A Study of Text Summarization Approaches: The Proper Way

Arisa Nguyen, KT Norton

April 2023

Abstract

While general large language models (LLMs) have taken the world by storm in recent years, natural language processing for specific tasks remains an important topic of research. Text summarization as a standalone natural language processing task is imperative for certain applications that aim to address the challenge of information overload. The “big data revolution” left companies with massive lakes of data; some of which remain untapped resources because they simply cannot be processed in their current unstructured states. Text summarization can help tackle this problem by providing a concise synopsis of the original long form text. One example that might be of particular interest for summarizing news articles, legal documents, or business reports is the task of summarizing text documents containing large amounts of proper nouns, such as the names of companies or organizations. These can be especially difficult for some text summarization approaches, which we will explore in this research.

Introduction

As large language models (LLMs) have become industry standard for natural language processing, they have made large advancements in text summarization and the ability to extract the most important information from a piece of text. However, even small pieces of missing or erroneous information can be extremely dangerous, as we’ve seen recently with OpenAI’s ChatGPT falsely accusing a law professor of sexual harassment.[1] Therefore, it’s imperative to understand the nuances of different text summarization approaches, as they relate to noun phrases or proper nouns, when choosing a model for a text summarization application.

Our goal for this project is to compare extractive and abstractive text summarization approaches to determine their strengths and weaknesses, particularly in the case of text documents containing noun phrases that the models have likely not seen during training or pre-training. Extractive text summarization approaches extract summaries from the original text, while abstractive text summarization aims to interpret the input text and generate a summary using new words and text. While extractive summarization is typically simpler and requires less computational cost to implement, it can often miss conceptual connections critical to the summarization of a body of text. Abstractive summarization is often more concise and fluent but more costly and is prone to hallucination[2].

We hypothesize that extractive text summarization approaches will yield better results than abstractive approaches because they are likely to extract the correct proper nouns from the text, while abstractive approaches might struggle to come up with the correct ones.

[1] "ChatGPT Falsely Accuses Jonathan Turley of Sexual Harassment, Concocts Fake WaPo Story to Support Allegation."

[2] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald

Related Work

Most of the research surrounding proper nouns in the natural language processing world has been conducted using Named Entity Recognition (NER),[3] including the improvements introduced by deep learning in recent years.[4] However, little research has been conducted to explore the impact of proper nouns on general LLMs or other natural language processing tasks, particularly text summarization. Our research aims to examine the impacts of proper nouns, particularly company or organization names, on a few state of the art text summarization models.

Methodology

We evaluated the summarization models on the abstractive dataset XLSum, consisting of 1.35 million article-summary pairs from BBC news. Of the 1.35 million pairs, 329,592 are in English.[5] In addition to being a large dataset with high quality summaries (as determined by human evaluation)[6], none of the models we used have seen XLSum before and therefore do not have an unfair advantage in our evaluations.

We chose the following models to use in our analysis:

- **Lead-3:** This is the first three sentences of the article. It is our baseline and also considered one of our extractive models. BART also used this model as their baseline.[7]
- **BERT Extractive:** This is an extractive model that uses BERT embeddings and K-means clustering to extract the sentences closest to the centroid for summarization. While the BERT embeddings in the BERT Extractive model are trainable, the model as a whole is not a transformer model and can't be trained.[8]
- **BART:** This is an abstractive model that is similar to BERT in its pre-training tasks and overall architecture but focuses on text generation. It utilizes a denoising autoencoder which essentially improves the model's flexibility for sentence length and ability to handle longer range transformations to the input.[9]
- **T5:** This is an abstractive model which utilizes a transformer model architecture that treats every task as a text-to-text problem. It is an encoder-decoder model pre-trained using both unsupervised and supervised techniques on tasks such as translation, summarization, question answering, and classification.[10]

With Lead-3 and BERT Extractive not being trainable on our dataset, this left us with a total of 6 models in our analysis: Lead-3 (as our baseline), BERT extractive, BART trained, T5 base

[3] L. F. Rau
[4] Li, Jing, et al.
[5] Hasan, Tahmid, et al
[6] Hasan, Tahmid, et al
[7] Lewis, Mike, et al.
[8] Miller, Derek.
[9] Lewis, Mike, et al.
[10] Raffel, Colin, et al.

(untrained), and T5 trained. The trained models refer to being trained with our XLSum training set. The T5 base model acts as another baseline for its trained counterpart, which has already been pre-trained on the summarization task. .

