

The Yahoo! logo is displayed in white, stylized, serif font against a purple background. The background of the slide features a series of concentric, light gray circles that create a ripple effect, centered behind the main text.

YAHOO!

Automated Content Labeling using Context in Email

Aravindan Raghuv
Yahoo! Inc, Bangalore.

YAHOO!

Show of hands!



at-least one email with an attachment last 2 weeks?

Introduction: The “What ?”

- Email attachments are a very popular mechanism to exchange content.
 - Both in our personal / office worlds.
- The one-liner:
 - The email usually contains a crisp description of the attachment.
 - **“Can we auto generate tags for the attachment from the email ?”**

Introduction: The “Why?”

- Tags can be stored as extended attributes of files.
- Applications like desktop search can use these tags for building indexes.
- Tags generated even without parsing content:
 - Useful for Images
 - In some cases, the tags have more context than the attachment itself.



Outline

- Problem Statement
- Challenges
- Overview of solution
- The Dataset : Quirks and observations
- Feature Design
- Experiments and overview of results
- Conclusion

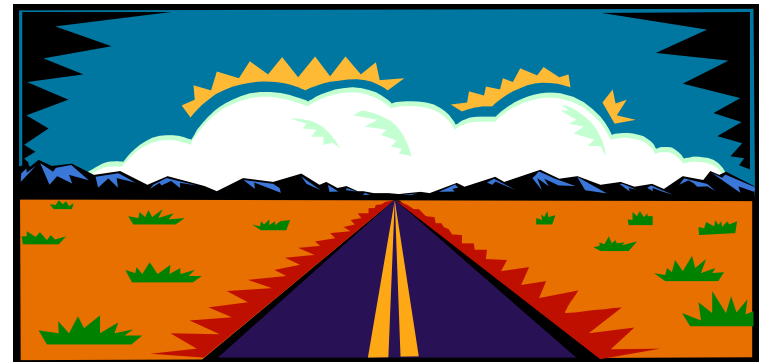


Problem Statement

“Given an email E that has a set of attachments A , find the set of words K_{EA} that appear in E and are relevant to the attachments A .”

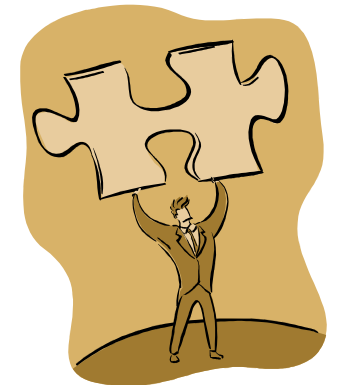
Subproblems:

- Sentence Selection
- Keyword Selection



Overview of Solution

- Solve a binary classification problem :
 - Is a given sentence relevant to the attachments or not?
- Two part solution:
 - What are the features to use?
 - Most insightful and interesting aspect of this work.
 - Focus of this talk
 - What classification algorithm to use?



Challenges : Why is the classification hard?

- Only a small part of the email is relevant to the attachment. Which part?
- While writing emails: users tend to rely on context that is easily understood by humans.
 - usage of pronouns
 - nick names
 - cross referencing information from a different conversation in the same email thread.

