Detecting Model Discrepancy Through Sentence Generation

Junyu Guo[1] ,Kaiwei Chang[2] , Nanyun Peng[3]

[1]*School of Mathematics, Sun Yat-sen University*
[2]*Dept. of Computer Science, UCLA*
[3]*Information Science Institute, USC*

Introduction

Motivation

As more machine learning models become available and be applied to downstream tasks as black box, it is important to know how they function differently.

Two Factors Contributing to Model Discrepancy:

- Algorithm Design:
  base model structure, optimization algorithm…
- Training Data:
  population, subset, biased and unbiased data…

Approaches toward Detecting Model Discrepancy:

- Traditional method
  - Statistical frequency analysis
  - Instance tests
- Proposed method
  - Automatic generation of discrepant instances, solving data insufficiency problem and reducing time wastage

Scenario

Based on yelp dataset
Gender annotated restaurant reviews
Categorical sentiment scores ranging from 1 to 5

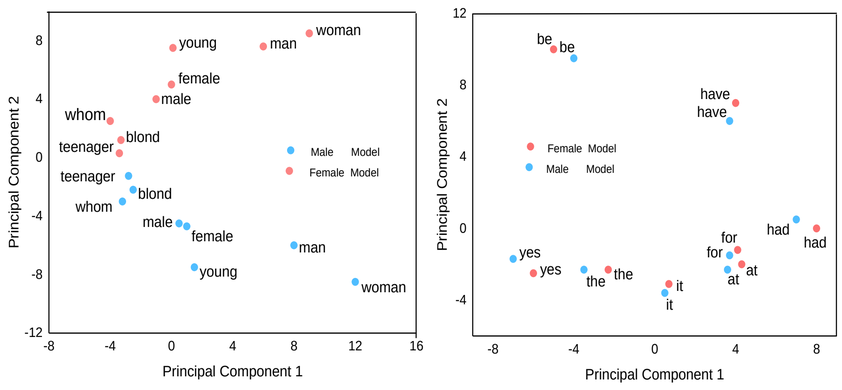
Goal

Given two models (in our scenario sentiment prediction models trained on different gender datasets with identical structure), we want to generate sentences which have very different predictions on two models and use these sentences to characterize model discrepancy. In this way we can also learn where exactly the relationship between expressions and sentiments differs between genders.

Understanding Model Discrepancy

1. Word Embedding Model

Training two Gensim Word2Vec Models on female and male corpus to get **F W2V Model** and **M W2V Model**



PCA analysis shows that gender related words have some symmetrical pattern and are further away from its counterpart in different models. Gender non-related words don't show this pattern and are close to their counterparts.

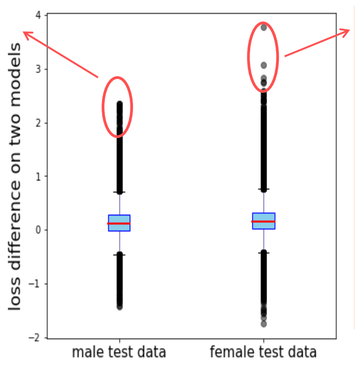
2. Awd-Language Model

Training two LSTM-RNN language models on female and male corpus to get F Model and M Model. Using test set from two groups to evaluate loss on two models.

$$loss = \frac{1}{N}\{-\log \mathrm{P}(\omega_1, \dots, \omega_N)\} \quad , \mathrm{P}(\omega_1, \dots, \omega_N) \approx \prod_{i=1}^{N} \mathrm{P}\left(\omega_i \middle| \omega_{i-(n-1)}, \dots \omega_{i-1}\right)$$

Avoid this place like the plague.
The service is second to none.
I love tilted kilt.
Stay away at all costs.
My wife had the cheeseburger.
This is a complete tourist trap!
I am not a bbq gourmet by any stretch of the imagination.

sentences with highest f-m loss difference on male test data



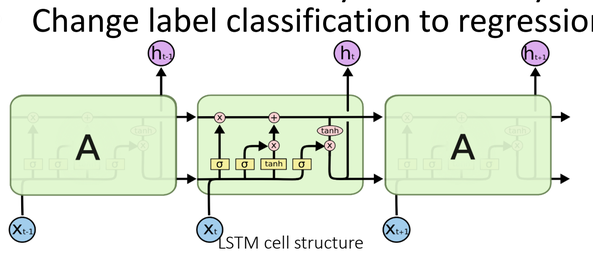
Love love love this place!
I am sooooo addicted!
Oh my gosh yum.
Yum yum yum yum!
Buy one get one free!
Antispasto was delish!
Um, delicious!
I love these bagels.

sentences with highest m-f loss difference on female test data

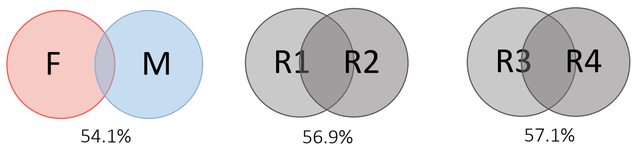
Sentiment Prediction Model Analysis

1. Main Structure

- Pre-trained glove embedding
- LSTM-RNN cell and fully connected layers
- Change label classification to regression



2. Prediction Accuracy distribution



The three Veen Diagram above shows the intersection of correct sentiment predictions of different models

- each circle denotes all the correct predictions of that model
- text in each circle indicates the training dataset of that model:
  F: female training data
  M: male training data
  Rx: random-split data with random seed x

The intersection between models trained on two distinct genders has smaller area than any other two random-split models. This indicates there exists observable difference between gender models

