Politeness Transfer: A Tag and Generate Approach

Aman Madaan*, **Amrith Setlur***, **Tanmay Parekh***, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, Shrimai Prabhumoye

School of Computer Science
Carnegie Mellon University

*First authors. Order decided randomly.

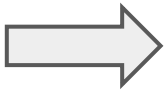
Outline

- Problem Statement: Politeness Transfer
- Proposed Methodology
- Experiments and Results

Problem Statement: Politeness Transfer

Converting non-polite sentences to polite sentences while preserving the meaning

Send me the data



Could you please send me the data?

Applications

- Automatic email responses
- Personal chatbots
- Education


Hi Rebekah

I wanted to take a minute to personally thank you for supporting Pasta Marco over the last 15 years.

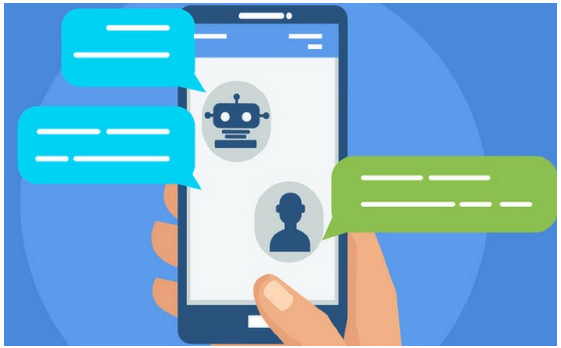
The pasta joint wouldn't be possible without amazing customers like you. As a special thank you, we're giving away free meatballs all month long. Because you're on our email list, you're the first to know!

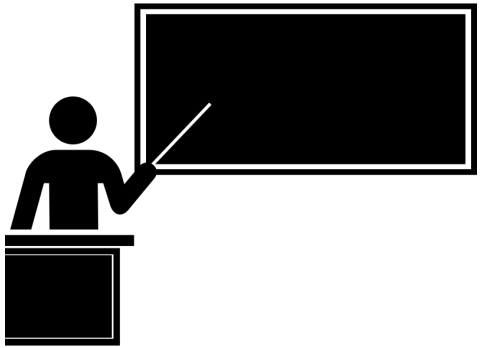
Also, I'd like to invite you to a special Customer Appreciation Day on Saturday, September 14. Join us for pasta tastings and a special spaghetti making class hosted by Chef Marco himself.

Spots are limited, so register here:

Thanks for all your years of support!

