LEARNED AUTOMATIC RECOGNITION EXTRACTION OF APPOINTMENTS FROM EMAIL

LAUREN PAONE
FACULTY ADVISOR: FERNANDO PEREIRA

ABSTRACT
• Try to automate the extraction of important information regarding appointments and events from email.

• 2 problems:
  ▪ Identifying whether the email contains information about an event.
  ▪ Extracting the title, date, time, and location of the event.

SYSTEM OVERVIEW

Components
• Parser reads the email and stores all of its information.
• Classifier takes in the parsed email and outputs whether the email contains an event or not.
• Extractor takes emails containing an event and outputs the titles, dates, times, and locations associated with the event.

Both the classifier and the extractor employ machine learning techniques.
• Trained and tested on pre-labeled emails.
• Have an array of specific features designed to improve accuracy.
• During training, each feature is assigned a weight that represents the likelihood that an email or token with that feature’s characteristics will be labeled correctly.

The extractor and classifier are being tested separately. The output of each is statistics that measure how well the component is doing.

Accuracy is measured best with:
• FMeasure - weighted average of precision and recall. (High FMeasure implies high accuracy)

The testing and training emails came from 2 sources:
• My SEAS account.
• A corpora released by Carnegie Mellon University after doing research of a similar nature.

THE CLASSIFIER
• Non-sequential maximum entropy model - direction based model that classifies emails independently of how previous emails were classified.

• Each email is associated with a linked list containing all features matching that email. If the feature is part of the testing data, based on the linked list of features and their weights, the classifier decides whether there is an event.

• Event-specific features added:
  ▪ Structured features test if there is formatted event information.
  ▪ Event Bigram features identify some 2-word strings that imply an event.
  ▪ Event Divider features look for character strings that may separate event information from the rest of the body.

CLASSIFIER RESULTS

• Expected to increase with features.
• Event Bigrams was best for my email but did not improve CMU results.

THE EXTRACTOR
• Sequential model - takes into account previous classifications. Implemented using Conditional Random Fields (probabilistic model to label sequence data).

• Input file is email bodies and subjects split up into tokens. Features are tagged to each individual token.

• The extractor was equipped with baseline features but many features were added to specifically recognize a title, date, time, and location.

EXTRACTOR RESULTS

• Overall, my features improved the extractor.
• Best with 30 iterations.

• Time and date results were best because they have easily identifiable formats.

CONCLUSION
• Both the classifier and extractor had decent results that could be improved by adding similar features or tweaking the existing features.

• Once both are improved, the two components can be connected and this will be a working system that can be integrated into an email client.

---

SENIOR PROJECT POSTER DAY 2006, CIS DEPT. UNIVERSITY OF PENNSYLVANIA