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

   ![Diagram of the two-stage framework for email generation](image)

   - **1st Stage: Modeling Sender Style and Topic Structure for Email Organization**
     - **Building Structure LM**
     - **Sender-Specific Model**
     - **Topic-Specific Model**
     - **Generating Email Structures**

   - **2nd Stage: Surface Realization**
     - **Building Content LM**
     - **Generating Text Content**
     - **Generating Email Structures**
     - **Filling Slots**

3. Training Data Preprocessing

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

- **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**

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

     - **2) Predicting Mixture Models**
       - $P_{ij}(l) = \alpha P^s_i(l) + (1 - \alpha) P^t_j(l)$

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

5. Surface Realization

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

- **Evaluation of Surface Realization**
  - Compare template-based generation (sentence-level NLG) and stochastic generation (word-level NLG) on the same email structures.

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).
   - The 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.