An analysis on different Text Summarization techniques with a focus on Yoda-style translations

Claudio Castorina
Niko Covic
Emmanuel Fernandez
Samuel Goldie
Egor Sementul

Maastricht University - Department of Advanced Computing Sciences

June 2023

Contents

Abstract

This paper focuses on exploring different methods for text summarization, with a focus on fine-tuning approaches for Large Language Models (LLMs) and a Graph-based model for comparison. Our study introduces an interesting twist by attempting to generate summaries in the style of the fictional character *Yoda* form *Star Wars*.
We analyse the difference in performance of pre-trained LLMs, namely T5, BART and Pegasus for this task and evaluate their summarization capabilities on the *Multi-News* dataset. We explore a *rule-based* approach and a *fine-tuning* approaches for translation in order to make the Yoda summaries.

The results demonstrate the effectiveness of fine-tuned models for both text summarization and Yoda-style translation. The paper concludes by suggesting future research directions, including the investigation of larger models for text summarization and Yoda-style translation.

1 Introduction

In recent times, due to its wide wide range of application and its capabilities in understanding human language, Natural Language Processing (NLP) has become increasingly important. One significant application of NLP is *text summarization*, as it plays a center role in extracting the most important and relevant information from large volumes of text. Another popular application relies on Natural Language Generation (NLG), which focuses on generating text that follows a specific pattern or mimics a particular individual.

In this paper, our goal is to examine different methods used for text summarization, specifically focusing on LLMs fine-tuning approaches for *news summaries*, and using graph-based approaches for comparison. Furthermore, we want to add an interesting twist to our study by attempting to generate summaries in the style of the famous fictional character from Star Wars, Yoda. As Yoda is known for reversing the order of sentences, this poses a fascinating challenge in the field of natural language processing. To address this, we will explore both a manually crafted method and the fine-tuning of a language model.

2 Related work

2.1 T5

The T5 transformer is an encoder-decoder model created to explore the possibilities and power of transfer learning [7]. It was pre-trained on a mix of supervised and unsupervised tasks converted into a text-to-text format. It is known to work well on a variety of NLP tasks, among which is summarization, which is the main focus of this report.

2.2 Pegasus

Pegasus is a *state-of-the-art* LLM developed by Google in 2018. It is a Transformer-based encoder-decoder model for abstractive text summarization [8]. Pegasus relies on *abstractive text summarization*, which is a technique that more closely resembles what humans do. Compared to *extractive* summarization, abstractive is generally more complex as it requires the model to understand the text and, based on that, generate new text to explain it [8].

2.3 BART

Bart is neural network based on a transformer-architecture developed by Facebook and it consists of a bidirectional encoder with a left-to-right decoder. It was presented as a *denoising* autoencoder, and its uses case was primarily aimed for pretraining sequence-to-sequence models [4].

2.4 Graph-based models

Graph-based models have emerged as a prominent approach for text summarisation, leveraging the concept of reductions to generate concise summaries. Graph-based models construct a representation of the input text as a graph, where sentences or phrases are nodes, and edges represent relationships between them. By applying graph-based algorithms, these models identify the most important nodes in the graph and generate summaries based on their connections and significance. A notable contribution in the field of graph-based summarization is the LexRank algorithm[2]: a stochastic graph-based method for computing relative importance of textual units.

2.5 Part of speech tagging

Part-of-speech (POS) tagging, which involves assigning grammatical categories to words in a text, has been extensively studied in the field of natural language processing (NLP). Over the years, various approaches, including rule-based systems and machine learning techniques, have been developed to tackle this task. With the advancements in NLP, POS tagging is considered a well-established and largely solved task.

3 Dataset

For fine-tuning pretrained models described above the *Multi-News* dataset was chosen. *Multi-News* is the first large scale *multi-document summarization* (MDS) dataset speficically designed for news articles, and it consists of source text and human-written summaries of these articles [3]. Each summary is written by professional editors and includes links to the original articles cited.
The two features of the dataset are:

- document: the original news article
- summary: corresponding summary

The dataset also includes a wide range of sources, thus it should allow for the summarization of diverse news articles efficiently as reported by the authors [3].

4 Methodology

4.1 Summarization with encoder-decoder models

In this research, we decided to explore the power of fine-tuning (i.e. transfer learning) as well as compare multiple different pre-trained models with each other. In order to match memory and time limitations, we decided to use the pre-trained based models of T5 [7] and BART [4].

