**Project Proposal**

**Building a Custom Email Parser for Extracting Meeting Information**

By Somesh Bagadiya

**Introduction**

This project proposes the development of a machine learning-based email parser that automates the extraction of meeting information from emails. By leveraging natural language processing and machine learning techniques, the system will intelligently identify and extract key details such as date, time, location, and participants from unstructured email text. The primary objective is to create a robust and efficient solution that streamlines the extraction process, saves time, and reduces manual effort. The resulting email parser will generate structured output, facilitating seamless integration with calendar applications and scheduling systems, ultimately enhancing productivity and efficiency in managing meeting information.

**Data Collection**

To train and evaluate the machine learning-based email parser, a diverse dataset of emails containing meeting information will be collected using two methods:

Synthetic Email Generation:

* Create email templates incorporating various structures and formats
* Generate synthetic emails programmatically, including meeting-related information
* Cover a wide range of scenarios and variations in the generated emails

Self-Collection and Annotation:

* Manually collect a set of real-world emails from personal or professional sources
* Annotate the collected emails with meeting-related information
* Ensure data privacy and security measures are implemented during collection and annotation

By combining these two data collection approaches, the project aims to create a comprehensive and diverse dataset for training and evaluating the email parser.

**Methodology**

This project will explore two machine learning approaches for extracting meeting information from emails:

1. Conditional Random Fields (CRF):

* Pros:
  + Models the sequential nature of email text.
  + Captures dependencies between different entities.
  + Provides interpretable results.
* Cons:
  + Requires careful feature engineering
  + May struggle with long-range dependencies
  + Limited ability to capture complex semantic information

2. Recurrent Neural Networks (RNN):

* Pros:
  + Captures long-term dependencies and contextual information
  + Automatically learns relevant features from the data
  + Handles variable-length sequences effectively
* Cons:
  + Prone to vanishing or exploding gradient problems
  + May struggle with very long sequences
  + Requires a large amount of labeled training data

The performance of these approaches will be evaluated and compared using appropriate evaluation metrics. The most effective approach will be selected based on its ability to accurately extract meeting information while considering factors such as computational efficiency and scalability.

**Acknowledging Challenges**

Implementing a machine learning-based email parser for extracting meeting information can present several challenges. Here are some common challenges you may encounter.

Data Acquisition and Annotation:

* Obtaining a large and diverse dataset of emails containing meeting information can be challenging.
* Annotating emails with meeting-related entities (e.g., date, time, location, participants) is time-consuming and requires careful guidelines to ensure consistency.

Handling Variations in Email Formats and Structures:

* Emails can have different formats, layouts, and structures, making it difficult to extract meeting information consistently.
* Dealing with various email clients, templates, and writing styles adds complexity to the parsing process.

Ambiguity and Incomplete Information:

* Meeting information in emails may be ambiguous, incomplete, or expressed in different ways.
* Resolving ambiguities and handling missing or partial information can be challenging.
* Dealing with informal language, abbreviations, and contextual references in emails.

Evaluation and Error Analysis:

* Defining appropriate evaluation metrics to measure the performance of the email parser.
* Conducting thorough error analysis to identify common patterns and sources of errors.
* Continuously iterating and improving the parser based on the evaluation results and user feedback.

**Conclusion**

In conclusion, this project proposal presents a well-conceived approach to developing a machine learning-based email parser for extracting meeting information from emails. By leveraging natural language processing techniques, specifically Conditional Random Fields (CRF) and Recurrent Neural Networks (RNN), and utilizing a combination of synthetic email generation and self-collection with annotation for data collection, the proposed system aims to automate the process of identifying and extracting crucial details from unstructured email text. Despite the acknowledged challenges, the successful implementation of this email parser has the potential to significantly reduce manual effort, enhance productivity, and streamline meeting management processes, ultimately delivering a high-impact and practical tool for individuals and organizations seeking to optimize their email management and meeting coordination workflows.

**References**

1. [DeepParse: A Trainable Postal Address Parser](https://ieeexplore.ieee.org/abstract/document/8615844)
2. [Annotating Needles in the Haystack without Looking: Product Information Extraction from Emails](https://dl.acm.org/doi/abs/10.1145/2783258.2788580)
3. [Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism](https://ieeexplore.ieee.org/abstract/document/8701426)
4. [Actionable Email Intent Modeling With Reparametrized RNNs](https://ojs.aaai.org/index.php/AAAI/article/view/11931)