Chatbots can be made more polite using style transfer & synthetic data

Polite Chatbot: A Text Style Transfer Application

**Sourabrats Mukherjee**
mukherjee@ufal.mff.cuni.cz

**Vojtěch Hudeček**
hudecek@ufal.mff.cuni.cz

**Ondřej Dušek**
odusek@ufal.mff.cuni.cz

**Presented at EACL 2023, Dubrovnik, Croatia.**

https://github.com/souro/polite_chatbot

**Supported by:** Charles University GAUK 392221, GAUK 30720, SVV 260575, ERC (No. 101039303 NG-NLG).

**Special thanks:** Zdeněk Kasner

---

**Our approach**

1. Politeness transfer model:
   - politeness markers: identified and removed from polite texts (TF-IDF-based)
   - BART trained to reconstruct politeness markers

2. Synthetic polite chat data:
   - generated by applying politeness transfer model

3. Polite dialogue model:
   - BlenderBot, DialoGPT, and GPT-2 as the dialog models
   - finetuned them using synthetic polite data for generating polite responses

**Experiments**

**Data**
- Politeness:
  - preprocessed & filtered Enron email sentences (Madaan et al., 2020)
- Dialog:
  - DailyDialog (DD): open-domain human-human dialogues (Li et al., 2017)

**Metrics**
- Automatic evaluation
  - politeness: politeness classifier
  - coherence: BLEU, SBERT cosine similarity (CS)
- Human evaluation
  - politeness, coherence, fluency
  - 1-5 Likert scales

**Results**

- Politeness transfer (Enron email data)
- Synthetic dialog data (DailyDialog)

**Conclusions**

Resulting model is end-to-end, no post processing needed

Compared to baselines, we achieve:
- Improved politeness
- Comparable content similarity, fluency, coherence

Future work: other stylized response generation tasks