Before we could compare these models, it was necessary to define a metric to compare them with. We settled with the ROUGE score metric, which is a popular metric used for text summarization tasks [5]. This metric requires us to have human main summaries of the data. It computes a score by counting the number of N-grams that occur both in the predicted summary, and in the human made summary. After doing so, a precision value is computed as follows

$$p = N/P$$

where N is the number of matching N-grams and P is the length of the predicted summary. Then, a recall value is computed as:

$$r = N/H$$

where H is the length of the human maid summary. Finally, the two scores are then combined to obtain an F-score. There are multiple variants, of the ROUGE score, such as ROUGE-1, ROUGE-2 etc, in which the main difference is the size of the N-grams being compared. For our purposes, we settled for the ROUGE-L score which finds the longest matching N-grams automatically, and counts them.

Now that we have defined a metric to measure the quality of summaries, we can define how we will compare the different models. Firstly, each of the base pre-trained models (T5 and BART) were evaluated on the test set taken from the *Multi-News* dataset. Secondly, fine-tuned the base BART and T5 models on the train set taken from the *Multi-News* dataset. The training parameters can be found in Table 2 of the appendix. Finally, an already fine-tuned BART and T5 models on the CNN-daily-news dataset were additionally used to compare the performance. For each of these models, the ROUGE precision, recall and F-score were saved for each of the instances in the test set.

4.2 Rule-based Yoda approach

At a baseline, the grammar used by Yoda can be identified in the structure *object-verb-subject* (OVS), in contrast to languages that mainly follow the *subject-verb-object* (SVO) structure. However Yoda doesn't necessarily follow this structure rigorously, and in many cases there are sentence specific adjustments that take place - mainly his misuse of auxiliary verbs at the end of sentences (e.g. "we arrived" → "arrived we **have**). In our model, we employ part-of-speech

(POS) tagging to analyze the syntactic structure of the input text. The procedure involves segmenting the text into sentences, further breaking them down into clauses, and applying Yoda's distinctive reordering of clauses. This approach allows us to transform the input text into Yoda's characteristic speech pattern, creating Yoda-like sentences while preserving the original meaning and content.

4.3 Fine-tuning models for Yoda syntax

Our aim for this part was to fine-tune pre-trained language models, to translate natural English sentences to their equivalent as if Yoda was speaking. The issue we faced was that there aren't any widely available datasets compatible to this task: because of this, we decided to use our previously rule-based Yoda method, and translate sentences from a dataset composed of simple English sentences[1].

In the beginning we thought that this would probably fine-tune the models to perform exactly like our method: however, we hypothesised that the LLM could pick up on some of the structure and interpolate that to harder sentences, using its pretrained semantic understanding. We decided to fine-tune Pegasus[8] and T5[7], to see if by learning how to translate from English to Yoda, they could potentially directly summarize directly into Yoda syntax.

For preprocessing, we applied several steps to prepare the found sentences dataset [1]. Initially, we filtered out sentences longer than 50 characters, since it can be observed that Yoda speaks in small sentences. We then looped through all these shorter sentences and computed their Yoda translations as supervised data. We then appended a task prefix to all the inputs, in order to let the models recognize when they are asked to translate to Yoda.

Following the preprocessing and train-test split, we proceeded to fine-tune the models based on the provided training parameters (as detailed in the appendix table 2). In order to test the performance of our models, ROUGE-1 and ROUGE-2 wouldn't be a good fit as Yoda translations only switch the order of the sentence, but generally keeps the same word. We found that the BLEU score[6] can capture the extent to which the generated translation successfully captures the essence of Yoda's speech pattern, as it evaluates the overall translation quality rather than relying solely on n-gram matching, using a technique known as Modified n-gram precision. Given that, we acknowledge that ROUGE-L might capture some aspects of correctness, but empirical tests show that, for very easy sentences, ROUGE-L tends to give a very high score, as the longest subsequence is generally short. This makes it revert to a ROUGE-N model where N is not large enough, and therefore would not be a good fit.

