Neural Response Ranking for Social Conversation: A Data-Efficient Approach

Igor Shalyminov, Ondřej Dušek, and Oliver Lemon
The Interaction Lab, Department of Computer Science
Heriot-Watt University, Edinburgh, EH14 4AS, UK
{is33, o.dusek, o.lemon}@hw.ac.uk

Abstract

The overall objective of 'social' dialogue systems is to support engaging, entertaining, and lengthy conversations on a wide variety of topics, including social chit-chat. Apart from raw dialogue data, user-provided ratings are the most common signal used to train such systems to produce engaging responses. In this paper we show that social dialogue systems can be trained effectively from raw unannotated data. Using a dataset of real conversations collected in the 2017 Alexa Prize challenge, we developed a neural ranker[1] for selecting 'good' system responses to user utterances, i.e. responses which are likely to lead to long and engaging conversations. We show that (1) our neural ranker consistently outperforms several strong baselines when trained to optimise for user ratings; (2) when trained on larger amounts of data and only using conversation length as the objective, the ranker performs better than the one trained using ratings – ultimately reaching a Precision@1 of 0.87. This advance will make data collection for social conversational agents simpler and less expensive in the future.

1 Introduction

Chatbots, or *socialbots*, are dialogue systems aimed at maintaining an open-domain conversation with the user spanning a wide range of topics, with the main objective of being engaging, entertaining, and natural. Under one of the current approaches to such systems, the *bot ensemble* (Serban et al., 2017; Yu et al., 2016; Song et al., 2016), a collection, or ensemble, of different bots is used, each of which proposes a candidate response to the user's input, and a *response ranker* selects the best response for the final system output to be uttered to the user.

In this paper, we focus on the task of finding the best supervision signal for training a response ranker for ensemble systems. Our contribution is twofold: first, we present a neural ranker for ensemble-based dialogue systems and evaluate its level of performance using an annotation type which is often used in open-domain dialogue and was provided to the Alexa Prize 2017 participants by Amazon (Ram et al., 2017): per-dialogue user ratings. Second and most importantly, we explore an alternative way of assessing social conversations simply via their *length*, thus removing the need for any user-provided ratings.

2 Data Efficiency in Social Dialogue

2.1 The Need for Data Efficiency

It is well known that deep learning models are highly data-dependent, but there are currently no openly available data sources which can provide enough high-quality open-domain social dialogues for building a production-level socialbot. Therefore, a common way to get the necessary data is to collect it on a crowdsourcing platform (Krause et al., 2017). Based on the model type and the development stage, it may be necessary to collect either whole dialogues, or some form of human feedback on how good a particular dialogue or turn is. However, both kinds of data are time-consuming and expensive to collect.

The data efficiency of a dialogue model can be split into two parts accordingly:

- *sample efficiency* – the number of data points needed for the model to train. As such, it is useful to specify an order of magnitude of the training set size for different types of machine learning models;
- *annotation efficiency* – the amount of annotation

[1] Code and trained models are available at https://github.com/WattSocialBot/alana_learning_to_rank

Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd Int'l Workshop on Search-Oriented Conversational AI, pages 1–8
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics
ISBN 978-1-948087-75-9

Variables	Pearson corr. coefficient
rating/length	0.11
rating/positive feedback	0.11
rating/negative feedback	0.04
length/positive feedback	0.67
length/negative feedback	0.49

Table 1: Correlation study of key dialogue aspects

effort needed. For instance, traditional goal-oriented dialogue system architectures normally require *intent*, *slot value*, and *dialogue state* annotation (e.g. Young et al., 2010), whereas end-to-end conversational models work simply with raw text transcriptions (e.g. Vinyals and Le, 2015).

2.2 Alexa Prize Ratings

The 2017 Alexa Prize challenge made it possible to collect large numbers of dialogues between real users of Amazon Echo devices and various chatbots. The only annotation collected was per-dialogue ratings elicited at the end of conversations by asking the user *"On a scale of 1 to 5, how much would you like to speak with this bot again"* (Venkatesh et al., 2017). Less than 50% of conversations were actually rated; the rest were quit without the user giving a score. In addition, note that a single rating is applied to an entire conversation (rather than individual turns), which may consist of very many utterances. The conversations in the challenge were about 2.5 minutes long on average, and about 10% of conversations were over 10 minutes long (Ram et al., 2017) – this makes the ratings very sparse. Finally, the ratings are noisy – some dialogues which are clearly bad can get good ratings from some users, and vice-versa.