Sincerely,
Bek at Pasta Marco





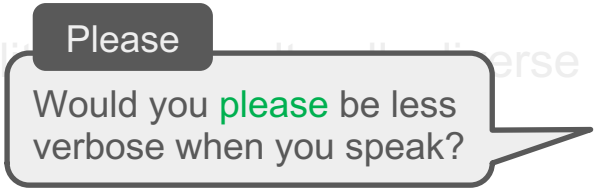
Challenges

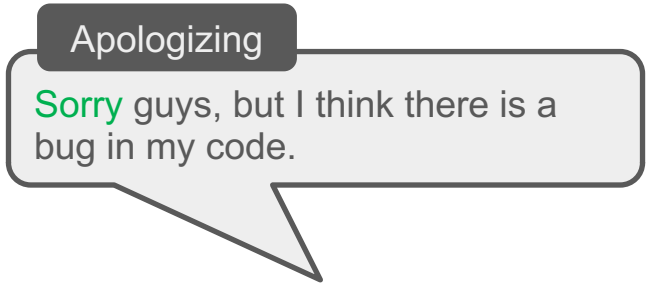
- Politeness is culturally diverse

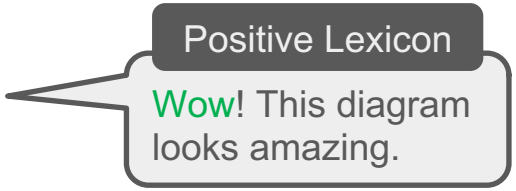
Could you please open the door

Challenges

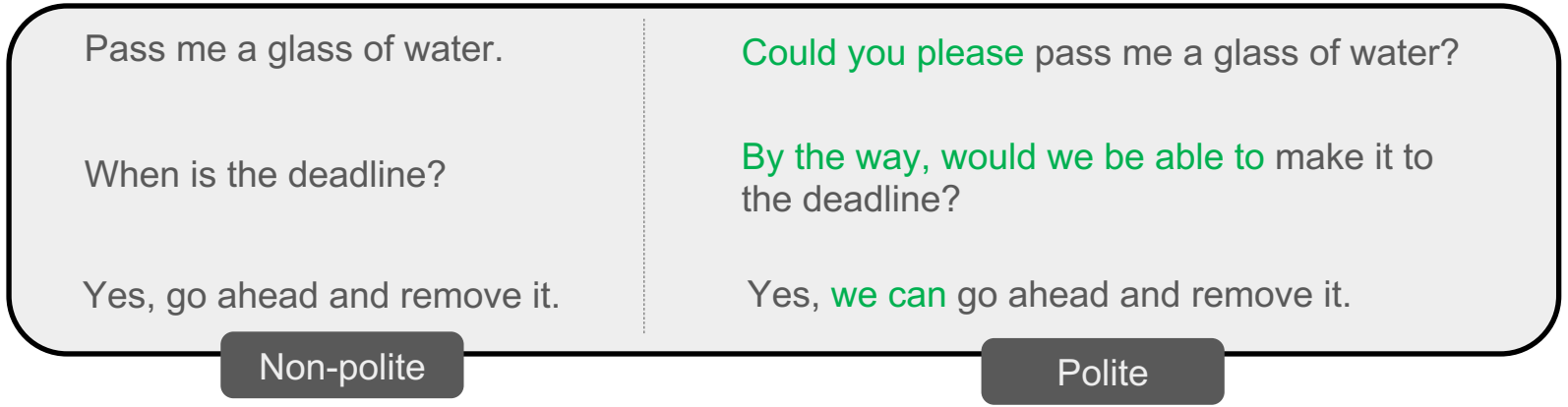
- Politeness is subtle. It isn't just "please" and "thank you" [Danescu-Niculescu-Mizil et al., 2013]







Challenges



- Ill-defined class: Non-polite

Easy to detect presence of politeness
Difficult to objectify absence of politeness

Challenges

- Politeness is culturally diverse
- Politeness is subtle. It is a spectrum and not binary [Danescu-Niculescu-Mizil et al., 2013]
- Ill-defined

No Labeled Data

- Data Paucity

Our Focus

Politeness accepted by North American English speakers + a formal setting (as defined in [Danescu-Niculescu-Mizil et al., 2013])

DIVERSE

DEFINITION?

SCOPE

Transfer only from polite to polite

Impolite = Insulting Phrases + Curse-words

Not all sentences can be made more polite

We focus on converting request or action-directives [Jurafsky, 1997] to polite requests (while preserving the meaning)

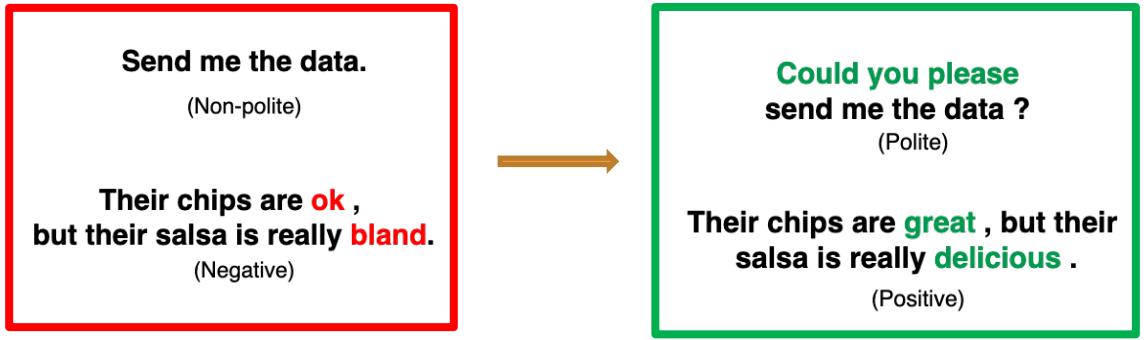
Outline

- Problem Statement: Politeness Transfer
- Proposed Methodology
- Experiments and Results

