Summarization of Resumes Using BART

Jenifer Lizethe Leiva Martín
Astrid Carolina Melo Guayacan
Juanita Palacios
Engineering Faculty
Universidad Militar Nueva Granada

***Abstract*—The aim of this paper is to use the pretrained BART model for both inference and finetuning in the Natural Language processing task of text summarization using resumes from a Kaggle dataset, with the purpose of making the model generate summaries more tailored to the context of resumes.**

I. Introduction

Today, technological advances allow many resumes to be sent to companies through digital media, in consequence, the time necessary for humans to extract useful information from these documents increases. Text summarization works as a solution to address this problem, since as a task of Natural Language Processing (NLP), it allows to obtain key information through coherent and concise summaries from large documents without human intervention.

Abstractive text summarization is the preferred type for this task, since it, contrary to extractive text summarization, generates new sentences and can use different words to generate the summary, which means that catch better the ideas of the text and is more similar to what a human would do.

In that sense, this project aims to use transfer learning to leverage the capabilities of the pre-trained transformer model BART from facebook (Meta) for the task of text summarization of resumes from a kaggle dataset.

II. Related work

Several research papers have explored the task of automated text summarization using pre-trained models based on transformer architecture.

Dharrao et al. (2024) [2] studied the abstractive summarization performance of three cutting edge models: BART, T5 and PEGASUS, on news business items from a BBC News articles dataset. They concluded according to the Rouge metric and overlap of n-grams that BART is a good choice when the goal is for the summary to be faithful to the original content. While T5 better preserves vocabulary and phrasing alignment according to Meteor score.

Mercan et al. (2023) [1] specifically addressed this task for resumes. In this study, the performance of pre-trained models: T5, Pegasus, BART and BART-LARGE, were evaluated on four open-source datasets of News and reviews of products, and a prepared resume dataset by the authors composed of 75 documents with information of skills, education, experience and personal information. They found that the BART-Large model fine-tuned with the resume dataset had the best performance.

III. Materials and Methods

A. Dataset characteristics

In this project, the Resume dataset is used, taken from Kaggle [5], which compiles more than 2,000 resumes from the livecareer website and organizes them into defined job categories in PDF format, the complementary CSV file includes:

- **I**D: Unique identifier and file name of the corresponding PDF.
- **R**esume str: Contains the text of the resume (plain text only).
- **R**esume html: Contains the HTML version of the resume, as obtained during web scraping.
- **C**ategory: Indicates the job category to which that resume belongs.

The data was obtained by web scraping individual resumes on livecareer.com. The scraping code is available in the original author's GitHub repository. The job categories included are: HR, Designer, Information-Technology, Teacher, Advocate, Business-Development, Healthcare, Fitness, Agriculture, BPO, Sales, Consultant, Digital-Media, Automobile, Chef, Finance, Apparel, Engineering, Accountant, Construction, Public-Relations, Banking, Arts and Aviation.

In this project, 2843 resumes were used in PDF format, the .zip dataset was downloaded from Kaggle and loaded in Google Colab.

B. Preprocessing

Before training the chosen model, a pre-processing process was applied over the dataset. Where first, the text was extracted from the PDF documents for create a .csv with a dataframe that contains as columns the filename, category and text of the resumes.

It also was applied the tokenizer of the model BART, which prepares the inputs by splitting text into tokens and

representing these with numerical IDs from the vocabulary, in order to ensure that the model can process the resumes. Where it was settled that the maximum number of tokens for the inputs would be 1024, which is the limit of the model by default, and that if the text is longer that that, it will be trimmed.

C. Model used

For the NLP task of summarization of resumes, BART (Bidirectional and Auto-Regressive Transformer) was chosen. BART, which was developed by Facebook (Meta) in 2019, is a pre-trained sequence to sequence, Encoder-Decoder transformer model that combines the strengths of the pretrained models BERT (Bidirectional Encoder Representations from Transformers) by Google and GPT (Generative Pre-trained Transformer) by OpenAI. BART is pretrained on english language and fine tuned on CNN/Daliy Mail dataset, which is composed of a collection of news articles and their respective multi-sentence summaries, both written by the original news authors. For pre-train the model, text is corrupted in different ways such as adding noise, missing words, etc, with the purpose that the model learns how to fix it. [4]

In the project, the model was first used in mode evaluation, to generate the summaries from the resumes since the dataset only gave the resumes and the category to which these documents belong. The generated summaries were appended as a column of the data frame, each summary had a maximum length of 150 tokens and the parameters were configured to produce concise summaries limited to a few sentences.

Then, the model was fine tuned with the resumes from kaggle as the inputs and the previous generated summaries as the labels. Where both were tokenized and it was applied padding to avoid problems during the training for different lengths of tokens, reason why an attention mask also was used to indicate the model that must focus its attention on actual tokens.

Finally, the model was used for inference by adding a space that allows uploading resumes PDFs.

```
Generated Summary:
My career of 34 years includes the graphic art field, fine arts, and elementary
art teacher. My personal evolution and vision helps me to know that I possessing
truth, talent and ability that is unique and highly creative. I am a visual
designer with an aptitude for experimental projects. I utilize the basic
elements of design: color, line, shape, space, texture and value. Dedicated art
professional with over 30 years of hands on experience. Proficient in Adobe
Photoshop, InDesign, Illustrator, Microsoft Word, Excel. Website development and
some HTML. Web maintenance. Part-time in Food and Beverage business for 7 years.
Designed souvenir mugs and t-shirts for large scale tourism distribution.
Clients included Sea World and Busch Gardens. Harrah's, Stratosphere, MGM Grand,
and numerous other Las Vegas attractions. I find everything an opportunity to be
solve problems, and am always coming up with creative solutions. Learned
Photoshop and Microsoft Word. Love gardening and growing things, healthful
creative cooking. Live out in the country on 35 acre farm.
```

Fig. 1: Example of a summary generated by the fine-tuned model for a Product and Web Designer's resume

IV. Conclusions

The developed model is able to generate summaries for the imported PDF resumes. However, it is evident that the fine-tuning was not fully effective, as the generated summaries, which were intended to be abstractive, tend to be more extractive, often reproducing entire phrases from the original resumes. Additionally, the default BART token limit of 1024 tokens poses challenges when generating summaries for long PDFs, as they get truncated. Furthermore, it was concluded that the hypothesis of training the model using summaries generated by the model itself is not effective, as it ultimately compromises the model's ability to produce high-quality summaries.

References

[1] Mercan, Ö., Cavsak, S., Deliahmetoglu, A., and Tanberk, S. (2023). *Abstractive Text Summarization for Resumes With Cutting Edge NLP Transformers and LSTM*. arXiv Preprint. https://doi.org/10.48550/arXiv.2306.13315

[2] Dharrao, D., Mishra, M., Kazi, A., Pangavhane, M., Pise, P., and Bongale, A. M. (2024). *Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction*. Research in Artificial Intelligence. https://doi.org/10.18280/ria.380311

[3] Charmy Dafda. (s.f.). *Text summarization using Facebook BART-large-CNN*. Techblog GeekyAnts. https://techblog.geekyants.com/text-summarization-using-facebook-bart-large-cnn

[4] Hugging Face. (2020). *BART — Transformers Documentation*. Hugging Face. https://huggingface.co/docs/transformers/model_doc/bart

[5] Bhawal, S. (2021). *Resume Dataset*. Kaggle. https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset/data