Two-Stage Stochastic Natural Language Generation for Email Synthesis by Modeling Sender Style and Topic Structure

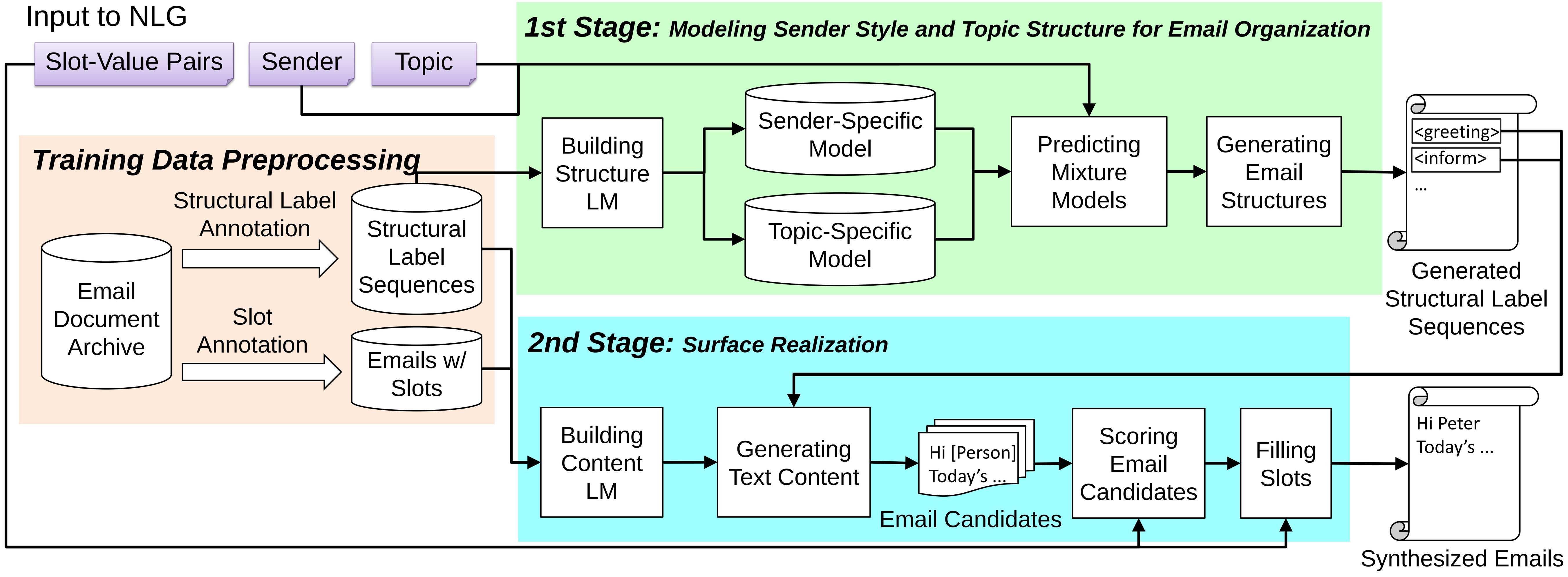
Yun-Nung (Vivian) Chen and Alexander I. Rudnicky

Carnegie Mellon, lti

1. The Task

- Motivation
  - Generate emails that reflect **sender style** and **intent of communication**
  - Provide emails as part of **synthetic evidence** of insider threats for purposes of training, prototyping, and evaluating anomaly detectors.
- Approach
  - Senders' characteristics are modeled based on their **writing patterns** (structure, politeness, etc.) instead of their attitudes
  - 1st Stage: modeling sender style and topic structure for email organization
  - 2nd Stage: stochastic generation of language for surface realization