Text Attribute Transfer by Learning to Tag and Generate

Transfer Desiderata:

1. Successful transfer into target style.



Text Attribute Transfer by Learning to Tag and Generate

Transfer Desiderata:

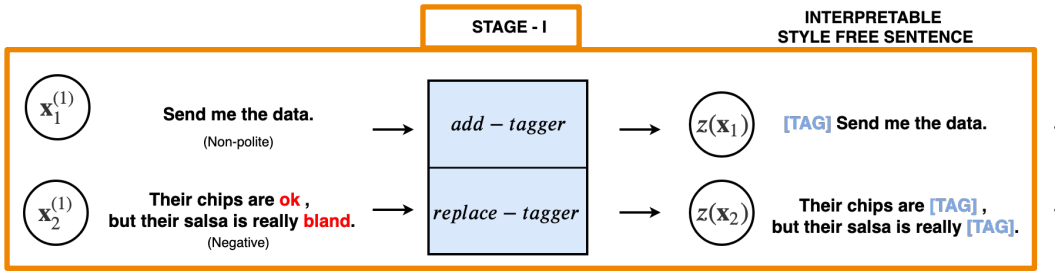
1. Successful transfer into target style.
2. Retaining content words (non-attribute markers).

[TAG] **Send me the data.**

Their chips are [TAG] **,**
but their salsa is really [TAG]**.**

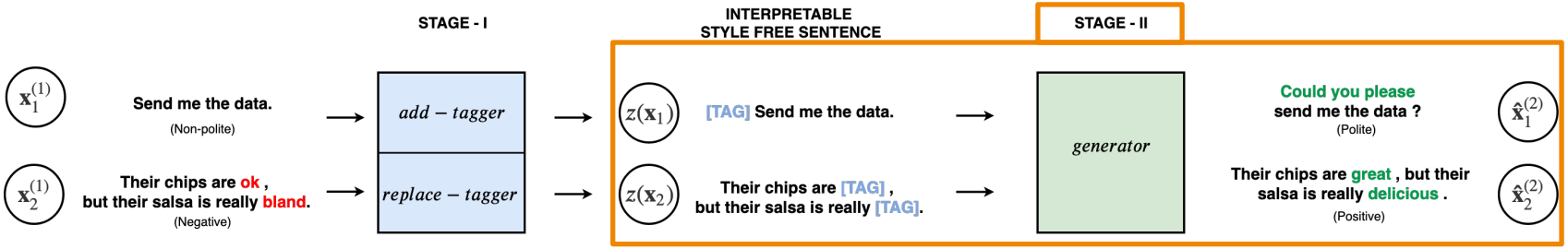
Tag and Generate Pipeline

Use tagger to TAG words indicative of the source style and get a **style neutral representation**.



Tag and Generate Pipeline

GENERATE *context appropriate* phrases
in the **target style**.



Creating Artificial Data for Training Tagger

$$\{\mathbf{x}_i^{(2)} \backslash a(\mathbf{x}_i^{(2)}) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\} \longrightarrow \{z(\mathbf{x}_i) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$$

(Input) (Output)

Step-I : Remove attribute markers

1. The chips are ok but their salsa is really bland

2. The service the last time I went was just terrible.

1. The chips are _____ but their salsa is really _____.

2. The service the last time I went was just _____.

Creating Artificial Data for Training Tagger

$$\{\mathbf{x}_i^{(2)} \backslash a(\mathbf{x}_i^{(2)}) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\} \longrightarrow \{z(\mathbf{x}_i) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$$

(Input) (Output)

Step-II : Generate Tags

1. The chips are _____ but their salsa is really _____.
2. The service the last time I went was just _____.

1. The chips are [TAG2] but their salsa is really [TAG3].
2. The service the last time I went was just [TAG3].

Creating Artificial Data for Training Generator

Use attribute markers of the style target to generate artificial parallel data.