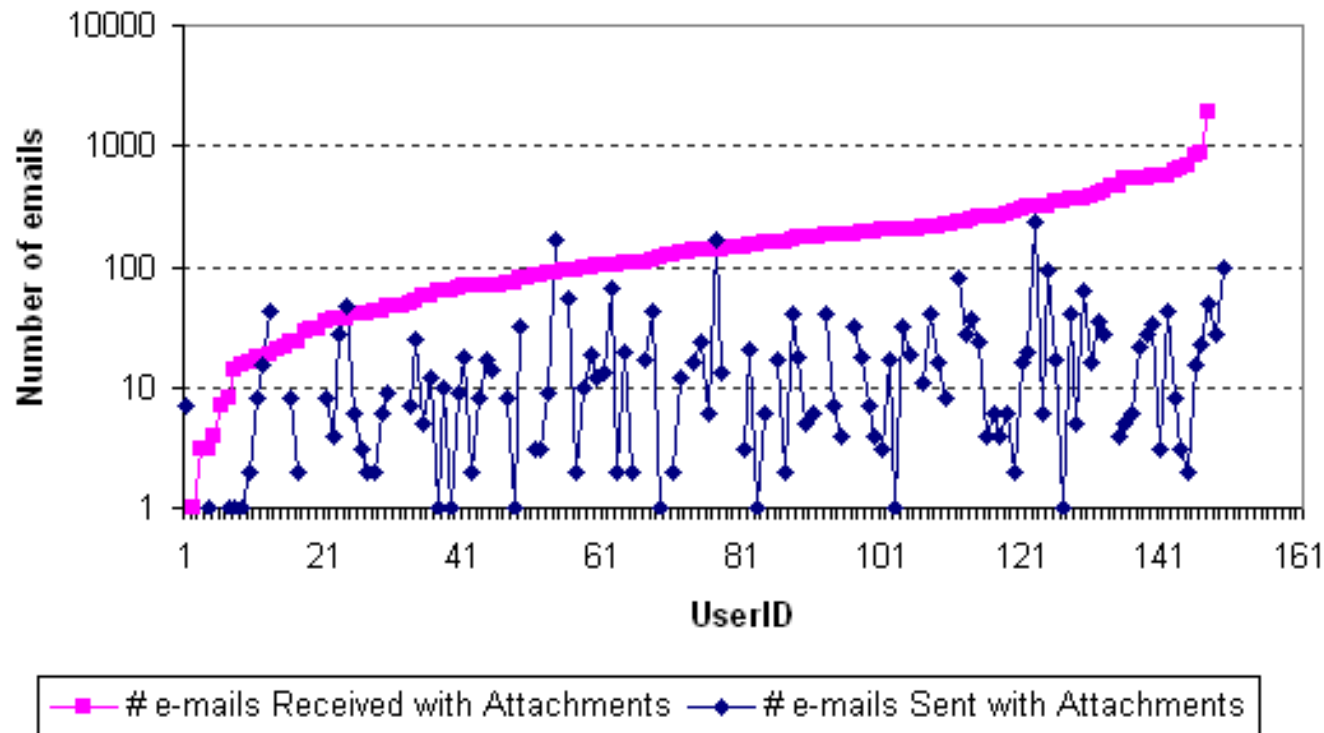
The Enron Email Dataset

- Made public by [Federal Energy Regulatory Commission](#) during its investigation [1].
- Curated version [2] consists of 157,510 emails belonging to 150 users.
- A total of 30,968 emails have at least one attachment

[1] <http://www.cs.cmu.edu/~enron/>

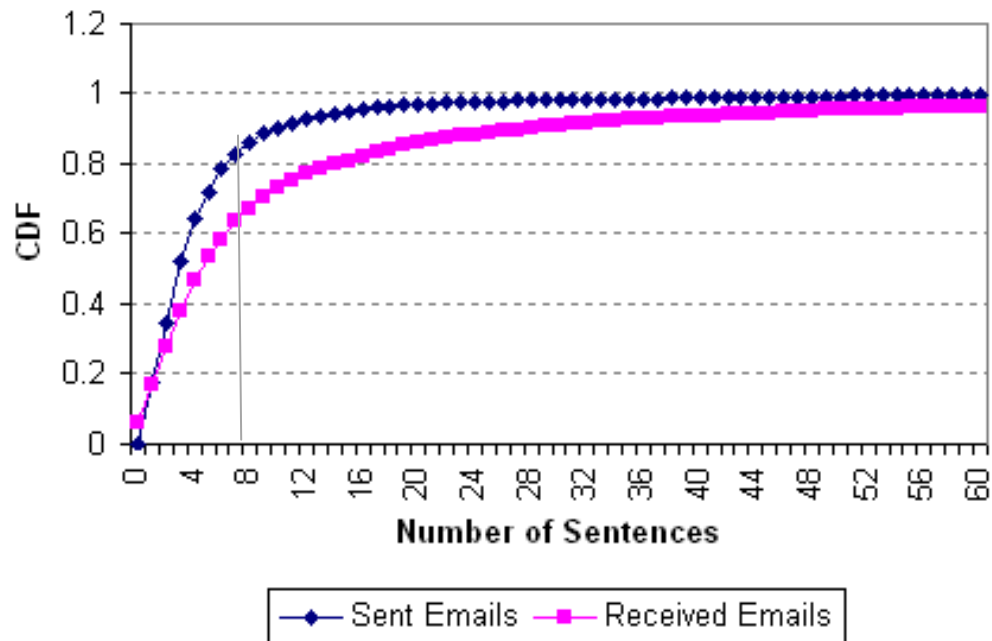
[2] Thanks to Mark Dredze for providing the curated dataset.