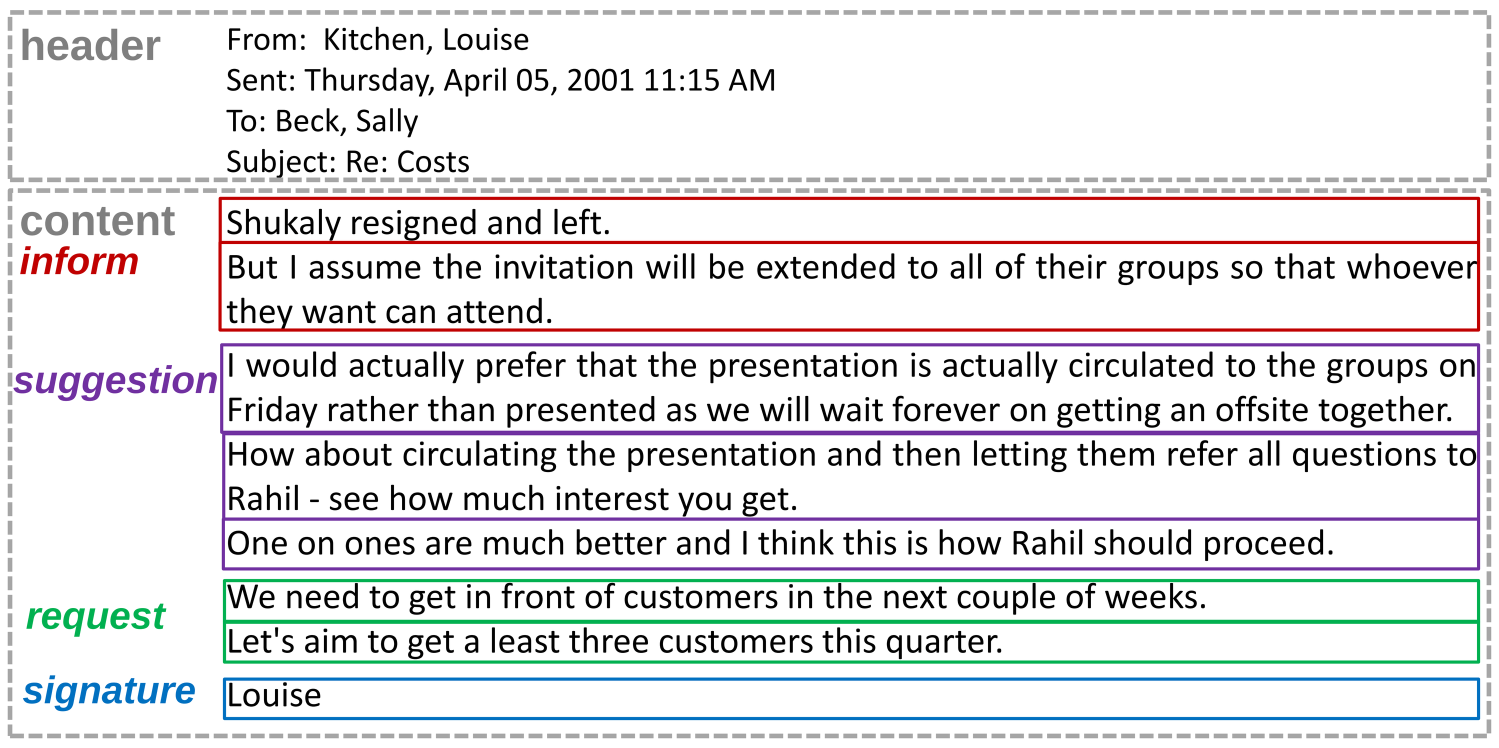
2. The Framework



- 1st stage structures the emails according to **sender style** and **topic structure** (high-level generation)
- 2nd stage synthesizes text content based on **the particulars of an email element** and **the goals of a given communication** (surface-level realization).

3. Training Data Preprocessing

- **Structural Label Annotation**
  - 10 email structure elements (*greeting, inform, request, suggestion, question, answer, regard, ack., sorry, sign*)



- Different senders tend to structure emails in different ways.