1. The chips are [TAG2] but their salsa is really [TAG3].
2. The service the last time I went was just [TAG3].

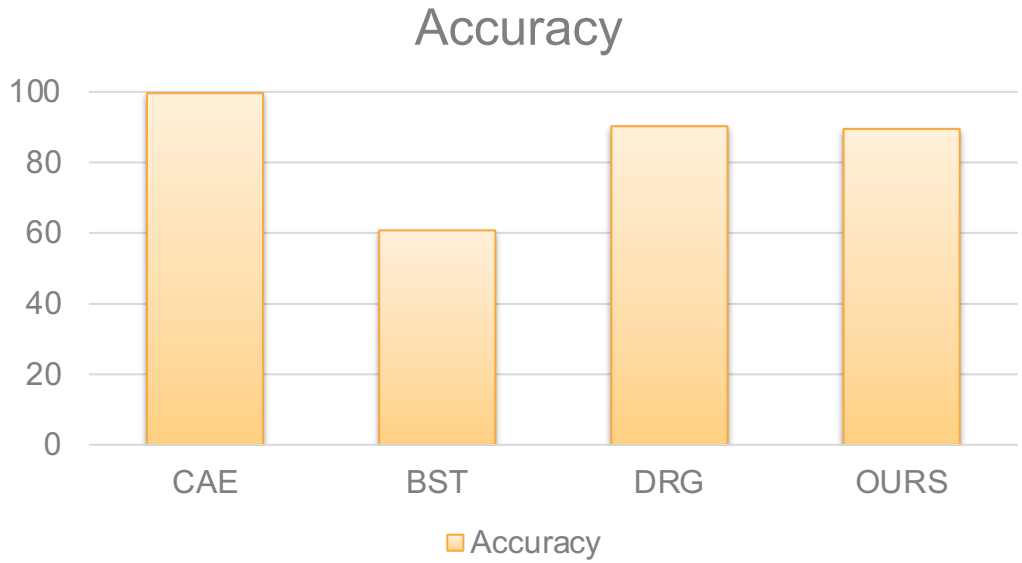
1. The chips are great but their salsa is really delicious.
2. The service the last time I went was just awesome.

Outline

- Problem Statement: Politeness Transfer
- Proposed Methodology
- Experiments and Results

Automatic Evaluation on Politeness

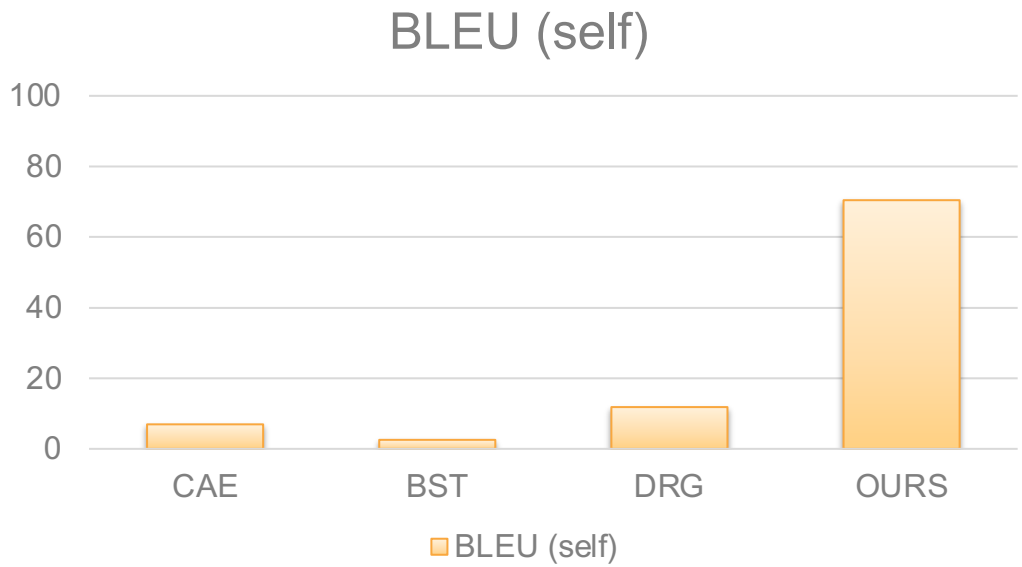
What % of the outputs are polite?



