Email Parser for Extracting Meeting Information using LLMs

Somesh Bagadiya[#1]

someshpushpkumar.bagadiya@sjsu.edu

Abstract— **This project develops a custom email parser to automate the extraction of meeting information from emails, primarily using OpenAI's GPT-3.5 model. The system is designed to accurately identify whether an email intends to schedule a meeting and then extract relevant details into a structured format. The parser was tested on a comprehensive dataset that includes synthetic data and real-world emails from the Enron Email Dataset. Results demonstrate the model's effectiveness in streamlining the process of managing meeting information, significantly reducing manual errors and improving productivity. Future enhancements will focus on further improving the accuracy and expanding the dataset to enhance the model's robustness and applicability.**

I. Introduction

A. Problem Statement

In the modern professional landscape, the management of meeting schedules and information largely depends on the interpretation of unstructured email texts. This task, typically performed manually, is not only time-consuming but also prone to human error, often resulting in scheduling conflicts and miscommunications. Recognizing the inefficiency of current practices, this project proposes the development of a machine learning-based email parser. The core aim is to automate the extraction of meeting information from emails, thereby enhancing productivity and reducing the manual effort involved in interpreting these communications. By leveraging advanced machine learning techniques, the project seeks to transform a traditionally arduous manual process into an automated, efficient, and reliable system.

B. Objectives

The primary objectives of this project are outlined as follows:

1. Develop an Automated Email Parser: Design and implement a LLM model that can automatically parse and interpret unstructured email data, replacing manual extraction methods.
2. Intent Analysis for Meeting Detection: Incorporate natural language processing (NLP) techniques to analyze the intent of emails, specifically to determine whether an email is attempting to set up a meeting. This will enable the system to selectively process emails pertinent to meeting scheduling.
3. Extraction of Meeting-Related Details: Once an email is identified as meeting-related, the system will extract critical information such as the time of the meeting, the location or platform where it will occur, and the participants involved. This feature aims to populate calendar events and alert systems automatically, streamlining the process of meeting management.

These objectives strive not only to improve the accuracy and efficiency of extracting meeting information from emails but also to significantly enhance user productivity by automating a previously manual and error-prone process.

II. Data Collection

To effectively train and evaluate the large language model (LLM) based email parser, acquiring a diverse dataset containing emails with detailed meeting information was paramount. This section outlines the methods employed to collect and prepare the data necessary for the project.

A. Synthetic Data Generation

Initially, the project involved creating a synthetic dataset of emails. The primary reasons for generating synthetic data were:

- Initial Model Training: Providing a controlled environment to test

preliminary versions of the LLM-based email parser.

- Data Privacy: Avoiding privacy issues associated with using real email communications.

Procedure:

- Faker Library: Leveraged the Faker() library to generate fake but realistic-looking email content, including random but plausible times for meetings.
- Phrase Construction: Developed a set of 20 phrases each for structured and unstructured meeting invites. These phrases were utilized to populate the body and subject lines of the emails.

Email Classification:

- Structured Emails: Designed to follow a consistent format, these emails are simpler to parse using basic pattern recognition.
- Unstructured Emails: Mimic natural language and vary in format, presenting more complexity and requiring sophisticated parsing techniques.
- Output: Successfully generated a total of 3,000 synthetic emails. This dataset served as the foundational set for initial stages of model training.

Challenges Encountered:

- Limited Diversity: The synthetic data, while useful, failed to capture the full complexity and variability found in real-world email communications.
- Scalability: Generating a sufficiently large dataset for robust training proved to be inefficient.
- Realism: Lacked the nuanced expressions and irregularities present in human-written communications, which are crucial for training a versatile parser.