Observation-1: User behavior for attachments



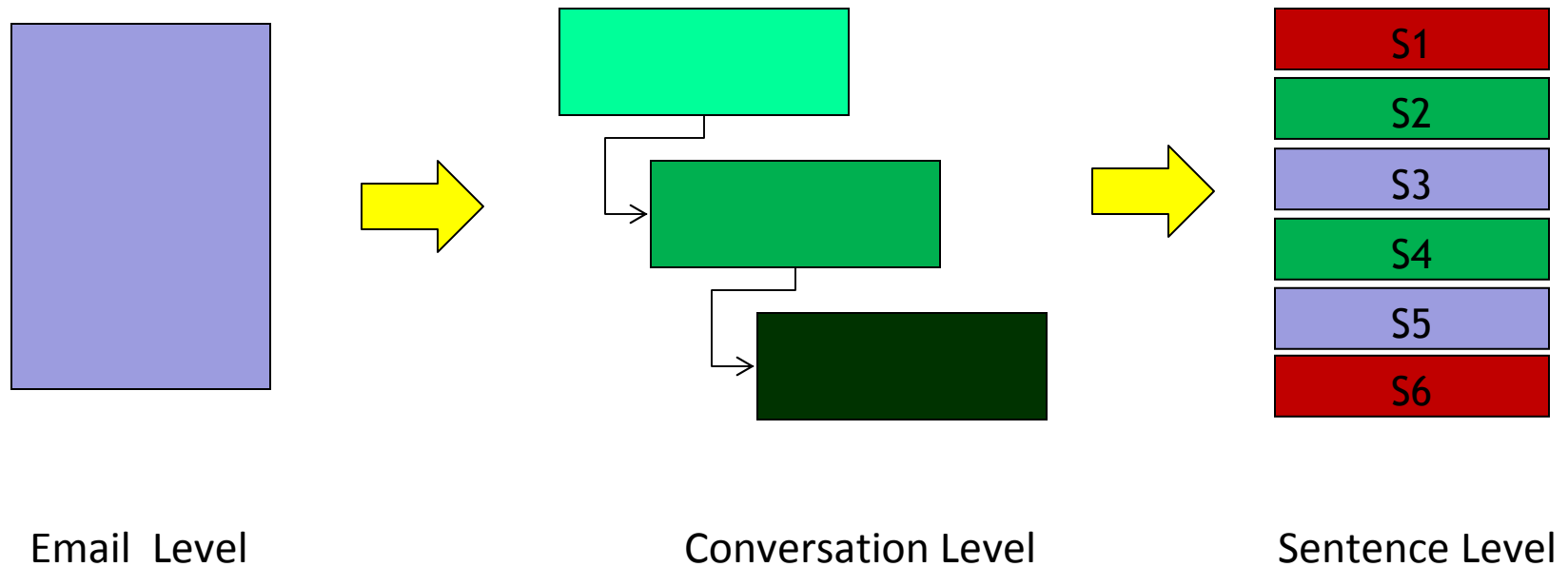
- Barring a few outliers, almost all users sent/received emails with attachments.
- A rich / well-suited corpus for studying attachment behavior