CAE: Shen et. al., 2017 BST: Prabhumoye et. al., 2017 DRG: Lo et. al., 2018

Automatic Evaluation on Politeness

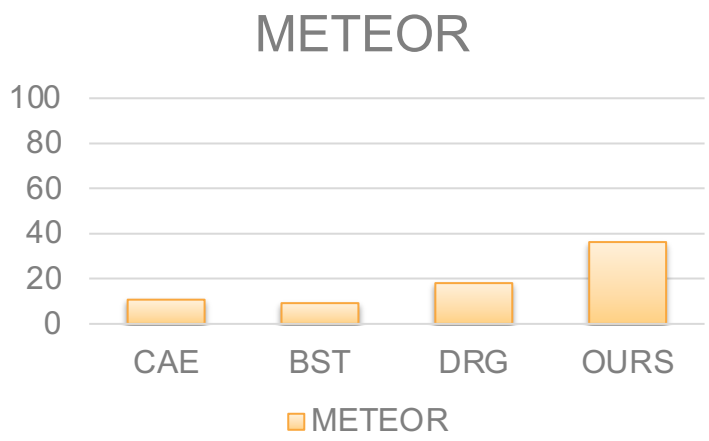
How much of the input content is preserved?

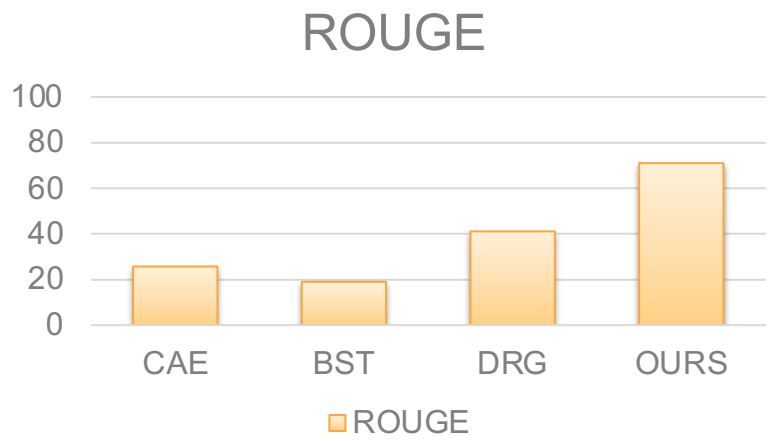


CAE: Shen et. al., 2017 BST: Prabhumoye et. al., 2017 DRG: Lo et. al., 2018

Automatic Evaluation on Politeness

How much of the input content is preserved?