B. Enron email dataset

To overcome the limitations of synthetic data, the Enron Email Dataset was incorporated into the project. This dataset, comprising over 500,000 emails from Enron Corporation executives, offered several advantages:

- Volume and Variety: The dataset provided a vast amount of data, depicting a wide range of scenarios and email interactions.
- Complexity and Realism: It includes complex email structures and informal language, reflecting true user interactions and thereby enhancing the LLM's ability to generalize to new, unseen emails.
- Impact on Training: The incorporation of real-world data significantly improved the training capabilities and the performance potential of the LLM-based email parser.

Data Preparation:

- HTML Removal: Stripped HTML tags and elements to cleanse the text.
- Field Extraction: Focused on extracting relevant information such as email addresses from 'To', 'Cc', and 'Bcc' fields.
- Column Pruning: Deleted non-relevant columns like 'file' and 'Mime-Version' to streamline the data.
- Email Filtering: Removed emails sent to oneself and applied text preprocessing techniques to ensure the dataset contained only relevant external communications.
- Word-Based Classification: Implemented to manage the large volume of data and prepare it for more effective intent analysis.

Enron Email Dataset Visualizations:

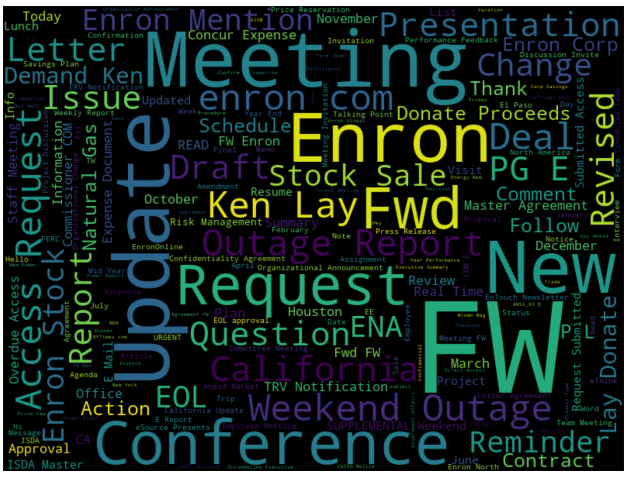

Fig. 1 Word Cloud for Subjects

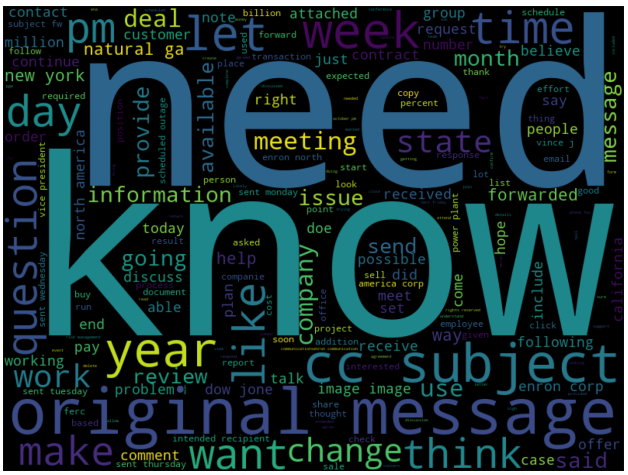

Fig. 2 Word Cloud for Body

The final dataset used for comprehensive model training combined the cleaned Enron emails with the generated synthetic emails, totaling approximately 370,000 emails. This robust dataset was instrumental in developing a highly effective LLM-based email parser capable of accurately interpreting and extracting meeting information.

III. Initial Methodology and Challenges

Model Selection

I fine tuned LLAM2 7B model for intent analysis, a transformer-based model renowned for its language understanding capabilities. Selected for its robust parameter set of 7 billion, this model was deemed ideal for the nuanced task of parsing and interpreting unstructured email content. This model was selected mainly due to its ability to deploy locally.