Given the main objective of social dialogue stated in the Alexa Prize rules as 'long and engaging' conversation, we tried to verify an assumption that user ratings reflect these properties of the dialogue. Apart from our observations above, we performed a correlation analysis of user ratings and aspects of dialogue directly reflecting the objective: dialogue length and explicit user feedback (see Table 1).

Although we have a significant number of dialogues which are both long and highly rated, the correlation analysis was not able to show any relationship between dialogue length and rating. Neither are ratings correlated with user feedback (see Section 6 for the details of user feedback collection). On the other hand, we found a promising moderate correlation between the conversation length and explicit positive feedback from users (specifically, the number of dialogue turns containing it). The respective length/negative feedback relationship is slightly weaker.

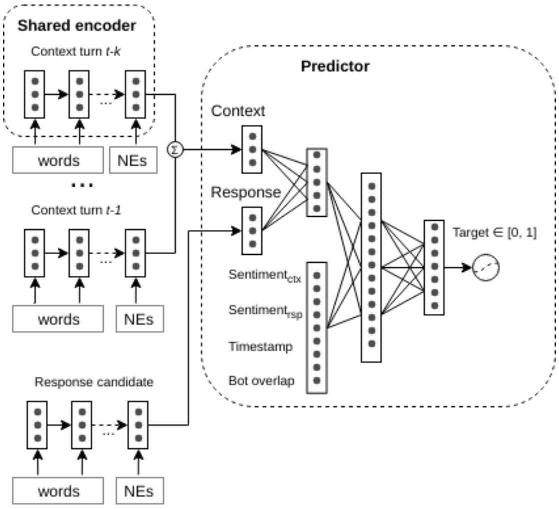

Figure 1: Neural ranker architecture

Therefore, we experiment with conversation length for approximating user satisfaction and engagement and use it as an alternative measure of dialogue quality. This allows us to take advantage of all conversations, not just those rated by users, for training a ranker. While some conversations might be long but not engaging (e.g. if there are a lot of misunderstandings, corrections, and speech recognition errors), training a ranker only using length makes it extremely *annotation-efficient*.

3 A neural ranker for open-domain conversation

The ranker described here is part of Alana, Heriot-Watt University's Alexa Prize 2017 finalist socialbot (Papaioannou et al., 2017). Alana is an ensemble-based model incorporating information-retrieval-based bots with news content and information on a wide range of topics from Wikipedia, a question answering system, and rule-based bots for various purposes, from amusing users with fun facts to providing a consistent persona. The rule-based bots are also required to handle sensitive issues which can be raised by real users, such as medical, financial, and legal advice, as well as profanities.

3.1 Ranker architecture

The architecture of our ranker is shown in Figure 1. The inputs to the model are 1-hot vectors of a candidate response and the current dialogue context (we use the 3 most recent system and user turns). They are encoded into a latent representation using a single shared RNN encoder based on GRU cells (Cho et al., 2014). The context embedding vectors are then summed up and concatenated with the response embedding (Eq. 1):

$$Enc(C, r) = \sum_i \mathit{RNN}(C_i) \oplus \mathit{RNN}(r) \quad (1)$$

where C is the dialogue context and r is a response candidate.

The context and the response are represented using combined word-agent tokens (where agent is either a specific bot from the ensemble or the user) and are concatenated with the lists of named entities extracted using Stanford NER (Finkel et al., 2005). All the word-agent tokens and named entities share the same unified vocabulary.

Encoder outputs, along with additional dialogue features such as context and response sentiment, timestamp, and bot names in the context and the response, go into the *Predictor*, a feed-forward neural network (MLP) whose output is the resulting rating (Eq. 2):

$$Pred(C, r) = \sigma(L(Sem(C, r) \oplus f(C, r))) \quad (2)$$

where: $L(x) = \mathit{ReLU}(Mx + b)$ is the layer used in the Predictor (the number of such layers is a model parameter),
$Sem = L(Enc(C, r))$ is the vector of semantic context-response features, and
$f(C, r)$ is a vector of the additional dialogue features listed above.