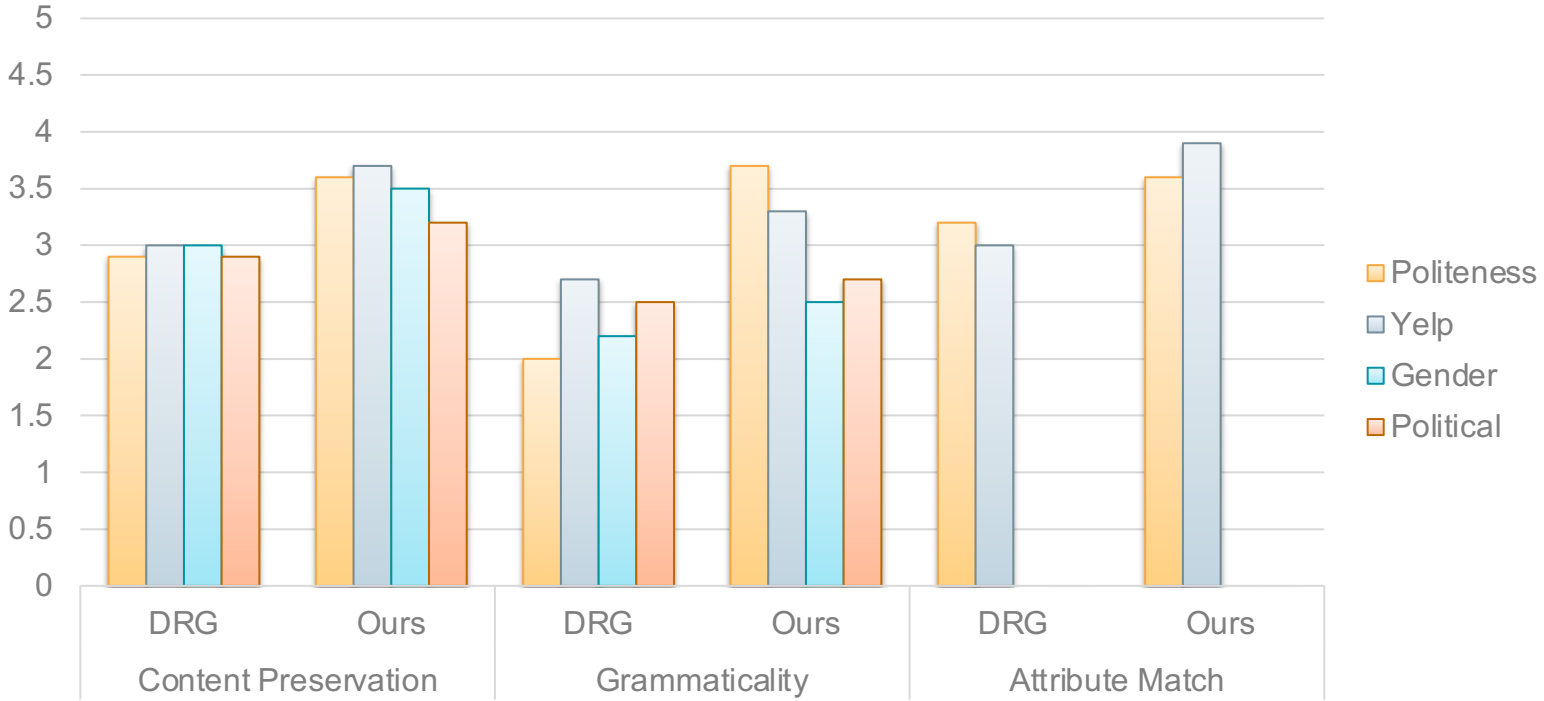




Results on Gender, Yelp Reviews, Political Slant, Amazon Reviews, and Captions in the paper!

CAE: Shen et. al., 2017 BST: Prabhumoye et. al., 2017 DRG: Lo et. al., 2018

Human Evaluation



Qualitative Examples

Input	Output	Politeness Strategy

Qualitative Examples

Input	Output	Politeness Strategy
what happened to my personal station?	could you please let me know what happened to my personal station?	Counterfactual modal

Qualitative Examples

Input	Output	Politeness Strategy
what happened to my personal station?	could you please let me know what happened to my personal station?	Counterfactual modal
yes, go ahead and remove it.	yes, we can go ahead and remove it.	First person plural

Qualitative Examples

Input	Output	Politeness Strategy
what happened to my personal station?	could you please let me know what happened to my personal station?	Counterfactual modal
yes, go ahead and remove it.	yes, we can go ahead and remove it.	First person plural
not yet-i'll try this wkend.	sorry not yet - i'll try to make sure this wk	Apologizing

Qualitative Examples

Input	Output	Politeness Strategy
what happened to my personal station?	could you please let me know what happened to my personal station?	Counterfactual modal
yes, go ahead and remove it.	yes, we can go ahead and remove it.	First person plural
not yet-i'll try this wkend.	sorry not yet - i'll try to make sure this wk	Apologizing
please check on metromedia energy,	please check on metromedia energy, thanks	Mitigating please start

Summary

- We introduce the task of **politeness transfer** and release **relevant dataset**.
- We propose an interpretable two-staged approach for content preserving style transfer – ***tag & generate.***
- Automatic and human evaluation shows that our approach **outperforms other state-of-the-art** models on content preservation metrics while retaining or improving the transfer accuracies.
- Code, data, and outputs at https://github.com/tag-and-generate

Thanks!