Ryan Lowe

McGill University
845 rue Sherbrooke
Montreal, Canada

ryan.lowe@cs.mcgill.ca
cs.mcgill.ca/~rlowe1

1 Research Interests

My primary research interests lie in the application of neural networks to building end-to-end dialogue systems. *End-to-end* refers to the fact that all the parameters of the model are jointly learned using a single objective function. This contrasts to the typical modular outlook on dialogue systems, where modules such as the natural language interpreter, dialogue state tracker, and natural language generator are learned separately. Note that we pursue this goal mostly in the context of chatbot-style systems, where the user has no specific goal.

One of the first to use such a paradigm in the context of neural networks was Ritter et al. (2011), who applied ideas from machine translation to predict the next utterance of a conversation on Twitter. Subsequent works have built upon this idea — now, more and more complicated neural network architectures are trained using maximum likelihood to predict the next utterance in a dialogue. The idea is that if a model assigns a high probability to the actual next utterance in a dialogue, it should be able to carry out a conversation reasonably well.

While I have been involved with creating some novel architectures for this problem (Lowe et al., 2015a; Serban et al., 2016), most of my work has focused on building *large datasets* that can be used to train such architectures, and finding ways to *evaluate* such architectures automatically when task completion rates are not available (as the user has no set ‘task’ to accomplish).

1.1 Large Dialogue Datasets

We first introduced the Ubuntu Dialogue Corpus (Lowe et al., 2015b), a large dataset of over 500,000 dialogues with an average of 7 turns. These dialogues consist of users on the Ubuntu IRC channel who are attempting to troubleshoot problems with their Ubuntu system. Other users respond and help the users to solve their problem based on their experience. While the original dataset is multi-party, we use a set of heuristics to disentangle the chat into dyadic (2-person) conversation. This dataset may be particularly useful for training models for technical customer service.

We also thoroughly surveyed the literature for existing datasets that could be used for building end-to-end dialogue systems (Serban et al., 2015), and found that there is a lack of very large datasets that can be used to train neural network models.

1.2 Dialogue Evaluation

Another direction of work is in trying to evaluate dialogue models automatically. While this has been well studied in the domain where task completion rates are given (such as the PARADISE framework (Walker et al., 1997)), it has been considerably less studied when such signals are not available. We proposed a recall-based method for evaluating these systems, called next utterance classification (NUC) (Lowe et al., 2015b), and analyzed how well humans performed on the task (Lowe et al., 2016). NUC evaluates how well a dialogue model can select the correct next utterance in a conversation, given a list of candidate utterances. This is a task that humans can perform quite well, while dialogue models still have some trouble, depending on the dataset.

We then investigated methods to evaluate the quality of responses generated by a model, given the context of a conversation. In particular, we examined the recent trend of using BLEU scores from machine translation to evaluate the quality of a response by comparing it to the ground-truth next utterance. We found that the BLEU score (as well as many other methods such as ROUGE, METEOR, and several word embedding based methods) correlate weakly or not at all with human judgement of the responses (Liu et al., 2016). Our current work is attempting to derive a system that can evaluate dialogue responses in a manner that correlates more significantly with human judgement of the response.

2 Future of Spoken Dialog Research

While there will always be a role for small dialogue systems in specialized domains, I believe that the next 5 to 10 years of dialogue research will bring about significant progress in building larger, end-to-end dialogue systems that can converse naturally about a variety of topics. Companies such as Google and Facebook are building large dialogue teams to tackle such problems using deep learning and reinforcement learning, and academia is sure to follow.

There are (at least) two large obstacles for such systems to be truly effective. The most immediate one is the response diversity problem — current neural network models tend to produce fairly generic utterances compared to the diversity of human responses. This could perhaps be overcome using better architectures, larger datasets, or moving away from maximum likelihood (perhaps using a generative adversarial framework (Goodfellow et al., 2014)).

The second problem is how to incorporate natural language reasoning, or 'common sense', into these conversational agents. This is a much more complex problem, and will almost certainly not be fully solved in the next decade. Indeed, there is no theoretical ceiling to how complex we can make the reasoning of such agents. Progress in this area is crucial if we are to build agents that can converse naturally with humans.

3 Suggestions for Discussion

- Methods for automatic evaluation of chatbot responses (in a setting where the user has no specific task to accomplish)
- Desiderata for large conversational datasets, and candidates for such datasets
- Can we build intelligent conversational agents (e.g. generally intelligent chatbots) using only conversational data? Do we need visual grounding or other sources of multi-modality?

References

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

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, Laurent Charlin, and Joelle Pineau. 2015a. Incorporating unstructured textual knowledge into neural dialogue systems. *NIPS Workshop on Machine Learning for Spoken Language Understanding*.

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

Ryan Lowe, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. *SIGDIAL*.

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*, pages 583–593. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

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*.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

Biographical Sketch

Ryan Lowe is a PhD student at McGill University, under the supervision of Joelle Pineau. Much of his work on dialogue is in collaboration with Iulian Serban, Yoshua Bengio, and Aaron Courville at the University of Montreal. He is also interested in reinforcement learning and causal models.

Ryan has previously worked at the Institute for Quantum Computing, the Max Planck Institute, and the National Research Council. His extra-curricular interests include rock climbing, reading fiction, writing, playing soccer, and hiking.