The metrics we used to evaluate our models are ROUGE-1, ROUGE-2, ROUGE-L, and CHRF. While ROUGE is a standard metric, CHRF is a score more recently developed to close the gap between human and machine-generated text evaluations. CHRF uses character n-gram[11], while ROUGE uses word n-gram to compare candidates to references. This allows CHRF to better compare “morphological variants of words”.[12] A 2020 study found that CHRF scores had stronger correlation with Direct Assessment (made by a human) and less errors than ROUGE and BLEU when evaluating machine translations between English and German.[13]

When training BART and T5 with our train set, the maximum article length and maximum summary length were 512 and 256 tokens, respectively, to decrease training time and prevent out-of-memory errors. These are the same lengths used by PEGASUS during pre-training.[14] All other models and their hyperparameters, including both training-specific (eg. epochs) and task-specific (eg. number of beam groups), were manually fine-tuned by choosing the parameters that resulted in the highest average ROUGE and CHRF scores on the validation set.

A total of 1200 random summary-pairs were taken from the English XLSum dataset, then split into train, validation, and test sets of 1000, 100, and 100, respectively. We theorize that trained models will score higher than untrained models across all scoring metrics, because previous research has shown that fine-tuning language models for downstream tasks with as little as 100 training examples dramatically improves model performance.[15] We also theorize that extractive models will have higher ROUGE scores and lower CHRF scores than abstractive models, because the extractive models will use words and phrases in the article, while abstractive models may abstract away from using exact words and phasing from the article.

Because we are particularly interested in how well our models generate summaries with proper nouns, we also created another sample of 100 manually-identified pairs that contained proper nouns, particularly company or organization names, in both the article and summary. We applied the fine-tuned models from the previous step to these 100 pairs to generate candidates. This allows us to investigate how well our models generate summaries when proper nouns are vital to the summarization.

We hypothesize that extractive models will have higher ROUGE and CHRF scores for the proper nouns dataset than the random dataset, because it would identify the importance of proper nouns and use phrases containing the proper noun from the article in its generated

[11] Popović, Maja.
[12] Mathur, Nitika, Timothy Baldwin, and Trevor Cohn.
[13] Mathur, Nitika, Timothy Baldwin, and Trevor Cohn.
[14] Zhang, Jingqing, et al.
[15] Zhang, Jingqing, et al.

summary. We also hypothesize the abstractive models will have lower ROUGE and CHRF scores on the proper nouns dataset than on the random dataset, because it may abstract too far from the reference.

Results

The tables below show the average scores for each model on the test sets.

Model Scores on Randomly Sampled Test Set

	Model	ROUGE-1	ROUGE-2	ROUGE-L	CHRF	Mean Candidate Word Length
Extractive	**Baseline**	0.191	0.029	0.127	27.0	59
	BERT Extractive	0.186	0.024	0.127	26.3	45
Abstractive	**BART**	0.293	0.094	0.235	27.4	19
	T5 Base	0.207	0.040	0.145	27.4	43
	T5	0.219	0.046	0.156	22.5	20

Models Scores on Articles Containing Proper Nouns

	Model	ROUGE-1	ROUGE-2	ROUGE-L	CHRF
Extractive	**Baseline**	0.188	0.022	0.120	26.7
	BERT Extractive	0.184	0.021	0.124	26.1
Abstractive	**BART**	0.311	0.095	0.246	28.6
	T5 Base	0.200	0.028	0.133	26.6
	T5	0.220	0.040	0.164	22.8

Based on the results of both test sets, the abstractive text summarization approaches outperform the extractive approaches. This is especially interesting for the T5 base mode, which was not trained on this dataset. This indicates that T5 generalizes well for summaries with proper nouns and could be a good model for applications that require text summarization on text containing proper nouns but do not have enough data to train the model further.

The BART trained model achieved the highest scores for all metrics on both test sets. The T5 trained model achieved higher scores for ROUGE overall, but ROUGE-2 scores increased only about half as much as ROUGE-1 scores when compared to T5 base and yielded considerably lower CHRF scores on both test sets. This indicates that the T5 trained model does well at generating summaries that are similar to the reference but are less morpho-syntactically similar, meaning they could be less grammatically similar or use different word order or word phrasing. This is expected from abstractive text summarization approaches. In the example below, we see a reference summary along with candidate summaries from the baseline and the BART trained model. BART's candidate actually captures the context of the summary and sounds more fluent and concise than the baseline. However, it has less similar wording to the reference summary and therefore has a lower CHRF score, while maintaining a higher ROUGE score.

