CaterExtract

Presented by: Arkadi Doktorovich, Hen Mandelbaum, Guy Yogev

Email Threads Extraction

When requesting catering quotes for events, it is possible to receive dozens of emails that are full of information about the quotes but also have additional and ineffective information, making it difficult to find the best quote. The goal of our project is to enable quick filtering of relevant details and setting and changing data based on email conversations between the event organizer and the caterer.

Project Goal

Our objective is to extract key features like **price, date, menu items, and constraints** from email threads. Inputs are raw email text, and outputs are structured JSON objects.

Novelty

This project generates realistic business email threads with structured feature annotations, filling a gap since real emails are private and rarely available for research.

Articles for Email Understanding and Feature Extraction

Title + Year	Task	Methods	Dataset	Results	Relation to Our Project
Email Summarizer: A Novel Hybrid Approach to Email Summarization (2025)	Summarization of individual emails and email content.	Hybrid extractive approach combining TF-IDF, LDA topic modeling, and MiniLM embeddings.	Real world email messages collected for summarization evaluation.	The hybrid model outperformed classical baselines such as LexRank and lead-based methods.	Demonstrates effective feature extraction from email text using combined statistical and semantic features.
Large Language Models Are Zero-Shot Text Classifiers (2023)	Zero-shot text classification without task-specific training.	Prompt based classification using LLMs with chain-of-thought reasoning.	Multiple benchmark text classification datasets (e.g., sentiment, spam).	LLMs achieved competitive or superior results compared to traditional supervised classifiers.	Motivates using LLMs for extracting structured features from catering emails with minimal annotation.
EMAILSUM: Abstractive Email Thread Summarization (2021)	Summarization of multi email threads.	Abstractive and extractive models using pretrained transformers such as T5.	EMAILSUM dataset with 2,549 annotated email threads.	Pretrained models significantly improved summary quality over non-pretrained baselines.	Highlights challenges and solutions for modeling email threads, which is essential for our feature extraction task.

Email Threads Dataset

Dataset Details

- **Source:** Generated via GPT-4o-mini.
- **Size:** 30 email threads, with 30 ground truth features.

Data Generation & Labeling

- Both the Email threads and ground truth are generated at the same time then separated into an email file and json file.
- The labels are assigned randomly to each email thread for diversity, that is also being regulated by the 0.7 temperature.
- The prompt takes in account most instances so bad email threads won't be generated.

Labels

event_type: string or null,
price_type: per_person | fixed | range | unknown,
final_price_value: number or null,
min_guests: number or null,
max_guests: number or null,
includes_vat: true | false | null,
is_kosher: true | false | null,
kosher_supervision: string or null,
includes_bar: true | false | null,
menu_type: meat | dairy | mixed | pareve | unknown,
dietary_options: [array of strings],
event_date: YYYY-MM-DD or null,
cancellation_policy: string or null,
menu_highlights": [array of strings],
extra_notes: string or null

Initial Approach & Evaluation

Baseline Approach

- **Tested models:**
  GLiNER - NER sees tokens and not decisions so it has issues recognizing values like total price or per person.
  Flan-T5 - outputs were inconsistent, often incomplete or invalid JSON.
- **Final Baseline Model: GPT-4o-mini**, zero shot prompt based feature extraction from entire email threads.
- **Custom Prompt Rules:** Enforced JSON schema, list fields, booleans, date format, and numeric fields.

Evaluation Results

Model/ Approach	Success/ Exact Match	Notes
GPT-4o-mini (zero-shot)	~68%	Captures thread context and most features accurately
GLiNER	~45%	Missed context, partial extractions
Flan-T5	0	Inconsistent, invalid JSON outputs

Error Analysis

- **JSON Parsing Issues:** GPT occasionally outputs invalid JSON or mis formats arrays.
- **List Variability:** Minor differences in menu/dietary item wording reduce exact match score.
- **Thread Length & Context:** Longer threads or complex negotiations may lead to missing details in extraction.

Project Plan

Step	Description	Deadline	Expected Outcome
Data Creation	Expand dataset of email threads, covering diverse event types, pricing, and menus.	Week 10	Dataset of 500–1500 high-quality email threads ready for feature extraction experiments.
Feature & Label Refinement	Test and improve labels/features, ensure consistency, and handle edge cases in the dataset.	Week 10	Clear, consistent labeling schema, dataset ready for reliable evaluation.
Prompt Optimization & Fine-Tuning	Experiment with GPT-4o-mini zero-shot vs few-shot prompts for feature extraction and Including fine-tuning a Ner model like GLiNER.	Week 11	Determine optimal prompt design and NER model optimization, improved extraction accuracy and robustness.
Evaluation & Analysis	Perform quantitative evaluation of GPT-4o-mini zero-shot and few-shot on the refined dataset. Analyze errors and feature level performance.	Week 11	Comprehensive report with exact match rates, and insights on strengths/weaknesses.
Prepare Final Presentation	Compile all results, visualizations, and insights into a final slide deck.	Week 12	Clear, data driven final presentation demonstrating dataset, methodology, and feature extraction performance.