 3 interviews.

- PEGASUS-base:
Andrew wants to study English Literature. **Mark** has an interview with a visa official today.

- BART-xsum:
Mark will walk Andrew to the visa office. Andrew has an interview with a visa official today. Andrew wants to study English Literature.

- PEGASUS-xsum:
Mark will walk Andrew to the visa office. Andrew has already been there for 3 interviews.

- BART-cnn/dm:
Mark will walk Andrew to the visa office. Andrew wants to study English Literature abroad. He has an interview, he's already been to 3 interviews. Mark will walk him to the office. **He will meet Andrew there**.

- PEGASUS-cnn/dm:
Mark will walk Andrew to the visa office. Andrew is applying **to study or work abroad**. He has already been here for 3 interviews. Andrew has not yet got his visa.

- The second test dialogue:

Leila: Hi Omar, Congratulations! Morocco beat Spain. The Moroccan goalkeeper is great!
Omar: Yes, I'm very proud of Bounou and the rest of the team. It's the first time Morocco ever made it to the quarter-finals.
Leila: This is a great achievement. I'm over the moon.
Omar: Haha. Next game is against Portugal on Saturday.
Leila: Portugal is dangerous.
Omar: I know. The Moroccan players have to remain focused and give all they've got.
Leila: If they win, Morocco will be the first African Nation to ever make it to the semi-finals.
Omar: Oh my God, I'm so excited. Let's watch the game together!
Leila: Great idea! I'll bring my giant Moroccan flag.

The generated summaries are as follow:

- **T5-base**:
Omar is proud of Bounou and the rest of the team. Morocco beat Spain in the

- **BART-base**:
Morocco beat Spain in the quarter-finals. Leila and Omar will watch the game together.

- **PEGASUS-base**:
Omar and Leila are are excited about Morocco making it to the quarter-finals.

- **BART-xsum**:
Morocco beat Spain and made it to the quarter-finals. Morocco's next game is against Portugal on Saturday. Leila will bring her giant Moroccan flag.

- **PEGASUS-xsum**:
Omar and Leila are proud of Morocco's **victory** over Spain. They will watch the game together on Saturday.

- **BART-cnn/dm**:
Morocco beat Spain. It's the first time Morocco made it to the quarter-finals. Next game is against Portugal on Saturday. Omar and Leila will watch the game together. They will bring a giant Moroccan flag. If they win, Morocco will be the first African nation to **reach** the semi-finals of the **tournament**.

- **PEGASUS-cnn/dm**:
It's the first time Morocco ever made it to the quarter-finals. Next game is against Portugal on Saturday. Leila will bring her giant Moroccan flag.

- The third test dialogue:

This conversation is similar to the previous one except that is a series of question/answer pairs to verify the model ability to concatenate the question with its respective answer, and summarize the concluded idea. The question-answer pairs in this example are:

- Which team are we playing against?
Portugal.
- When?
Saturday at 7am.
- Wanna watch the game together?
Absolutely!

Leila: Hi Omar, Congratulations! Morocco beat Spain. The Moroccan goalkeeper is great!
Omar: Yes, I'm very proud of Bounou and the rest of the team. It's the first time Morocco ever made it to the quarter-finals.
Leila: This is a great achievement. I'm over the moon. Which team are we playing against next?
Omar: Portugal.
Leila: When?
Omar: Saturday at 7am.
Leila: If they win, Morocco will be the first African Nation to ever make it to the semi-finals.
Omar: Oh my God, I'm so excited. Wanna watch the game together?
Leila: Absolutely! I'll bring my giant Moroccan flag.
Omar: That would be dope!

The generated summaries are as follow:

- **T5-base**:
Omar and Leila will watch Morocco's game against Portugal at 7am.

- **BART-base**:
Morocco beat Spain. Bounou and the rest of the team made it to the quarter-finals. Leila and Omar will watch the game together.

- **PEGASUS-base**:
Omar and Leila will watch the game together on Saturday at 7am.

- **BART-xsum**:
Morocco beat Spain to reach the quarter-finals of the **European Championship**. They will play Portugal on Saturday at 7 am. Leila will bring her giant Moroccan flag..

- **PEGASUS-xsum**:
It's the first time Morocco has made it to the quarter-finals **of a major tournament**.