We use $ReLU$ activation for the hidden layers because it is known to be highly efficient with deep architectures (Glorot et al., 2011). Finally, we use sigmoid activation σ for generating the final prediction in the range $[0, 1]$.

3.2 Training method

We use either dialogue rating or length as the prediction target (as discussed in Sections 5 and 6). The model is trained to minimize the Mean Squared Error (MSE) loss against the target using the Adagrad optimizer (Duchi et al., 2011). In our training setup, the model learns to predict per-turn target values. However, since only per-dialogue ones are available in the data, we use the following approximation: the target value of a context-response pair is the target value of the dialogue containing it. The intuition behind this is an assumption that the majority of turns in “good” dialogues (either length- or rating-wise) are “good” in their local contexts as well – so that given a large number of dialogues, the most successful and unsuccessful turns will emerge from the corresponding dialogues.

4 Baselines

We compare our neural ranker to two other models also developed during the competition: *handcrafted* and *linear* rankers — all three were deployed live in the Alana Alexa Prize 2017 finalist system (Papaioannou et al., 2017), and were therefore of sufficient quality for a production system receiving thousands of calls per day. We also compare our model to a recently published *dual-encoder* response selection model by Lu et al. (2017) based on an approach principally close to ours.

4.1 Handcrafted ranker

In the handcrafted approach, several turn-level and dialogue-level features are calculated, and a linear combination of those feature values with manually adjusted coefficients is used to predict the final ranking. The list of features includes:

- coherence, information flow, and dullness as defined by Li et al. (2016);
- overlap between the context and the response with regards to named entities and noun phrases;
- topic divergence between the context turns and the response – topics are represented using the *Latent Dirichlet Allocation* (LDA) model (Hoffman et al., 2010);
- sentiment polarity, as computed by the NLTK Vader sentiment analyser (Gilbert and Hutto, 2014).[2]

4.2 Linear ranker

The linear ranker is based on the VowpalWabbit (VW) linear model (Agarwal et al., 2014). We use

[2] http://www.nltk.org/howto/sentiment.html

the MSE loss function and the following features in our VW ranker model:

- bag-of-n-grams from the dialogue context (preceding 3 utterances) and the response,
- position-specific n-grams at the beginning of the context and the response (first 5 positions),
- dialogue flow features (Li et al., 2016), the same as for the handcrafted ranker,
- bot name, from the set of bots in the ensemble.

4.3 Dual-encoder ranker

The closest architecture to our neural ranker is that of (Lu et al., 2017), who use a dual-encoder LSTM with a predictor MLP for task-oriented dialogue in closed domains. Unlike this work, they do not use named entities, sentiment, or other input features than basic word embeddings. Dialogue context is not modelled explicitly either, and is limited to a single user turn. We reproduced their architecture and set its parameters to the best ones reported in the original paper.

5 Training data

Our data is transcripts of conversations between our socialbot and real users of the Amazon Echo collected over the challenge period, February–December 2017. The dataset consists of over 200,000 dialogues (5,000,000+ turns) from which over 100,000 dialogues (totalling nearly 3,000,000 turns) are annotated with ratings. From this data, we sampled two datasets of matching size for training our rankers, using the per-turn target value approximation described in Section 3.2 – the *Length* and *Rating* datasets for the respective versions of rankers.

The target values (length/rating) in both sets are normalized into the $[0, 1]$ range, and the *Length* set contains context-response pairs from long dialogues (target value above 0.7) as positive instances and context-response pairs from short dialogues (target value below 0.3) as negative ones. With the same selection criteria, the *Rating* set contains context-response pairs from highly rated dialogues (ratings 4 and 5) as positive instances and context-response pairs from low-rated dialogues (ratings 1 and 2) as negative ones. Both datasets contain 500,000 instances in total, with equal proportion of positive and negative instances. We use a 8:1:1 split for training, development, and test sets.