Observation-2: Email length

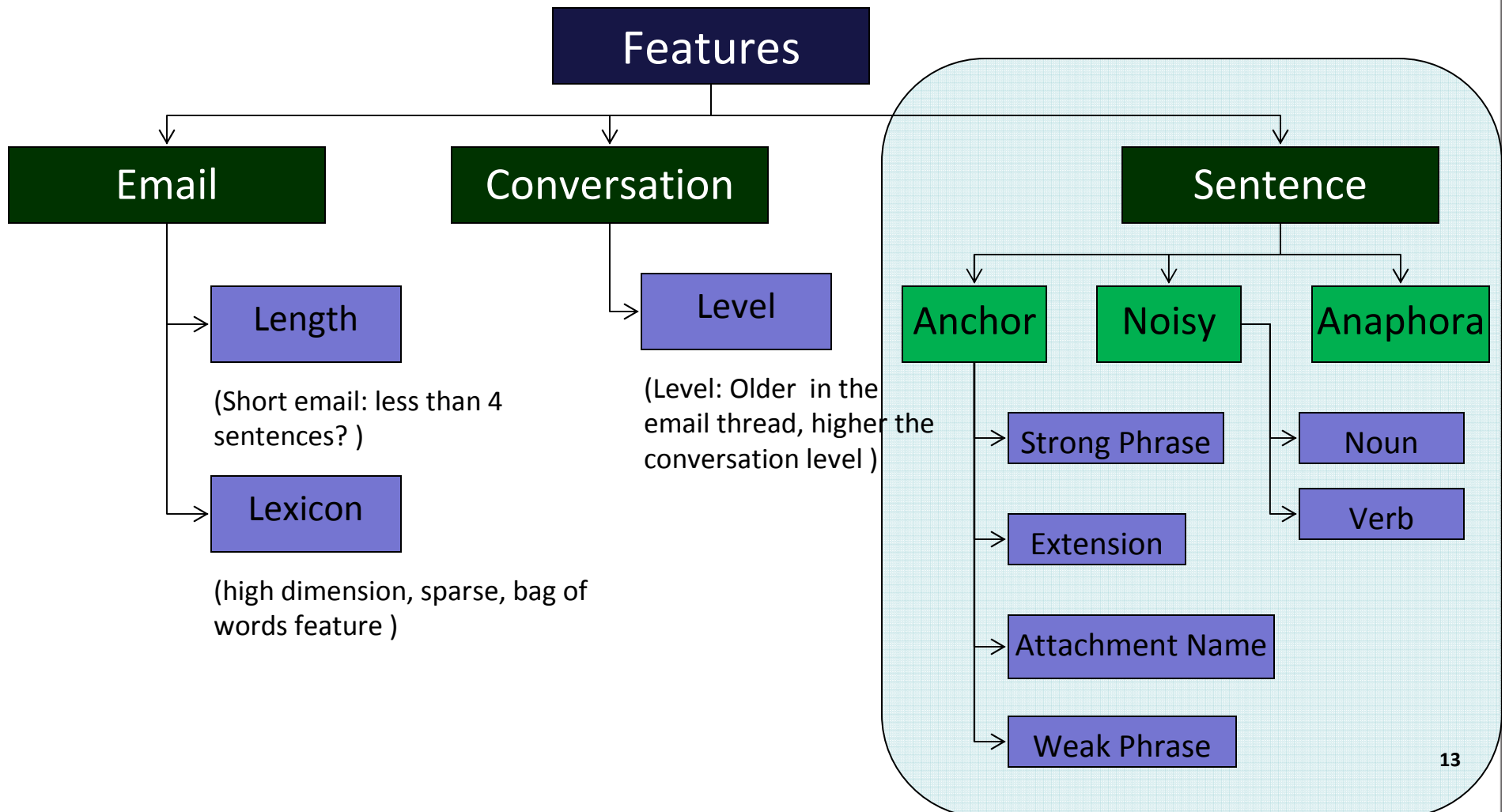


- Roughly 80% of the emails have less than 8 sentences.
- In another analysis: even in emails that have less than 3 sentences, not every sentence is related to the attachment!!

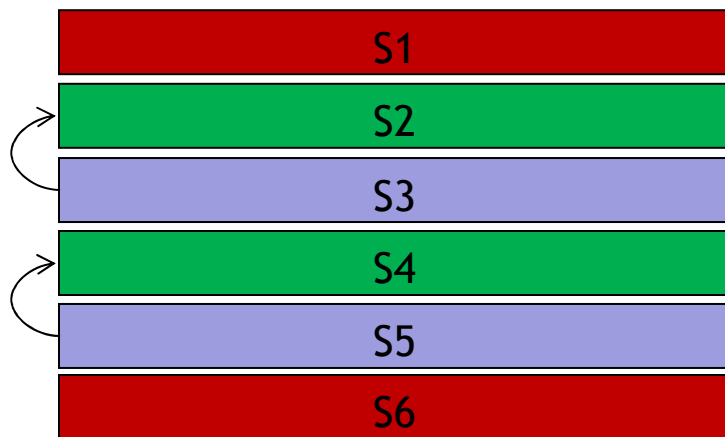
Feature Design : Levels of Granularity



Feature Taxonomy:



Feature Design : Sentence level



Anchor Sentences:
Most likely positive
matches.



