DERP : Deep Evaluation for Response Predictors
Project Proposal - COL864

Akshay Kumar Gupta **Shantanu Kumar** **Surag Nair** **Barun Patra**
Indian Institute of Technology, Delhi
{cs5130275,ee1130798,ee1130504,cs1130773}@iitd.ac.in

Abstract

Automatic evaluation of dialogue response generation systems has been a fundamentally difficult task faced by researchers in the field. It has been shown that most automatic metrics that are used either do not correlate or correlate very weakly with human scoring of a dialogue system. We present a proposal for a novel automatic method of evaluation. We hope that this method addresses the issues faced by traditional evaluation systems, and aligns better with human scoring. We discuss the datasets which we will utilize for this task, and our approach, along with how we intend to evaluate our evaluation system. Finally we present a tentative timeline for the same.

1 Introduction

Traditional response generation systems are trained to produce a relevant and fluent response given a set of previous utterances as context. An important area of consideration to create any good dialogue system is how to evaluate such a dialogue system. We focus on unsupervised dialogue generation systems, which do not rely on human labels for training. These systems are gaining importance in recent times with the advent of neural models (Serban et al., 2016)(Sordoni et al., 2015)(Vinyals and Le, 2015) that can be trained without the need for task specific supervised labels. However, how to automatically evaluate these systems is still an open area of research. The unreliability of existing metrics means that researchers are forced to resort to human evaluation, which is potentially slow and expensive, and which can act as a roadblock to fast-paced research.

In this report, we propose a new deep learning based evaluation metric for dialogue systems, given a context and ground truth response. We plan to compare this metric to human evaluators, and hope to show better correlation than that of existing metrics. This work is important because if successful, it would simplify and accelerate research in this area.

2 Related Work

A number of automatic metrics have been previously used to evaluate unsupervised dialogue systems. The most often used is BLEU (Papineni et al., 2002), which has traditionally been used for machine translation, but has also been used for dialogue (Ritter et al., 2011). Recently proposed deltaBLEU (Galley et al., 2015), a modified version of BLEU, which correlated moderately with human annotators. They generate triples of Twitter conversation, with multiple ground truths per context, and using human annotators to rate these ground truths, they develop a test set for any dialogue system to test on. However, it is often not possible to obtain such human annotation, and their test set itself is limited to Twitter, so cannot be used to test goal-driven responses (such as on the Ubuntu (Lowe et al., 2015) corpus) nor can it be used to test responses in the presence of long contexts.

(Liu et al., 2016) conducted a survey wherein they evaluated embedding based metrics like greedy matching and word overlap based metrics like BLEU, ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) for dialogue systems. They showed that these metrics correlate very weakly or not at all with human evaluators, while human evaluators correlate well with each other. The main reason is that there can be significant variation in possible responses that cannot be

captured by word overlap metrics.

Additionally, perplexity has also been used as a metric (Serban et al., 2015), but perplexity is not a model-independent metric and cannot be used to rate individual responses. Finally, we do not consider retrieval based metrics such as re-ranking and precision-recall because we are looking at response generation.

3 Approach

Our proposed approach is to train a model which, given the context and the ground truth response, gives a score to the response generated by a dialogue system. This model is explained in detail in Section 3.2. In order to train the model, we need a dataset which has multiple valid responses for a given context. We propose an unsupervised automatic extraction method to obtain such a dataset (Section 3.1).

3.1 Data Extraction

Given a pair of semantically similar conversational contexts, we hypothesize that the responses to these contexts are interchangeable i.e they can be considered valid responses to the context. We intend on using Skip-Thoughts (Kiros et al., 2015) sentence-level embedding vectors on the Reddit Comments Data to obtain similar conversational contexts. This task would yield 4-tuples of the form `(Context, Gold Response, Alternate Response, Score)`, which can then be used as training data for our evaluation system.

3.2 Model

The model we intend on using for this task of response evaluation is a Siamese-style Recurrent Neural Network model. A recurrent architecture would be used to encode the context first and then read the input response, which may attend on the context during encoding. The Siamese architecture, trained on the data obtained from the data collection task, would be used to score the similarity between the ground truth response and proposed response for the same context. The Siamese architecture is trained with pairs of responses (4-tuples) extracted in the above task, with the aim of encoding them in a space where valid responses for the given context are similar.

3.3 Evaluation

The aim of the project is to develop an evaluation metric that agrees with the human annotations for the dialogue system responses. Thus we would evaluate our evaluation system based on its correlation with human annotations. In particular, we would compute the Pearson correlation, which estimates linear correlation, and Spearman correlation, which estimates any monotonic correlation (as used by (Liu et al., 2016)).

4 Datasets

4.1 Reddit Comments Dataset[1]

This is a 250 GB public dataset of Reddit comments. We currently possess a 1 month 5GB subset of this data (54 million comments) which we will use for data extraction for training and evaluation. Each comment has a link to its parent and the subreddit it belongs to.

4.2 Quora Similar-Questions Dataset[2]

This dataset contains a set of over 400,000 question pairs and a 0 or 1 label indicating whether the questions are the same or not. An example pair is "What is the most populous state in the USA?" and "Which state in the United States has the most people?" We might use this data to pre-train our context matching network.

4.3 Twitter Alternate Responses Dataset

(Galley et al., 2015) built a multi-reference dataset using 29M Twitter context-message-response triples, wherein the message of the triples was matched using bag-of-words similarity function and all the responses from the matched triples were assigned to one context-message pair as alternate responses. However, this dataset was further hand labeled by human annotators to give a score to each alternate response and was thus a very small dataset comprising of only 4232 (high quality) triples. We will use this either as a small high quality training set or as an additional way to evaluate our system.

5 Budget

Given that we require human annotation for the evaluation of our system, we are planning to post a

[1] https://redd.it/3bxlg7

[2] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

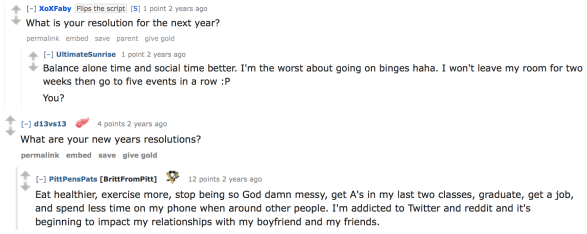

Figure 1: Two Reddit comments whose responses can be interchanged

task on the Amazon Mechanical Turk for the same. A budget of $100 should be sufficient for the task.

6 Timeline

The project will take part in two stages. Stage one would be the data collection task, while stage two would be the model building task. We intend on completing stage one by the end of February and Stage 2 by April.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. volume 29, pages 65–72.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863* .

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. pages 3294–3302.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, volume 8.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* .

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909* .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 583–593.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. *arXiv preprint arXiv:1507.04808* .

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069* .

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714* .

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* .