Prior to creating both datasets, we filtered out of the dialogue transcripts all system turns which cannot be treated as natural social interaction (e.g. a quiz game) as well as outliers (interaction length $\geq$ 95th percentile or less than 3 turns long).[3] Thresholds of 0.3 and 0.7 were set heuristically based on preliminary data analysis. On the one hand, these values provide contrastive-enough ratings (e.g. we are not sure whether the rating in the middle of the scale can be interpreted as negative or positive). On the other hand, they allow us to get enough training data for both Length and Rating datasets.[4]

6 Evaluation and experimental setup

In order to tune the neural rankers, we performed a grid search over the shared encoder GRU layer size and the Predictor topology.[5] The best configurations are determined by the loss on the development sets. For evaluation, we used an independent dataset.

6.1 Evaluation based on explicit user feedback

At the evaluation stage, we check how well the rankers can distinguish between good responses and bad ones. The criterion for 'goodness' that we use here is chosen to be independent from both training signals. Specifically, we collected an evaluation set composed of dialogue turns followed by explicit user feedback, e.g. "great, thank you", "that was interesting" (we refer to it as the *User feedback* dataset). Our 'bad' response candidates are randomly sampled across the dataset.

The user feedback turns were identified using sentiment analysis in combination with a whitelist and a blacklist of hand-picked phrases, so that in total we used 605 unique utterances, e.g. *"that's pretty cool", "you're funny", "gee thanks", "interesting fact", "funny alexa you're funny"*.

'Goodness' defined in this way allows us to evaluate how well our two approximated training signals can optimize for the user's satisfaction as explicitly expressed at the turn level, thus leading

[3] Some extremely long dialogues are due to users repeating themselves over and over, and so this filter removes these bad dialogues from the dataset. Dialogues less than 3 turns long are often where the user accidentally triggered the chatbot. These outliers amounted to about 14% of our data.

[4] Using more extreme thresholds did not produce enough data while less ones did not provide adequate training signal.

[5] We tested GRU sizes of 64, 128, 256 and Predictor layers number/sizes of [128], [128, 64], [128, 32, 32].

to our desired behaviour, i.e., producing long and engaging dialogues.

The *User feedback* dataset contains 24,982 $\langle context, good\_response, bad\_response\rangle$ tuples in total.

To evaluate the rankers on this dataset, we use *precision@k*, which is commonly used for information retrieval system evaluation (Eq. 3).

$$P@k(c, R) = \frac{\sum_{i=1}^{k} Relevant(c, R_k)}{k} \tag{3}$$

where c is dialogue context, R is response candidates list, and $Relevant$ is a binary predicate indicating whether a particular response is relevant to the context.

Precision is typically used together with recall and F-measure. However, since our dialogue data is extremely sparse so that it is hard to find multiple good responses for the same exact dialogue context, recall and F-measure cannot be applied to this setting. Therefore, since we only perform pairwise ranking, we use *precision@1* to check that the good answer is the top-ranked one. Also due to data sparsity, we only perform this evaluation with *gold positive* responses and *sampled negative* ones – it is typically not possible to find a good response with exactly the same context as a given bad response.

6.2 Interim results

The results of our first experiment are shown in Table 2. We can see that the neural ranker trained with user ratings clearly outperforms all the alternative approaches in terms of test set loss on its respective dataset as well as pairwise ranking precision on the evaluation dataset. Also note that both versions of the neural ranker stand extremely close to each other on both evaluation criteria, given a much greater gap between them and their next-best-performing alternatives, the linear rankers.

The dual-encoder ranker turned out to be not an efficient model for our problem, partly because it was originally optimized for a different task as reported by Lu et al. (2017).

7 Training on larger amounts of data

A major advantage of training on raw dialogue transcripts is data volume: in our case, we have roughly twice as many raw dialogues as rated ones (cf. Section 5). This situation is very common in data-driven development: since data annotation is a very expensive and slow procedure, almost always there is significantly more raw data than annotated data of a high quality. To illustrate this, we collected extended training datasets of raw dialogues of up to 1,000,000 data points for training from the length signal. We trained our neural ranker and the VW ranker using the same configuration as in Section 6.[6]

Model	P@1 (eval set)	Loss (test set)
Handcrafted	0.478	—
VowpalWabbit@length	0.742	0.199
VowpalWabbit@rating	0.773	0.202
DualEncoder@length	0.365	0.239
DualEncoder@rating	0.584	0.247
Neural@length	0.824	0.139
Neural@rating	**0.847**	**0.138**

Table 2: Ranking models evaluation: pairwise ranking precision on the independent *User feedback* dataset and loss on the *Length/Rating* test sets (Section 5) for the corresponding trainset sizes of 500,000.