Noisy Sentences:
Most likely negative
matches.



Anaphora Sentences: have linguistic relationships
to anchor sentences

Feature Design: Sentence Level → Anchor

- Strong Phrase Anchor: Feature value set to 1 if sentence has any of the words:
 - attach
 - here is
 - Enclosed
- of the 30968 emails that have an attachment, 52% of them had a strong anchor phrase.

Feature Design: Sentence Level → Anchor

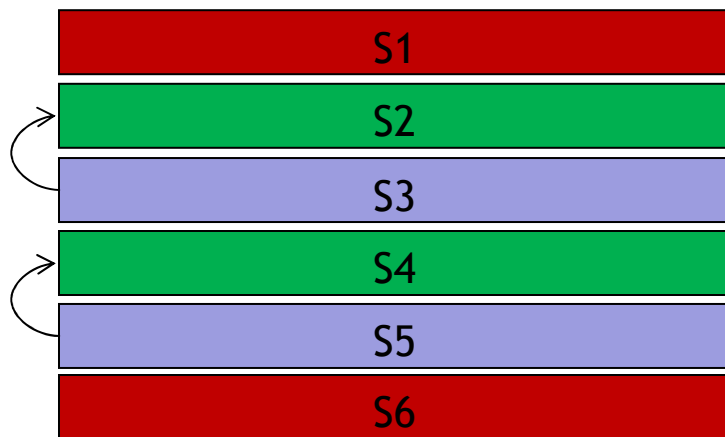
- Behavioral Observation: Users tend to refer to an attachment by its file type
- **Extension Anchor:** Feature value set to 1 if sentence has any of the extension keywords:
 - xls → spreadsheet, report, excel file
 - jpg → image, photo, picture
- **Example:**


“Please refer to the attached spreadsheet for a list of Associates and Analysts who will be ranked in these meetings and their PRC Reps.”


Feature Design: Sentence Level → Anchor

- Behavioral Observation: Users tend to use file name tokens of the attachment to refer to the attachment
- **Attachment Name Anchor:** Feature value set to 1 if sentence has any of the file name tokens.
 - Tokenization done on case and type transitions,
- **Example:**
 - attachment name “Book Request Form East.xls”
 - “These are book requests for the Netco books for all regions.”

Feature Design : Sentence level



 Anchor Sentences:
Most likely positive matches.

 Noisy Sentences:
Most likely negative matches.

 Anaphora Sentences: have linguistic relationships to anchor sentences

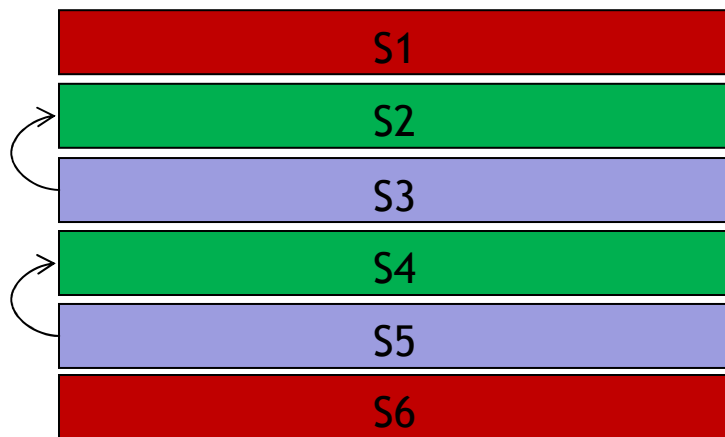
Feature Design : Sentence level → Noisy


- Noisy sentences are usually salutations, signature sections and email headers of conversations.
- Two features to capture noisy sentences
 - Noisy Noun
 - Noisy Verb


- **Noisy Noun:** marked true if more than 85% of the words in the sentence are nouns.