- **Slot Annotation**
  - General class: 7-class extracted by Named Entity Recognition (*location, person, org., time, money, percent, date*)
  - Topic class: 3-class extracted by keywords (*meeting, issue, discussion*)

4. Modeling Sender Style and Topic Structure for Email Organization

- **Each email can be treated as a structural label sequence**

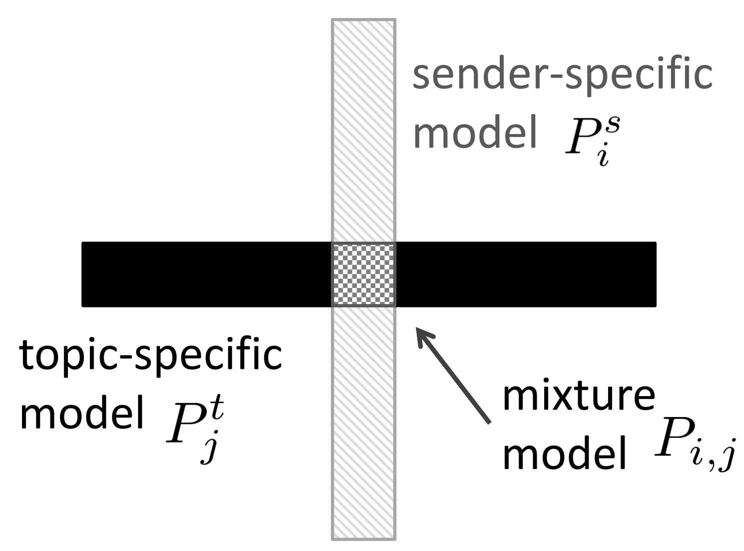
For each structural label:

1) **Building Structure Language Models**
   - Sender-specific structure LM (trigram w/ smoothing)
   - Topic-specific structure LM (trigram w/ smoothing)

- A sender may have personal style about email structure.
- Emails about the same topic may have similar structures.

2) **Predicting Mixture Models**

$$P_{i,j}(l) = \alpha P_i^s(l) + (1-\alpha) P_j^t(l)$$



3) **Stochastically Generating Email Structures**
   - Generate structural label sequences randomly according to dist. of mixture models

5. Surface Realization

For each structural label:

- **Build Content Language Model**
  - Cross-sender content LMs (5-gram w/o smoothing)

For each generated structural label:

1) **Stochastically Generate Text Content**
2) **Score Email Candidates**
   - We penalize the synthesized email if it:
     - contains slots without provided values
     - doesn't have the required slots
3) **Fill Slots**
   - Tomorrow's [meeting] is at [location].

   → Tomorrow's speech seminar is at Gates building.

(%)	Template	Stochastic	No Diff
Coherence	36.19	**38.57**	25.24
Fluency	28.10	**40.48**	31.43
Naturalness	35.71	**45.71**	18.57
Preference	36.67	**42.86**	20.48
Overall	34.17	**41.90**	23.93

The ratio of subjects' preference according to different criteria

6. Experiments

- **Evaluation of Sender Style Modeling**
  - Rate synthesized emails for each sender on a scale of 1 (highly confident that email is not from the sender) to 5 (highly confident that email is from the sender)
  - Average normalized scores the corresponding senders receive: **45% > 33%** [for 3 senders]
- Sender style can be noticed by subjects based on greeting usage, politeness, the length of email, etc.
- **Evaluation of Surface Realization**
  - Compare template-based generation (sentence-level NLG) and stochastic generation (word-level NLG) on the same email structures.
- The word-based stochastic generation outperforms the template-based algorithm and requires less effort in terms of knowledge engineering.

7. Conclusions

- We propose a two-stage stochastic NLG process for email synthesis that models sender style and topic structure.
- Subjects can detect sender style and can differentiate template-based (sentence-level) and stochastically-generated sentences (word-level).
- This technique can be used to create realistic emails and that email generation could be carried out using mixtures containing additional models based on other characteristics.
- The current study shows that email can be synthesized using a small corpus of labeled data; however these models could be used to bootstrap the labeling of a larger corpus which in turn could be used to create more robust models.