Fine-Tuning Approach

Fine-tuning the LLAM2 7B model involved several specific modifications to adapt it to the unique requirements of email intent analysis:

- Transfer Learning: Started with a pre-trained model to leverage its existing knowledge base, which significantly reduces the amount of new data required and shortens the training time.
- Data-Specific Tuning: Incorporated domain-specific vocabulary and phrases commonly found in business communications and meeting setups. This helped the model better recognize and interpret relevant terms and contexts.
- Hyperparameter Optimization: Adjusted key hyperparameters such as learning rate, batch size, and number of training epochs to optimize performance. A lower learning rate was tested to fine-tune model weights carefully without losing pre-learned useful features.
- Task-Specific Layers: Added custom layers specifically designed to enhance the model's ability to extract and classify information pertinent to meeting intents, such as datetime entities and participant details.

Setup and Challenges

- Computational Limitations: Local GPU Constraints: Deploying the model on a local NVIDIA GTX 1060 6Gb GPU resulted in frequent out-of-memory errors, indicating inadequate hardware capabilities for handling extensive computational loads required by the model.
- Quantization Efforts: Implemented model quantization by reducing precision from 16-bit to 4-bit to manage resource usage. Although this reduced computational demands, it significantly impaired the model's accuracy and output quality, proving ineffective for the project's needs.
- Cloud Computing Barriers: Transitioned to Google Colab Pro for better hardware access but faced limitations due to the platform's compute unit cap, which was quickly exhausted under the heavy demands of ongoing training and testing cycles.

Technical Adjustments and Conclusion

The fine-tuning efforts highlighted the delicate balance between model complexity and computational efficiency. The initial methodology illuminated the practical

challenges of deploying large-scale NLP models and the need for robust computational resources. This experience necessitated a reevaluation of both the computational strategy and the feasibility of using such a sophisticated model within the constraints of available technologies.

The iterative fine-tuning and the challenges encountered underscored the importance of selecting appropriate models and tuning strategies that align with both the technical objectives and the operational capacities.

IV. Integration of OpenAI API for Intent Recognition

As part of the ongoing development and enhancement of the email parser, the project incorporated the OpenAI API to improve the intent recognition capabilities of the system. This advanced API, based on cutting-edge language models, was utilized to process a significant subset of the email dataset.

API Integration

- Access and Configuration: Secured an OpenAI API key to utilize the GPT-3.5-turbo-0125 model, which provides advanced capabilities for generating and understanding language based on context.

Temperature Settings:

- Low (0.1 - 0.3): Produces conservative and predictable outputs. This setting was considered too restrictive for the project's needs, as it might not handle the variability in email content effectively.
- Medium (0.5 - 0.7): Selected a temperature setting of 0.7 to achieve a balance between coherence and variety in outputs. This setting was deemed ideal for parsing and interpreting varied yet coherent text, which is essential for accurately identifying intent in business communications.
- High (0.8 - 1.0): Yields creative but potentially less coherent outputs. This range was avoided as it could lead to irrelevant or incorrect interpretations not suitable for the project's precision requirements.

Email Processing with OpenAI API

- Data Subset: Out of the total dataset of approximately 370,000 emails, the OpenAI API was employed to analyze about 12,500 emails, representing a diverse mix of structured and unstructured formats.
- Processing Time: The task was completed over a period of about 7 hours, which underscores the efficiency of the API in handling large volumes of data with complex language patterns.
- API Utilization: Details of the specific prompt used for processing are outlined in the subsequent slide, which includes the configuration of the API call and the parameters set to optimize accuracy and response time.

This section details the application of the OpenAI API in the project, highlighting its role in enhancing the email parser's ability to accurately identify and categorize intents in emails, a critical component in automating the extraction of meeting-related information.

V. Future Work and Enhancements

As the project moves forward, several areas have been identified for enhancement and development to refine the capabilities of the email parser. These enhancements aim to improve accuracy, efficiency, and scalability of the system.