- **Noisy Verb:** marked true if no verbs in the sentence

Feature Design : Sentence level



 Anchor Sentences:
Most likely positive matches.

 Noisy Sentences:
Most likely negative matches.

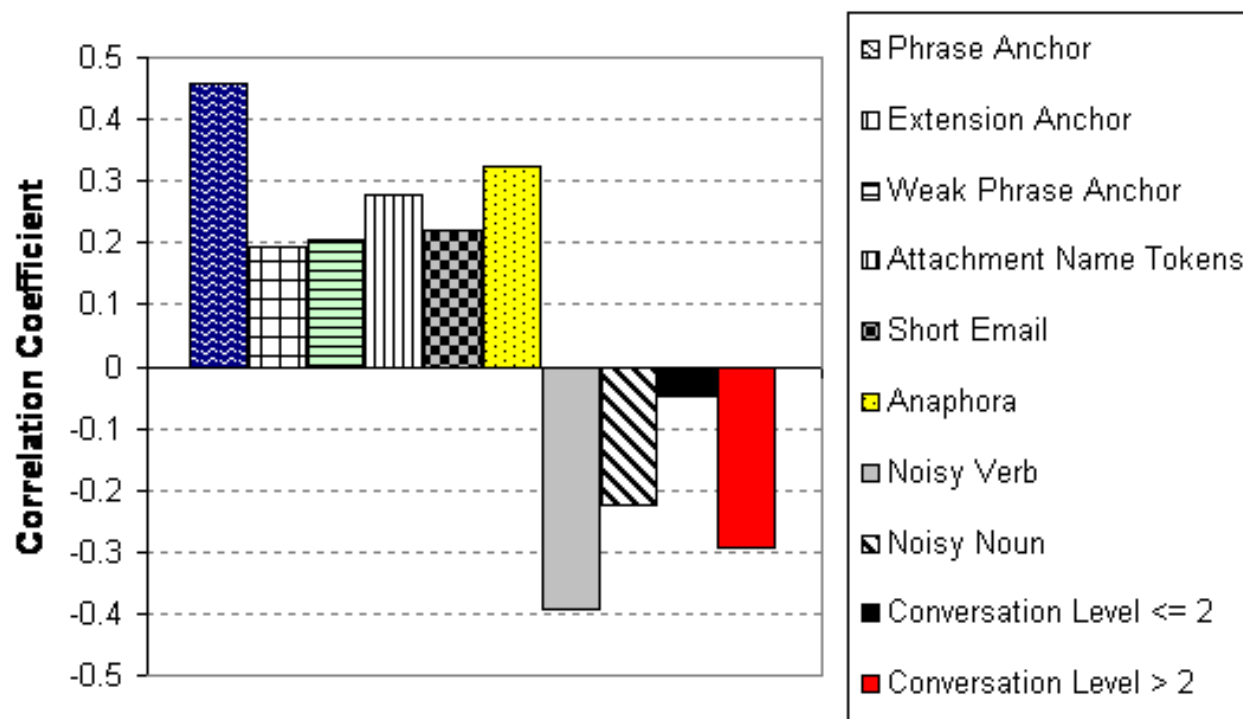


Anaphora Sentences: have linguistic relationships to anchor sentences

Feature Design : Sentence level → Anaphora

- Once anchors have been identified:
 - NLP technique called anaphora detection can be employed
 - Detects other sentences that are linguistically dependent on anchor sentence.
 - Tracks the hidden context in email.
- Example:
 - “Thought you might me interested in the report. It gives a nice snapshot of our activity with our major counterparties.”

Correlation Analysis



Best positive correlation:

- Strong phrase anchor
- Anaphora feature

Short email \rightarrow low correlation coefficient.

noisy verb feature \rightarrow good negative correlation

conversation level ≤ 2 feature has lower negative correlation when compared to the conversation level > 2 feature.