5 Experiments and Results

5.1 Summarization using graph-based models

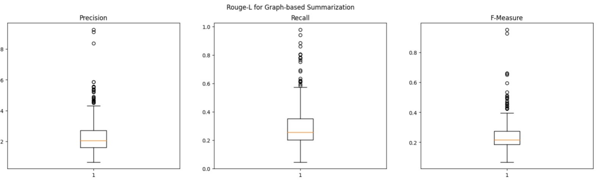

Figure 1: ROUGE score for Graph reduction models

As figure 1 shows, the graph reduction model provides a very decent and quite consistent baseline for summarization. Here, precision and recall seem to be non-significantly different, resulting in a near-harmonic balance.

5.2 Summarization using encoder-decoder models

As seen in Figure 2, the fine-tuning seems to have increased the performance of the BART model in all three ROUGE metrics, compared to the base model. When it comes to the fine-tuned model on the CNN-Daily-Mail dataset however, the precision

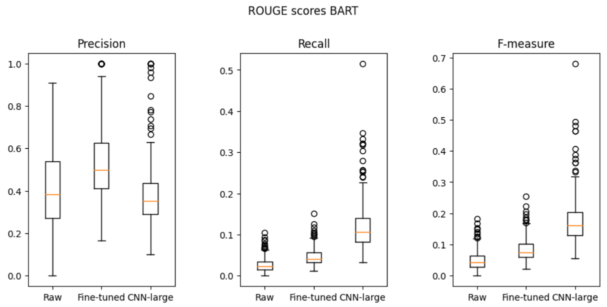

Figure 2: ROUGE score for BART models

of the model seems to have dropped, while the recall and F-score seem to have become substantially higher.

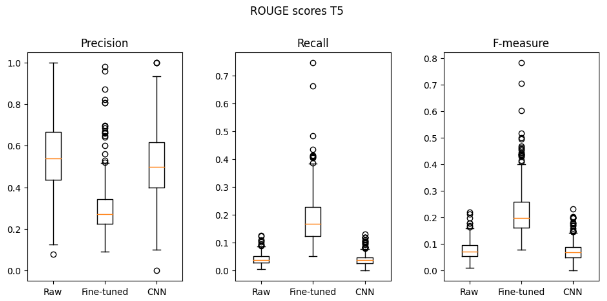

Figure 3: ROUGE score for T5 models

In figure 4, we observe that the highest performing model is the fine-tuned model in all three aspects of the ROUGE metric, except for precision, by a large margin. The model fine-tuned to the CNN-Daily-News dataset seems to have not even outperformed the base T5 model.

5.3 Fine-tuned models for Yoda syntax

As we can see from Table1 both fine-tuned models perform very similarly well in terms of exactly predicting the test dataset, with an exact accuracy of 84%.

In terms of BLEU score, since most values are crammed to the score of 100 in Table 4, the box is not visible, but we can again see that both models have very similar capabilities to predict the test dataset.

Table 1: Average Exact Predictions for Fine-Tuned Models T5 and Pegasus

Model	**Average Exact Value (%)**
Yoda-T5	0.84375
Yoda-Pegasus	0.8333

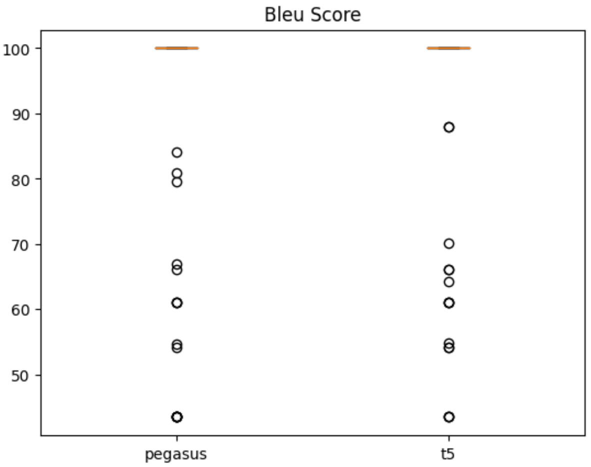

Figure 4: BLEU score for fine-tuned models for Yoda syntax

6 Discussion

6.1 Graph reduction summarization

The results show an impressive baseline, especially considering the non-supervised nature of the technique. As previously mentioned, we can observe a near-harmonic balance, which indicates that the model is performing strongly. Interestingly enough, this method, which was originally intended as a baseline, outperformed most of the Deep Learning methods.