3. Counter Examples

Sentence	Fscore	Mscore	label
"Dude, this is a good deal!"	1.24	4.51	5
"this place has absolutely no sense of urgency. it shouldn't take an hour to get 5 balloons."	0.96	4.3	1
"place sucked, go to Carolinas. meats taste like can and they can't follow simple order."	4.64	1.03	1

Sentence Level

These donuts are absolutely ridiculous! Individually handmade to order, perfect presentation, fresh and hot. I never thought such artistic excellence could be applied to a mere donut. I drive 30 minutes each way for these beautiful babies! (true label: 5)

Fscore: 4.93 Mscore: 1.21

Fscore: -0.81 ↓ Mscore: +3.52 ↑

Phrase Level

……fresh and hot. I never thought such artistic excellence could be applied to a mere donut. ~~I drive 30 minutes each way for these beautiful babies!~~

Fscore: 4.12 Mscore: 4.73

Fscore: +0.04 ↑ Mscore: +3.71 ↑

Token level

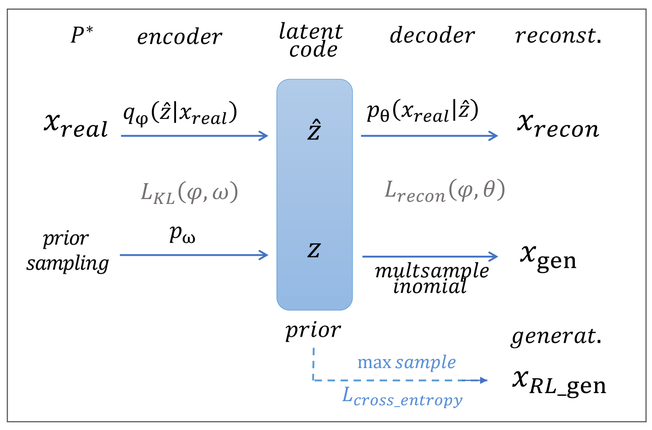
These donuts are absolutely ~~ridiculous~~! Individually handmade …… applied to a mere donut. I drive 30 minutes each way for these beautiful ~~babies~~!

Fscore: 4.97 Mscore: 4.92

Manually detecting sentence components contributing to major model prediction change

Sentence Generation

1. Reinforced Variational Autoencoder



2. Descriptive Objective

For VAE Part:

- Maximize likelihood between reconstruction sentences and real sentence distribution
- Minimize KL distance between encoded latent code z and prior z

For Reinforcement Part:

- Maximize likelihood of generating desired sentences with big prediction difference on two models (shifting data distribution)

3. Formulated Objective

$$Training \begin{cases} x_{recon} \sim \mathrm{Recon}(\hat{z}) = P_\theta\left(x_{recon}|\hat{z}\right) = \prod_t (\hat{x}_t|\hat{x}^{<t}, \hat{z}) \\ (\hat{x}_t - softmax(o_t/\tau), o_t - logit\ vector, \tau - temperature) \\ \hat{z} \sim Encode(x_{\mathrm{real}}) = q_\varphi(\hat{z}|x_{real}) \end{cases}$$

$$Evaluation\ \&\ Generation \begin{cases} z \sim Sample\ Prior P_\omega \\ (unit\ spherical\ /\ standard\ normal\ ) \\ x_{\mathrm{gen}} \sim Decode(z) \end{cases}$$

Loss Function:

$$\mathrm{L_{VAE}}(\varphi, \omega, \theta) = \mathrm{L_{recon}}(\varphi, \theta) + L_{KL}(\varphi, \omega)$$
$$= -\mathrm{E}_{q_\varphi(z|x)}[\log \mathrm{P}_\theta(\mathrm{x}|\mathrm{z})] + \mathrm{KL}(\mathrm{q}_\varphi(\mathrm{z}|\mathrm{x})||\mathrm{p}(\mathrm{z}))$$

$$\mathrm{L_{RL}}(\theta^*) = -E_{p_\omega(z)}[log\ P_\alpha(x|z)]$$

$log\ P_\alpha(x|z)$ refers to desired sentence distribution, favoring sentences with big prediction difference between the two sentiment models

Some Generated Examples:

Generated Examples	Fscore	Mscore
"This place is closed."	4.96	1.23
"Went there last night. Eating leftovers today! A lot of food. Food in a Mexican food."	1.90	4.84

Conclusions

Contributions

- Bringing up concept of automatic detection for model discrepancy.
- Proposing sentence generation as a method to achieve this goal.
- Similar to previous work of Style Transfer, where people usually use binary labels for added features and transfer sentence style by training a discriminator/critic simultaneously. In our case, the added feature becomes continuous which can't be achieved by binary discriminator and reinforcement approach is introduced to solve the problem

Future Plan

- Carrying out experiments to quantitatively evaluate the quality of generated sentences
- A more challenging goal would be, given an existed sentence with very different prediction scores from the two models, can we make a slight change to the original sentence to get similar scores?

References

- Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In ICML, 2017.
- Junbo, Zhao, Kim, Yoon, Zhang, Kelly, Rush, Alexander M., and LeCun, Yann. Adversarially Regularized Autoencoders. arXiv:1706.04223 [cs], June 2017. URL http://arxiv.org/abs/1706.04223. arXiv: 1706.04223.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically Equivalent Adversarial Rules for Debugging NLP Models. In ACL, 2018.
- Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," arXiv preprint arXiv:1710.11342, 2017.
- Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation.