COL772: Project Proposal

Suyash Agrawal (2015CS10262)
Aman Agrawal (2015CS10210)

March 20, 2018

1 Goal

Our aim is to create a model that can learn to generate reply for tweets. Further, we would also experiment with other domains like email and dialogues.

2 Dataset

We plan to use the Marsan Mas curated twitter corpus of $800,000$ context-responses[1] for our baseline model testing. Later, we can actually use the twitter streaming API to mine tweets and their corresponding replies to create our own dataset in a day or so.
We can later extend the dataset by merging it with the Enron Email Dataset, which contains about $500k$ emails of 150 employees of Enron Corporation[2].
If, time permits, we can concatenate the Cornell Movie-Dialogs corpus [3] in addition to the Twitter corpus to train a better model.

3 Literature

Recently, google released a paper "Smart Reply"[4], which is their model for email auto responder. Apart from this, there are various papers on text summarization[5] where we plan to take inspiration from.
We have not found a paper that exactly does twitter replies and thus we would be writing code from scratch.

4 Evaluation

For quantitative evaluation, we are currently planning on using the perplexity as our loss measure. The gold labels, would be a held out test data from the whole dataset. We would also be reporting the BLEU, METEOR and ROGUE metric of our predictions.
For qualitative evaluation, we can just look at some randomly sampled results and judge their quality.

References

[1] https://github.com/marsan-ma/chat_corpus.

[2] Enron email dataset: https://www.cs.cmu.edu/ enron/.

[3] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[4] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, and Vivek Ramavajjala. Smart reply: Automated response suggestion for email. *CoRR*, abs/1606.04870, 2016.

[5] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.