6.2 Summarization using encoder-decoder models

As shown in the results, we can observe that the best-performing BART model is the model fine-tuned on the CNN-Daily-Mail dataset, while the best-performing T5 model is the one fine-tuned on the *Multi-News* dataset. An interesting observation is that the best-performing models always seem to contain the lowest precision values. This is likely because the models learned that the summaries in their respective fine-tuning datasets contain longer summaries, so they ended up producing longer summaries themselves.

Looking at the values of the results shown in Figures 2 and 4, we can observe that the best T5 model seems to have slightly outperformed the best BART model. It is again visible that the precision in the T5 model is lower than that of the BART model, which indicates that the T5 model is closer to producing summaries on appropriate weight.

Additionally, it is worth mentioning a little more context about the CNN-Daily-Mail dataset models. The BART model is the official model trained by Facebook, and uses a BART-large mode, which is larger in comparison to the BART-base model. The T5 model was fine-tuned independently, and uses the model equal in size to the T5-base model. It is therefore likely that the better performance of the BART CNN-Daily-Mail model is not due to the data it was fine-tuned on, but due solely due to the size of the model.

6.3 Fine-tuning models for Yoda syntax

Overall we were able to fine-tune 2 large language models to be able to efficiently translate natural English sentences to their equivalent as if Yoda was speaking. We also tested to see if they could summarize directly into a Yoda syntax, the outputs can be found in the appendix. We did not run experiments on this as we didn't have ground truths for Yoda summaries, or a metric to see how good they are. However we can clearly see as humans that the Yoda-pegasus managed to summarise the text into Yoda speech quite effectively, however, Yoda-t5 wasn't able to keep the reversed syntax.

7 Extension

Our research has laid the groundwork for further investigation and development. To improve efficiency and quality of text summarization and Yoda-style translation, the following extensions are suggested in order to achieve better results.

7.1 Investigation of Larger Models

Future work can explore the application of larger models such as BERT-large, T5-large or even GPT-3 for the task of text summarization. The increased model complexity might contribute to a quality increase in summarization. Also, we recognize that our deep learning models are slightly underfitting, caused by the lack of extensive computation resources.

7.2 Improvement of Yoda-style Sentence Generation

The Yoda-style sentence generation method used in this study has the potential for improvement. Training a sequence-to-sequence model specifically for Yoda-style translations using an augmented dataset could potentially enhance the model's capacity to handle more complex sentences. An interesting avenue for further research would be to investigate why Pegasus, as a pre-trained model, demonstrated a better ability to grasp the underlying logic behind Yoda translation. This analysis could shed light on the strengths and weaknesses of different language models in capturing and reproducing specific linguistic styles.

7.3 Adaptation to Different Text Genres

While our work has been oriented towards news article summarization, future studies can customize our approach for different genres such as scientific articles, novels and legal documents. Based on our Yoda approach result, it's clear that the potential of LLMs has yet to be fully explored.

References

[1] Genericskb: A knowledge base of generic statements. Allen Institute for AI, 2020.

[2] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, dec 2004.

[3] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] 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.

[5] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[7] 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, 2020.

[8] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

8 Appendix

Table 2: Training parameter for fine-tuning models for yoda translations

Parameter	Value
output_dir	’./output’
num_train_epochs	5
per_device_train_batch_size	8
per_device_eval_batch_size	8
warmup_steps	500
weight_decay	0.01
logging_dir	’./logs’
logging_steps	100
evaluation_strategy	’steps’
eval_steps	500
save_strategy	’steps’
save_steps	500
learning_rate	1e-4
overwrite_output_dir	True
predict_with_generate	True
load_best_model_at_end	True
height	

Table 3: Training parameter for fine-tuning models for summarization

Parameter	Value
num_train_epochs	5
max_len	1024
summary_len	batch_size
EPOCHS	
shuffle	True
num_workers	0
height	

Original Text for Yoda translations

Donald Trump, the 45th President of the United States, is a polarizing figure who has left an indelible mark on American politics. Known for his larger-than-life personality, Trump’s presidency was characterized by controversial policies, fiery rhetoric, and a penchant for unconventional communication through social media.
During his time in office, Trump pursued an ”America First” agenda, aiming to prioritize the interests of the United States in areas such as trade, immigration, and foreign policy. His administration implemented significant tax cuts, deregulation measures, and pursued a more assertive stance on international trade agreements.