Fine-Tuning Advanced Language Models

- Explore BERT: Investigate the feasibility of fine-tuning BERT (Bidirectional Encoder Representations from Transformers) for a more nuanced understanding and extraction of meeting-related information from emails. BERT’s architecture is particularly suited for understanding the context of words in emails due to its bidirectional training.
- Experiment with MISTRAL 7B and LLAMA2: These models, known for

their robust intent analysis capabilities, will be adapted to better recognize and process the specific data structures found in business communications. Enhancing their training to focus on the domain-specific nuances can significantly improve their precision.

Enhancing Model Performance

- Dataset Expansion: To increase the robustness and accuracy of the models, the project plans to expand the diversity and size of the training datasets. This involves incorporating more varied examples of email interactions to cover a broader range of scenarios.
- Algorithm Optimization: Continue refining algorithms to handle more complex scenarios, which will reduce the likelihood of misinterpretations or errors during data extraction. Focus will be on improving the models' ability to discern subtle cues within the email text.

Computational Strategies

- Cloud Computing Solutions: Explore more cost-effective and scalable cloud computing options to support the processing demands of large language models. This may involve negotiations with cloud providers for better rates or services that cater specifically to AI-driven projects.
- Distributed Computing: Consider setting up a distributed computing environment that can parallelize processing tasks, thereby enhancing overall performance and reducing processing time.
- Model Compression Techniques: Research into advanced model compression techniques will be pursued to reduce the memory usage and computational demands of the models without compromising their accuracy.

Long-Term Goals

- User Feedback Integration: Develop a mechanism to integrate user feedback into the continuous improvement cycle of the email parser. This will allow the system to adapt and evolve based on actual user experiences and needs.
- Real-Time Processing: Aim to enhance the system's capabilities to allow for real-time processing of emails. This would enable instant extraction and integration of meeting information, significantly improving user productivity and system responsiveness.

VI. Results

This section presents the outcomes of using the GPT-3.5 model for intent analysis in email parsing. The model was tasked with determining if the emails were attempting to schedule meetings and, if so, extracting the relevant details in a specified format.

Methodology

The GPT-3.5 model was provided with a prompt that required it to:

- Determine the intent of scheduling a meeting within the email.
- Provide a binary response ('Yes' or 'No').
- If 'Yes', output the details of the meeting in a structured JSON format.

The JSON format was specified to include:

- sender: Email address of the sender.
- receiver: Email address of the receiver.
- subject: Subject of the email.
- description: A brief summary of the email.
- platform or location: Platform or Location of the meeting.
- data: Additional data in key-value pairs.

Implementation and Processing

The model processed emails and returned outputs based on the criteria set in the prompt. The following instructions were embedded in the prompt to ensure the precision of the output:

- Only return JSON formatted data if the intent to schedule a meeting was confirmed.
- Structure all fields as strings, with dates and times formatted as 'dd-mm-yyyy hh:mm'.

- Ensure parsable JSON responses without the need for further processing or inclusion of special characters outside the JSON body.

Results and Outputs

The implementation showcased the GPT-3.5's capability to handle complex natural language processing tasks efficiently. Below are sample results demonstrating the model's responses:

Example 1:

Subject: Meeting

Body: dave delainey has asked to have the following meeting arranged date friday february 2 2001 at 4 50 pm am location eb 3321 topic california update please let me know if you have any questions or conflicts thanks kay 3063

Yes,

{
"sender": "kay.chapman@enron.com",
"receiver":["dan.leff@enron.com", "william.bradford@enron.com"],
"subject": "Meeting",
"description": "dave delainey has asked to have the following",
"platform or location": "",
"data": {
"Date": "2001-02-01",
"Time": "04:49:00"
}
}

Example 2:

Subject: ABB contract

Body: ben is trying to set up a conference call with abb for next wednesday dale are you available pete thompson is on vacation next week but i suspect he would be glad to call in kay forwarded by kay manncorpenron on112220001256 pm andrew kelemenect 12220001249 pm to sheila tweedhouectect kay manncorpenronenron dale rasmussenhouectect cc subject abb contract we have set a line in the sand with aes for 1501 we need this contract asap for aes to review

