Final Project: Disentangled Conversation Embeddings

Emily Tseng
et397@cornell.edu

March 23, 2020

1 Area

Automated analysis of social media conversation is of increasing importance to such pressing social problems as the radicalization of young people towards hate-fueled violence, the congregation of intimate partner abusers in relationship forums, or the spread of misinformation during a deadly pandemic. Yet, the speed and scale of conversation on Facebook, Twitter, Reddit and other online discussion platforms mean human efforts to grasp and moderate what is being said on the Internet are, to put it mildly, intractable. At hand are the entwined problems of *analysis* and *detection*—we want methods to quickly consume ever-increasing quantities of information, and in parallel methods that flag or risk-score worrying situations for human intervention. Keyword-based detection suffers from a lack of scalability, needs constant updating, and struggles to capture conversation-level dynamics; simultaneously, supervised learning approaches suffer from a need for extensively labeled datasets and a lack of cross-domain generalizability.

Recent advances in representation learning offer some promise. Consider the common procedure of unsupervised *pre-training* of structured representations of inputs from massive amounts of data, coupled with *fine-tuning* of those representations in a model built to a specific supervised task, e.g. predicting ratings from reviews. Chang and Danescu-Niculescu-Mizil (2019) showed that *conversation embeddings* built with this approach could forecast the derailment of online conversations with a substantial improvement in accuracy over existing rule-based baselines, suggesting that the pre-trained representations were learning semantics not captured previously. However, as with many models that rely on neural representations, the approach suffers from a lack of human interpretability: while they showed the embeddings were able to capture conversational ordering, they did not explore further what, precisely, about a conversation was encoded.

In this project, we propose to explore development of an interpretable and structured neural topic model for online conversations. We will begin by performing a set of experiments with the embeddings generated in Chang and Danescu-Niculescu-Mizil (2019) to understand what conversational dynamics are captured by the form. (In this, we will follow some of the analysis methods outlined in Clark et al. (2019)'s examination of BERT's attention). Using learnings from this effort, we will then investigate the extension of this conversational embedding method with structured aspect learning, as applied in Esmaeili et al. (2019). Our hope is to achieve a model capable of producing disentangled conversational representations with predictive power on downstream tasks.

Using the ChangeMyView dataset (Chang et al., 2019), we will then visualize the interpretability of the new model by highlighting what it learns as topics, and show its predictive power on the

derailment prediction task reported in Chang and Danescu-Niculescu-Mizil (2019). We will then show the flexibility of the approach for unsupervised exploration of conversations by highlighting the aspects and topics learned on a novel dataset of online discussions of potentially abusive relationships. Taken together, our work will highlight the promise of representation learning for conversational datasets, with an eye towards assisting human moderators seeking to understand behavior in online discussions and, importantly, to capture their unsavory parts.

2 Papers

- Major paper: Chang and Danescu-Niculescu-Mizil (2019) Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop
- Minor paper 1: Serban et al. (2016) Building end-to-end dialogue systems using generative hierarchical neural network models
- Minor paper 2: Esmaeili et al. (2019): Structured Neural Topic Models for Reviews
- Minor paper 3: Clark et al. (2019): What Does BERT Look At?

3 Baseline

Chang and Danescu-Niculescu-Mizil (2019).

4 Team and Time

My team thus far consists of just me: Emily Tseng, et397@cornell.edu. I signed up to present on April 20th.

References

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., and Danescu-Niculescu-Mizil, C. (2019). Convokit: The cornell conversational analysis toolkit. Retrieved from http://convokit.cornell.edu.

Chang, J. P. and Danescu-Niculescu-Mizil, C. (2019). Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of EMNLP*.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Esmaeili, B., Huang, H., Wallace, B., and Meent, J.-W. v. d. (2019). Structured neural topic models for reviews. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 3429–3439. PMLR.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.