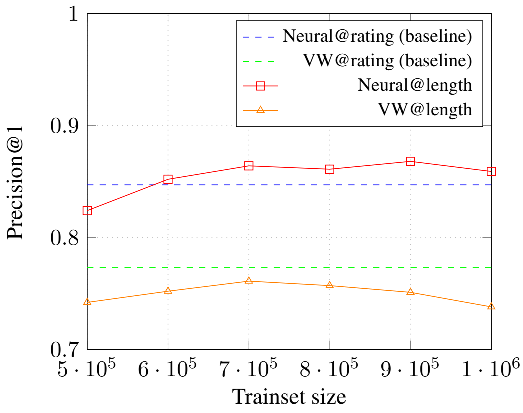

Figure 2: Comparison of rankers trained on extended datasets

The results are shown in Figure 2, where we see that the neural ranker trained on the length signal consistently outperform the ratings-based one. Its trend, although fluctuating, is more stable than that of VW – we believe that this is due to VW's inherent lower model capacity as well as its training setup, which is mainly optimised for speed. The figure also shows that VW@length is worse than VW@rating, regardless of training data size.

8 Discussion and future work

Our evaluation results show that the neural ranker presented above is an efficient approach to re-

[6] We were not able to train the dual encoder ranker on all the extended datasets due to the time constraints.

sponse ranking for social conversation. On a medium-sized training set, the two versions of the neural ranker, length and ratings-based, showed strongly superior performance to three alternative ranking approaches, and performed competitively with each other. Furthermore, the experiment with extended training sets shows that the accuracy of the length-based neural ranker grows steadily given more unannotated training data, outperforming the rating-based ranker with only slightly larger training sets.

The overall results of our experiments confirm that dialogue length, even approximated in quite a straightforward way, provides a sufficient supervision signal for training a ranker for a social conversation model. In future work, we will attempt to further improve the model using the same data in an adversarial setup following Wang et al. (2017). We also plan to directly train our model for pairwise ranking in the fashion of Burges et al. (2005) instead of the current pointwise approach. Finally, we are going to employ contextual sampling of negative responses using approximate nearest neighbour search (Johnson et al., 2017) in order to perform a more efficient pairwise training.

9 Related work

Work on response ranking for conversational systems has been been growing rapidly in recent years. Some authors employ ranking based on heuristically defined measures: Yu et al. (2015, 2016) use a heuristic based on keyword matching, part-of-speech filters, and Word2Vec similarity. (Krause et al., 2017) apply standard information retrieval metrics (TF-IDF) with importance weighting for named entities. However, most of the recent research attempts to train the ranking function from large amounts of conversational data, as we do. Some authors use task-based conversations, such as IT forums (Lowe et al., 2015) or customer services (Lu et al., 2017; Kumar et al., 2018), while others focus on online conversations on social media (e.g. Wu et al., 2016; Al-Rfou et al., 2016).

The basic approach to learning the ranking function in most recent work is the same (e.g. Lowe et al., 2015; Al-Rfou et al., 2016; Wu et al., 2016): the predictor is taught to rank positive responses taken from real dialogue data higher than randomly sampled negative examples. Some of the approaches do not even include rich dialogue contexts and use only immediate context-response pairs for ranking (Ji et al., 2014; Yan et al., 2016; Lu et al., 2017). Some authors improve upon this basic scenario: Zhuang et al. (2018) take a desired emotion of the response into account; Liu et al. (2017) focus on the engagement of responses based on Reddit comments rating; Fedorenko et al. (2017) train the ranking model in several iterations, using highly ranked incorrect responses as negative examples for the next iteration. Nevertheless, to our knowledge, none of the prior works attempt to optimise for long-term dialogue quality; unlike in our work, their only ranking criterion is focused on the immediate response.

