Title: Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances
Authors: Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Fei, Jie Zhou
Affiliations: Key Laboratory of Intelligent Information Processing; Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS); University of Chinese Academy of Sciences; Pattern Recognition Center, WeChat AI, Tencent Inc, China

Problem

Many open-domain dialogue models are designed such that the concatenated dialogue history is directly passed as input to the model. The authors refer to this design as the *flat pattern* and claim that it ignores how information dynamically flows between utterances of a dialogue. In simpler terms, the authors' goal is to make a dialogue model that appropriately considers how much semantic and contextual information has been contributed by each message in a dialogue history as opposed to simply passing the dialogue history as input to the model.

Prior Work

The authors categorize earlier dialogue models into two groups. The more common of the two is the aforementioned flat pattern models that simply take the concatenated dialogue history as input. This has been shown in other research to ignore the contextual dynamics of a conversation across the dialogue history. Essentially, this reduces a model's ability to adapt to the shifting contexts of a conversation. However, this has been made up for in the past via the development of large-scale pre-training approaches and access to large amounts of conversational data. Most of the recent advances in chatbot technology can be attributed to these two factors. The authors specifically cite DialoGPT as a flat pattern model that has benefitted from large amounts of pre-training and data, and they use DialoGPT later in the paper as a baseline for evaluating their own work. The second group of dialogue models is that of hierarchical models. In these models, each utterance in the dialogue history is separately encoded as input to the model. According to the authors, this approach ignores information regarding the history of the conversation.

Contributions of Paper

The authors propose the DialoFlow model to model how information flows across utterances in a dialogue history. At a high level, DialoFlow is a combination of three separately trained models that perform the following steps: *context flow modeling* to model how conversational context is passed between utterances, *semantic influence modeling* to model how much each utterance impacts the context of the conversation, and finally *response generation modeling* to produce an output response to the provided dialogue history.

The authors define the context of a dialogue at the *k*th utterance as some dense representation of the dialogue history up to that point. In other words, the dialogue history up to a certain utterance is represented as a vector that is not mostly composed of zeroes. The semantic influence of the *k*th

utterance is defined as the vector-difference between the context at the *k*th utterance and the context at the *(k+1)*th utterance. DialoFlow produces responses by first taking the context calculated at all previous utterances and predicting what the next context vector will be. Using the prediction for the new context, the model simply calculates what the semantic influence for its output response should be. Then, the model takes the predicted semantic influence and the dialogue history to produce an output response that would result in that semantic influence.

DialoFlow also uses slightly different training objectives. Traditional training approaches train dialogue models using context-response pairs, where models get only the immediate context as input. DialoFlow instead opted to use the entire dialogue history as input. Each model in DialoFlow is also trained using its own loss function. For context flow modeling, the loss function minimizes the distance between the predicted context and the actual context. For semantic influence modeling, the loss function is based on maximizing the probability of seeing particular words in an utterance given the semantic influence of that utterance. Finally, for response generation modeling, the loss function is based on the probability of seeing a particular utterance given the semantic influence of that utterance and the contents of all previous utterances in the dialogue.

The authors propose their own metric for chatbot dialogue evaluation called the Flow Score. The Flow Score is a measure of the perplexity of the dialogue with respect to the semantic influence of the chatbot's utterances. Perplexity refers to a model's ability to make predictions. For each utterance made by the chatbot, a similarity measure is calculated between the predicted semantic influence of that utterance and the actual semantic influence of the chatbot's output response. These similarity measures are combined to create the Flow Score for the entire dialogue. A lower Flow Score indicates that the chatbot is more human-like and produces higher-quality responses.

Evaluation

The authors chose DialoGPT pre-trained on OpenAI GPT-2 as the baseline for evaluating DialoFlow. The performance of DialoFlow generally increased as the size of the model increased, whereas DialoGPT performed the best with a medium-sized model. When evaluated using the NIST metric, which penalizes the use of common phrases such as "I don't know," DialoFlow performs better than DialoGPT. This would indicate that DialoFlow tends to produce less general responses than DialoGPT. Furthermore, this demonstrates that modeling semantic flow does indeed improve the chatbot's ability to create more relevant and informative responses. Entropy, which measures lexical diversity, is similar between both DialoFlow and DialoGPT. The authors also performed human evaluation, where prompt-response pairs from both DialoFlow and DialoGPT were randomly presented to human judges tasked to pick the more human-like response. In about 25% of cases, judges stated DialoFlow was about as human-like as DialoGPT, and in about 46% of cases, judges stated that DialoFlow was more human-like. Across different metrics, DialoFlow generally performed better than DialoGPT when operating on long dialogue histories.

The proposed Flow Score metric was evaluated by measuring its correlation to human ratings of chatbot performance. Correlation was measured via Pearson correlation and Spearman correlation. Where other metrics show a low to moderate correlation with human ratings using both measures, Flow Score demonstrates a Pearson correlation of .91 and a Spearman correlation of .9. This shows that the authors' proposed Flow Score is a good metric for measuring the quality of a chatbot's responses.

Citations

The paper itself has 22 citations on Google Scholar. Among the authors, Zekang Li has 418 citations, Jinchao Zhang has 727 citations, Zhengcong Fei has 126 citations, Yang Feng has 1372 citations, and Jie Zhou has 21836 citations. Of these, Jie Zhou is the most cited author.

Conclusion

The paper is important for two major reasons. First, the authors discovered a new way to structure dialogue models to produce more human-like output. By training three separate models to perform distinct tasks – context prediction, semantic influence prediction, and response generation – the authors improved on existing solutions that were prone to making responses that were too general and did not carry information in a manner natural to human conversation. Second, the authors created the Flow Score, a new metric for evaluating chatbots. Flow Score closely correlates with how human judges tend to score chatbot responses while still being completely automatic, so future researchers can evaluate their own research with Flow Score to get a cursory measure of how human-like their chatbots are.