Trump's approach to governance often drew both fervent support and vehement opposition. Supporters praised his efforts to revitalize the economy, prioritize national security, and challenge traditional political norms. Critics, on the other hand, raised concerns about his handling of sensitive issues, including immigration, climate change, and racial tensions.
Beyond policy, Trump's leadership style and unfiltered communication drew considerable attention. His prolific use of Twitter became a defining characteristic of his presidency, allowing him to directly communicate with his base and express his thoughts, often generating controversy and media frenzy.
Trump's presidency was not without challenges and controversies, including investigations into Russian interference in the 2016 election and subsequent impeachment proceedings. These events further deepened the divisions within the country and fueled passionate debates about the state of democracy and the role of the presidency.

Yoda summary by yoda-pegasus fine-tuned model

A divisive figure who has left an indelible mark on American politics, Donald Trump is.

Yoda summary by yoda-t5 fine-tuned model

Donald Trump, the 45th President of the united states, is a polarizing figure who has left an indelible mark on american politics.

Original Text for summarization (best models)

Stallone Superfans We Dropped $445 To Pose With Our Hero Sly Stallon. Sly was one of the attractions at NYC Comic Con this weekend, where the going rate for a snapshot with the Italian Stallion ran just south of $500. Autographs were $395.But the price didn't scare off legions of Stallone-aholics, who willingly shelled out the cash for a moment with their hero.Stallone wasn't the only celeb charging for pics. A photo withran $400 bucks, and people dropped $200 for a shot withThe moral of the story – nerds got bank, yo. Yesterday afternoon, Sylvester Stallone turned up at New York Comic-Con, partly to promote his upcoming film, Escape Plan, and partly to sell a tiny fragment of his existence for a steep fee. A very steep fee: $395 an autograph, $445 a photo. ""He's flying here straight from Bulgaria,"" a woman selling tickets for Stallone's signature Saturday afternoon kindly explained. ""This is a very limited opportunity."" At 3:30 pm, she said there were only 20 tickets left unsold, though she couldn't say out of how many had been available. Stallone's appearance was orchestrated by Celebrity Authentics, a memorabilia company that also brought Sigourney Weaver and Carrie-reboot star Chlo¡unk¿ Grace Moretz to the convention. A breakdown of Stallone's fees: Meanwhile, an in-person moment with Alien icon Sigourney Weaver cost less than half of that. David Duchovny's signature was $80; a photo with him was $95. Gillian Anderson's signature was $60, but it would cost you an additional $20 if you wanted her to scribble something for your cat. By contrast, a sampling of what other nerd-friendly luminaries charged: Pokemon voice actress Veronica Taylor, $10 True Blood's Kristin Bauer, $30 Andrew McCarthy, $40 Joel Grey, $40 Anthony Daniels, C-3PO, $50 William Shatner, $75 Patrick Stewart, $75 As for photos, here's a price list of the talent booked by Comic Con: TMZ has snapshots of some of the lovely people who paid to take photos with Stallone.

Graph-based summary

Sly was one of the attractions at NYC Comic Con this weekend, where the going rate for a snapshot with the Italian Stallion ran just south of $500. By contrast, a sampling of what other nerd-friendly luminaries charged: Pokemon voice actress Veronica Taylor, $10 True Blood's Kristin Bauer, $30 Andrew McCarthy, $40 Joel Grey, $40 Anthony Daniels, C-3PO, $50 William Shatner, $75 Patrick Stewart, $75 As for photos, here's a price list of the talent booked by Comic Con: TMZ has snapshots of some of the lovely people who paid to take photos with Stallone.

T5 (fine-tuned)

A photo with Sly Stallone cost $395, an autograph was $395. A woman selling tickets for Stallone's signature said she had only 20 tickets left unsold at the time of the event. Then there were those who paid more than $500 to pose with their hero. And that's not all. There are plenty of other celebrities who have been booked by Comic Con. Here's a price list of some of the stars who took photos with Stallone. ”He's flying here straight from Bulgaria,” said a woman selling tickets for Stallone's autograph. ”This is a very limited opportunity,” she said. At 3:30 pm, there were only 20 tickets left unsold, though she couldn't say out of how many available.