- **BART-cnn/dm**:
Moroccan team beat Spain in the quarter-finals. They are playing Portugal on Saturday at 7 am. Omar and Leila will watch the game together. Leila will bring her giant Moroccan flag. If they win, Morocco will be the first African nation to make it to the semi-final.

- **PEGASUS-cnn/dm**:
It's the first time Morocco ever made it to the quarter-finals **of the World Cup**. They will play Portugal on Saturday at 7 am. Leila will bring her giant Moroccan flag.

Results are discussed in the section below.

4.0.4 Conclusion & Future Work

- Conclusion:

The quantitative scores were all in favor of **BART-xsum**, but generated summaries that are too short to convey enough information. The qualitative results of the implemented models are summarized as follow:

- The **base** models were able to generate correct information, except that the summaries were short, and did not contain the entire scope of the conversation.

- **BART-base** failed in some occasions to complete its sentences.

- The models pretrained on the **xsum** dataset did not necessarily generate better summaries that the **base** models. Their summaries also consisted of two sentences. This could easily be attributed to the short summaries of the **xsum** observations.

- On few occasions, **PEGASUS-base**, **BART-cnn/dm**, **BART-xsum** outputted information was either inacurrate, or fabricated, as is highlighted in **red**.

- As expected, the **cnn/dm** models outputted summaries consisting of more sentences than their counterparts. This could intuitively be explained by the long summaries of the **cnn/dm** observations.

- On few occasions, the **xsum** and **cnn/dm** models showcased abstractive summarization behaviour by introduced words that were not specifically mentioned in the subject dialogue (i.e. victory, reach, tournament, major tournament, Moroccan team, World Cup) as is highlighted in **blue**.

- Beside false data generation, the **cnn/dm** model summaries proved to be capable of understanding the conversation topic, extracting valuable information from question-answer pairs, using unseen words in the source text (abstractive behavior), being of adequate length to communicate more information than the strict minimum.

- Future Work:

To tackle the technical limitations of the results, a number of future improvements could be added:

- We could combine **w**ith other dialogue-summary datasets from different fields (i.e. finance, sports, medecine) to allow the fine-tuned model to generalize well on a larger array of conversation topics.

- We could implement different models such as ASGARD [27]. As a matter of fact, Zhang et al [27] clearly state in their abstract, that sequence-to-sequence models "commonly suffer from fabricated content, and are often found to be near-extractive". Our results do corroborate this statement as a number of the implemented models in this study generated fabricated data, and only showcased a mild abstractive behavior.

References

[1] Derek Dean and Caroline Webb. Recovering from information overload. *McKinsey Quarterly*, 27, 2011.

[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. Text summarization techniques: A brief survey. 2017.

[3] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Hegel: Hypergraph transformer for long document summarization. 2022.

[4] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. 2016.

[5] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. Assessing sentence scoring techniques for extractive text summarization.

[6] Pei-ying Zhang and Cun-he Li. Automatic text summarization based on sentences clustering and extraction. pages 167–170, 2009.

[7] K. Kaikhah. Automatic text summarization with neural networks, 2004.

[8] Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. page 20–es, 2004.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019. URL `https://arxiv.org/abs/1910.10683`.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2019. doi: 10.48550/ARXIV.1910.13461.

[11] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. 2019.

[12] URL `https://commoncrawl.org/`.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL `https://arxiv.org/abs/1706.03762`.

[14] URL `https://huggingface.co/datasets/cnn_dailymail`.

[15] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation, 2019. URL `https://arxiv.org/abs/1905.03197`.

[16] URL `https://huggingface.co/datasets/samsum`.

[17] URL `https://huggingface.co/datasets/xsum`.

[18] URL `https://huggingface.co/google/t5-v1_1-base`.

[19] . URL `https://huggingface.co/facebook/bart-base`.

[20] . URL `https://huggingface.co/google/pegasus-x-base`.

[21] . URL `https://huggingface.co/facebook/bart-large-xsum`.

[22] . URL `https://huggingface.co/google/pegasus-xsum`.

[23] . URL `https://huggingface.co/facebook/bart-large-cnn`.

[24] . URL `https://huggingface.co/google/pegasus-cnn_dailymail`.

[25] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. July 2004.

[26] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.449. URL `https://aclanthology.org/2021.findings-acl.449`.

[27] Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. July 2020.