Reference	Baseline	BART Candidate
Performances at the Latitude Festival were suspended for about an hour after lightning and heavy rain caused the power to be turned off.	Organisers of the event at Henham Park, Suffolk, said performances were paused on the Obelisk Arena Stage, the BBC Sounds Stage and the Lake stage, as a precaution. They were waiting for "inclement weather to pass" they said on Twitter. "Lightning is the concern, not rain" it said after people complained of delays.	A festival in Suffolk has been shut down for more than an hour due to lightning.

The results also show that the T5 trained model and BART trained model produced shorter summarization candidates than the other models' candidates, yielding less characters to potentially overlap with the reference summary. This could explain why the T5 trained model has lower CHRF scores and also makes the BART trained model's high average CHRF scores that much more significant.

See Appendix II for an example of a target summary with each model's candidate.

Conclusion

The treatment of proper nouns in text summarization approaches is an area that requires further investigation and research. The experiments in this research indicate that neither abstractive nor extractive text summarization approaches as a whole outperform the other. However, given the results of the experiments, it is clear that BART trained on a target dataset will yield the best results for text summarization applications where the target dataset is summarized using proper nouns. Additionally, if data limitations are a concern, the T5 base model is a great prospect as an off the shelf model.

Future work should explore the impact of fine-tuning and training models directly on the articles with reference summaries containing proper nouns, as we suspect that this may enhance model results. Additionally, future work should explore the accuracy of the content of the model candidates to examine how well the models' candidates summarize the original text, rather than being scored against a specific reference summary.

Appendix I - Sources

Related Work

- "ChatGPT Falsely Accuses Jonathan Turley of Sexual Harassment, Concocts Fake WaPo Story to Support Allegation." *GW Law*, 10 April 2023, https://www.law.gwu.edu/chatgpt-falsely-accuses-jonathan-turley-sexual-harassment-concocts-fake-wapo-story-support.
- Hasan, Tahmid, et al. "XL-sum: Large-scale multilingual abstractive summarization for 44 languages." arXiv preprint arXiv:2106.13822 (2021).
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, Online. Association for Computational Linguistics.
- L. F. Rau, "Extracting company names from text," *[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, Miami Beach, FL, USA, 1991, pp. 29-32, doi: 10.1109/CAIA.1991.120841.
- Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).
- Li, Jing, et al. "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34.1 (2020): 50-70.
- Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. "Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics." arXiv preprint arXiv:2006.06264 (2020).
- Miller, Derek. "Leveraging BERT for extractive text summarization on lectures." arXiv preprint arXiv:1906.04165 (2019).
- Popović, Maja. "chrF: character n-gram F-score for automatic MT evaluation." Proceedings of the tenth workshop on statistical machine translation. 2015.
- Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.
- Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." International Conference on Machine Learning. PMLR, 2020.

Repositories

- GitHub
  - https://github.com/arisanguyen/w266_project/tree/master
- HuggingFace BART Fine-tuned Model
  - https://huggingface.co/arisanguyen/finetuned-BART-all-categories/tree/main
- HuggingFace T5 Fine-tuned Model
  - https://huggingface.co/arisanguyen/finetuned_T5_all_categories/tree/main
- HuggingFace Transformers Repository
  - https://github.com/huggingface/transformers/tree/main/examples/pytorch/summarization
- HuggingFace XLSum Dataset
  - https://huggingface.co/datasets/csebuetnlp/xlsum

Appendix II - Model Candidates Example

Model	**Candidate**
Target	Celebrations have been held to mark the 200th anniversary of the opening of the Monmouthshire and Brecon Canals.
Baseline	Enthusiasts recreated the original opening of the waterway in 1812 at Pontymoile Basin with church bells sounding along the route. The 49-mile, canal built to transport coal and iron, winds its way south from Brecon to Cwmbran, Newport and Cwmcarn. Today it is popular with tourists on canal boat holidays.
BERT Extractive Summarizer	The 49-mile, canal built to transport coal and iron, winds its way south from Brecon to Cwmbran, Newport and Cwmcarn. There are a lot of canal companies that hire weekly holiday trips for visitors.
Trained BART	The anniversary of the opening of the Monmouthshire Canal has been marked by a number of celebrations.
T5 Base	The original opening of the waterway in 1812 was recreated with church bells . the 49-mile, canal built to transport coal and iron, winds its way south from Brecon to Cwmbran . today it is popular with tourists on canal boat holidays .
Trained T5	The canal is credited with kick-starting the Industrial Revolution in towns near its route.