10 Conclusion

We have presented a neural response ranker for open-domain 'social' dialogue systems and described two methods for training it using common supervision signals coming from conversational data: user-provided ratings and dialogue length. We demonstrated its efficiency by evaluating it using explicit positive feedback as a measure for user engagement. Specifically, trained on ratings, our neural ranker consistently outperforms several strong baselines; moreover, given larger amounts of data and only using conversation length as the objective, the ranker performs better the ratings-based one, reaching 0.87 Precision@1. This shows that conversation length can be used as an optimisation objective for generating engaging social dialogues, which means that we no longer need the expensive and time-consuming procedure of collecting per-dialogue user ratings, as was done for example in the Alexa Prize 2017 and is common practice in conversational AI research. Per-turn user ratings may still be valuable to collect for such systems, but these are even more expensive and problematic to obtain. Looking ahead, this advance will make data collection for social conversational agents simpler and less expensive in the future.

Acknowledgements

This research received funding from the EPSRC project MaDrIgAL (EP/N017536/1). The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. 2014. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15(1):1111–1133.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational Contextual Cues: The Case of Personalization and History for Response Ranking. *CoRR*, abs/1606.00372.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of ICML*, pages 89–96.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Denis Fedorenko, Nikita Smetanin, and Artem Rodichev. 2017. Avoiding Echo-Responses in a Retrieval-Based Conversation System. *CoRR*, abs/1712.05626.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370.

C. J. Gilbert and Erric Hutto. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*, pages 216–225, Ann Arbor, MI, USA.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of AISTATS*, pages 315–323.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An Information Retrieval Approach to Short Text Conversation. *CoRR*, abs/1408.6988.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *CoRR*, abs/1702.08734.

Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie Webber. 2017. Edina: Building an Open Domain Socialbot with Self-dialogues. In *1st Proceedings of Alexa Prize*, Las Vegas, NV, USA. ArXiv: 1709.09816.

Girish Kumar, Matthew Henderson, Shannon Chan, Hoang Nguyen, and Lucas Ngoo. 2018. Question-Answer Selection in User to User Marketplace Conversations. *CoRR*, abs/1802.01766.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proc. EMNLP*, pages 1192–1202.

Huiting Liu, Tao Lin, Hanfei Sun, Weijian Lin, Chih-Wei Chang, Teng Zhong, and Alexander Rudnicky. 2017. RubyStar: A Non-Task-Oriented Mixture Model Dialog System. In *1st Proceedings of Alexa Prize*, Las Vegas, NV, USA. ArXiv: 1711.02781.

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. In *Proceedings of SIGDIAL*, pages 285–294.

Yichao Lu, Phillip Keung, Shaonan Zhang, Jason Sun, and Vikas Bhardwaj. 2017. A practical approach to dialogue response generation in closed domains. *CoRR*, abs/1703.09439.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L. Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dusek, Verena Rieser, and Oliver Lemon. 2017. An ensemble model with ranking for social dialogue. In *NIPS Workshop on Conversational AI.*

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2017. Conversational AI: The Science Behind the Alexa Prize. In *1st Proceedings of Alexa Prize*, Las Vegas, NV, USA. ArXiv: 1801.03604.

Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A deep reinforcement learning chatbot. *NIPS Workshop on Conversational AI*, abs/1709.02349.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *CoRR*, abs/1610.07149.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2017. On Evaluating and Comparing Conversational Agents. In *NIPS 2017 Workshop on Conversational AI (ConvAI)*, Long Beach, CA, USA. ArXiv: 1801.03625.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *ICML Deep Learning Workshop 2015*.

Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of SIGIR*, pages 515–524, Shinjuku, Japan.

Bowen Wu, Baoxun Wang, and Hui Xue. 2016. Ranking responses oriented to conversational relevance in chat-bots. In *Proceedings of COLING*, pages 652–662, Osaka, Japan.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of SIGIR*, pages 55–64, Pisa, Italy.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Zhou Yu, Alexandros Papangelis, and Alex Rudnicky. 2015. TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness. In *Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium*, pages 108–111, Palo Alto, CA, USA.

Zhou Yu, Ziyu Xu, Alan W Black, and Alex I. Rudnicky. 2016. Strategy and Policy Learning for Non-Task-Oriented Conversational Systems. In *Proc. SIGDIAL*, Los Angeles, CA, USA.

Yimeng Zhuang, Xianliang Wang, Han Zhang, Jinghui Xie, and Xuan Zhu. 2018. An Ensemble Approach to Conversation Generation. In *Natural Language Processing and Chinese Computing*, volume 10619, pages 51–62. Springer International Publishing. DOI: 10.1007/978-3-319-73618-1_5.