Yes

{
"sender": "kay.mann@enron.com",
"receiver":"sheila.tweed@enron.com, dale.rasmussen@enron.com",
"subject": "ABB contract",
"description": "Ben is trying to set up a conference call",
"platform or location": "Conference call",
"data": {}
}

Example 3:

Subject: Yesterday's power outage

Body: as you are painfully aware we had a power outage wednesday morning in the enron building this outage was caused by localized structural failure of the raised floor in our 34th floor data center this resulted in disruption to the power distribution system servicing the phone switch and a number of ena servers as a cautionary move enron online was interrupted while the extent of the failure was assessed resulting in the system being unavailable from 1123 to 1139 all enron building telephones and voicemail were unavailable for approximately 1 hour and 20 minutes and certain ena trading systems were unavailable for over 2 hours during this time the power was stabilized and systems were restored immediate steps are being taken to correct this problem we apologize for this inconvenience if you have any questions in regard to this outage please call philippe bibi at x37698 or bill donovan at x35459

No

These results illustrate the model's effectiveness in discerning the intent behind the emails and extracting detailed information when relevant. The precision in formatting and content extraction confirms the potential of using advanced language models like GPT-3.5 for automated email parsing tasks.

Conclusion

The successful application of the GPT-3.5 model for intent analysis and detail extraction from emails highlights its robustness and

accuracy in performing targeted NLP tasks. This bodes well for further enhancements and practical implementations of the system in real-world scenarios, potentially transforming how organizations manage and utilize email communications for scheduling and information extraction.

VII. Conclusion

The project "Building a Custom Email Parser for Extracting Meeting Information" has made significant strides in addressing the challenges associated with manually processing unstructured email texts. By leveraging advanced language models, specifically the GPT 3.5 Turbu and exploring the potential of the OpenAI API, this project has demonstrated the feasibility and efficiency of automating the extraction of meeting-related information from emails.

The development and fine-tuning of the LLAM2 7B model, despite encountering computational and scalability challenges, highlighted the critical need for appropriate computational resources and more refined model optimization strategies. The integration of the OpenAI API proved to be a pivotal step in enhancing the project's capability to process a substantial subset of the dataset effectively, thereby validating the model's practical utility in real-world applications.

Future Directions:

Going forward, the project will focus on expanding the training dataset, refining the model's ability to understand and process complex and nuanced language constructs, and exploring cost-effective cloud computing solutions to overcome current computational barriers. The long-term goal includes integrating real-time processing capabilities and developing a continuous feedback loop to dynamically improve the model based on user interactions and feedback.

In conclusion, this project not only advances the field of automated email parsing but also provides a scalable framework for further research and development in automated text processing and natural language understanding. The insights gained and the methodologies developed lay a solid groundwork for future enhancements, poised to revolutionize how businesses handle communication and information management.

VIII. References and Resources

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI, 2019.

[2] “Zero-Shot Intent Classification Using a Semantic Similarity Aware Contrastive Loss and Large Language Model,” in IEEE Conference Publication, IEEE Xplore, 2020.

[3] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” arXiv preprint arXiv:1801.06146, 2018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., “Attention Is All You Need,” arXiv preprint arXiv:1706.03762, 2017.

[5] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?” arXiv preprint arXiv:1905.05583, 2019.

[6] “Actionable Email Intent Modeling With Reparametrized RNNs,” AAAI, [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11931.

[7] OpenAI, "OpenAI API Documentation," [Online]. Available: https://platform.openai.com/docs/introduction.

[8] Enron Email Dataset, [Online]. Available: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset.

[9] Hugging Face, "LLama 2," [Online]. Available: https://huggingface.co/blog/llama2.

[10] Mistral AI, "Announcing Mistral 7B," [Online]. Available: https://mistral.ai/news/announcing-mistral-7b/.