Experiments



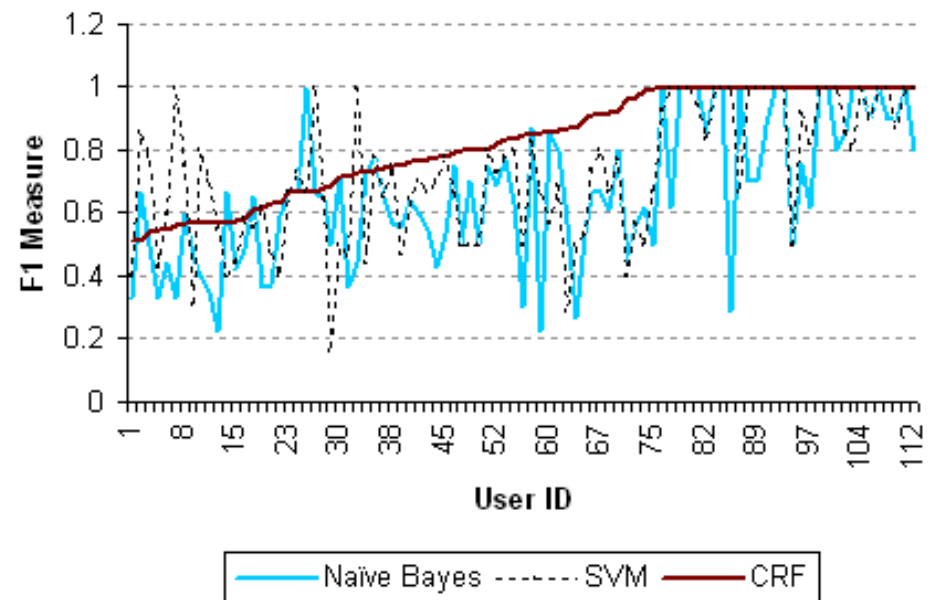
- Ground Truth Data
 - randomly sampled 1800 sent emails.
 - Two independent editors for producing class labels for every sentence in the above sample.
 - Reconciliation:
 - Discarded emails that had at least one sentence with conflicting labels.
- ML Algorithms studied : Naïve Bayes, SVM, CRF.

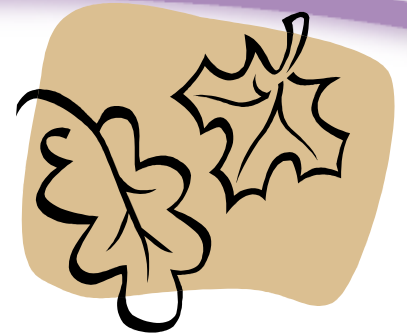
Summary of Results: F1 measure

- With all features used, F1 scores read:
 - CRF : 0.87
 - SVM: 0.79
 - Naïve Bayes: 0.74
- CRF consistently beats the other two methods across all feature sub-sets
 - Sequential nature of the data works in CRF's favor.
- The Phrase Anchor provides best increase in precision
- The Anaphora feature provides best increase in recall.

Summary of Results: User Specific Performance

- In majority of the cases, CRF outperforms
- With same set of features:
 - the CRF can learn a more generic model applicable to a variety of users





In the paper . . .

- A specific use-case for attachment based tagging:
 - Images sent over email
 - A user study based on survey
- More detailed evaluation studies / comparison for Naïve Bayes, SVM, CRF
- TagSeeker : A prototype for proposed algorithms implemented as a Thunderbird plugin.

Closing Remarks

- Improvement of retrieval effectiveness due to keywords mined:
 - Could not be performed because the attachment content is not available.
- Working on an experiment to do this on a different dataset.
- Thanks to the:
 - Reviewers for the great feedback!
 - Organizers for the effort putting together this conference!





Conclusion

- Presented a technique to extract information from noisy data.
- The F1 measure of the proposed methodology
 - In the high eighties. Good!
 - Generalized well across different users.
- For more information on this work / Information Extraction @ Yahoo!
 - aravindr@yahoo-inc.com

YAHOO!

Thanks

The word "Thanks" is written in a bold, purple, sans-serif font with a thick black outline. The letter 'h' is replaced by a tan-colored hand with fingers spread, pointing upwards. To the right of the hand, there are three green exclamation marks. The entire graphic is set against a light blue, tilted rectangular background.

Related Work

- Email Organization