Making Chatbots Better by Training on Less Data

Richard Krisztian Csaky
Department of Automation
and Applied Informatics
Budapest University
of Technology and Economics
ricsinaruto@hotmail.com

Gabor Recski
Department of Automation
and Applied Informatics
Budapest University
of Technology and Economics
recski.gabor@aut.bme.hu

Abstract

Current conversational models lack diversity and generate boring responses to open-ended utterances (Li et al., 2015; Wei et al., 2017; Shao et al., 2017). *Priors* provide additional information to dialogue models to aid response generation, but annotating a dataset with priors such as persona (Li et al., 2016a), emotion (Zhou et al., 2017), or topic (Xing et al., 2017), is expensive and such annotations are rarely available. We present a method for improving chatbots' responses to open-ended utterances by removing those utterances from training data using a simple entropy-based approach that does not require human supervision. We show that training on this filtered dataset results in better conversational quality as chatbots learn to output better and more diverse responses to these utterances.

1 Introduction

Current open-domain neural conversational models (NCM) are trained on pairs of source and target utterances in an effort to maximize the likelihood of each target given the source (Vinyals and Le, 2015). However, real-world conversations are much more complex, and a plethora of suitable targets (responses) can exist to a given input. We remedy this issue by excluding from the training set inputs which are the "most open-ended", determined by calculating the entropy of the distribution over target utterances. We show that dialogue models can be improved by using this simple filtering method, effectively training on less data.

Previous approaches to the issues mentioned are discussed in Section 2, and Section 3 describes our method in detail. Section 4 presents an analysis of the filtered dataset, dialogue systems trained on the new datasets are evaluated in Section 5. Section 6 concludes and presents future work.

2 Background

Current open-domain NCMs are based on neural architectures developed for machine translation (MT). Conversational data differs greatly from MT data in that targets to the same source may vary not only grammatically but also semantically (Wei et al., 2017; Tandon et al., 2017); consider plausible replies to the question: *What did you do today?*. Dialogue datasets also contain responses that appear after many different inputs, e.g. answers such as *yes*, *no* and *i don't know* appear after a large and diverse set of inputs. Following the approach of modeling conversation as a sequence to sequence (`seq2seq`) (Sutskever et al., 2014) transduction of single dialogue turns, these issues can be referred to as the *one-to-many*, and *many-to-one* problem, respectively. Since `seq2seq` architectures are inherently deterministic, meaning that once trained they can't output different sequences to the same input sequence, they are not suited to deal with the ambiguous nature of dialogues. Related to these issues is the evaluation of open-domain chatbots. Currently, there is no well-defined automatic evaluation method (AVM) (Liu et al., 2016), and many researchers resort to human evaluation (Vinyals and Le, 2015; Serban et al., 2017a; Ram et al., 2018). Some AVMs that correlate with human judgement have been proposed recently (Li et al., 2017a; Lowe et al., 2017), but they are harder to conduct than perplexity or BLEU (Papineni et al., 2002).

The focus of this work is the *one-to-many*, and *many-to-one* problem, previous approaches to which can be grouped into three categories. First, the encoding procedure can be modified by feeding more information into the model, like dialogue history (Serban et al., 2016), persona information (Li et al., 2016a; Joshi et al., 2017; Zhang et al., 2018), mood/emotion category (Zhou et al., 2017;

Li et al., 2017b), topic category (Xing et al., 2017; Liu et al., 2017), etc. Second, some approaches augment the decoding process, with e.g. latent variable sampling (Serban et al., 2017b; Zhao et al., 2017) or beam search (Goyal et al., 2017; Wiseman and Rush, 2016; Shao et al., 2017). Finally, directly modifying the loss function (Wiseman and Rush, 2016) or training procedure of the model, by using reinforcement (Li et al., 2016c; Serban et al., 2017a; Li et al., 2016b; Lipton et al., 2017) or adversarial learning (Li et al., 2017a) are also among the solutions proposed.

3 Methods

In this work the *one-to-many*, *many-to-one* issue is approached from a different perspective: instead of adding more complexity, we try simple data filtering methods to exclude source-target utterance pairs that have high entropy, since we believe that these cause dialogue models to output safe but boring responses. Entropy of utterances has also been used before for evaluation purposes (Serban et al., 2017b). The entropy of a source/target utterance is calculated based on the distribution of the target/source utterances that it is paired with in the dataset. In essence, the learning task is formulated in a way for which the maximizing likelihood approach is more suitable. NCMs have been shown to produce better qualitative results after they overfit the training data (Csáky, 2017; Tandon et al., 2017). This also supports the claim that the loss function is not capturing conversational goals, since a neural network model should perform best when the validation loss is minimal. Our experiments suggest that when training NCMs on our filtered datasets, validation loss becomes a better indicator of the model's performance.

Of the 72 000 unique source utterances in the DailyDialog dataset (see Section 4 for details), 60 000 occur with only a single target. For these it seems straightforward to maximize the conditional probability $P(T|S)$, S and T denoting a specific source and target utterance. However, in the case of sources that appear with multiple targets in the dataset, models are forced to learn some "average" of observed responses. This is the *one-to-many* problem. We can similarly formulate the *many-to-one* problem, where a diverse set of source utterances are observed with the same target. This may be a less prominent issue in training NCMs, since the probability of source utterances given some target doesn't appear in standard loss functions (although it is used in some special objective functions (Li et al., 2015)). Still, we shall experiment with excluding such targets (e.g. *I don't know*), since conversational models generate these quite frequently and they are typically uninformative and unengaging (see Section 5.1 on evaluation principles).

For each source utterance s in the dataset we calculate the entropy of the distribution $T|S = s$, i.e. given a dataset D of source-target pairs we define the *target entropy* of an utterance s as

$$E_{\text{tgt}}(s, D) = -\sum_{(s,t_i)\in D} p(t_i|s) \log_2 p(t_i|s)$$

Similarly, *source entropy* of an utterance can be defined as

$$E_{\text{src}}(t, D) = -\sum_{(s_i,t)\in D} p(s_i|t) \log_2 p(s_i|t)$$

The probabilities are calculated based on the observed relative frequency of sentence pairs in the data. After calculating source and target entropies for each utterance in a corpus, we filter the training data using one of 3 strategies. TARGET-BASED, where pairs are filtered if the source utterance has high target entropy. SOURCE-BASED, where we filter based on the source entropy of the target utterance. Finally, the ST-BASED dataset is obtained by filtering pairs based on both entropy values.

We also experimented with clustering utterances based on Jaccard similarity between their bags of words (BoW), measuring the entropy of clusters as opposed to individual utterances. Our intuition was that such a cluster-based entropy may better reflect the semantic diversity of all target/source utterances associated with some source/target. In practice, however, this method proved inferior, bringing only noise to the simpler approach in identifying the utterances that give rise to the most diverse sets of responses. This may be due to the BoW model's inability to capture semantic similarity. In Section 6 we briefly present our plans for experimenting with different metrics in the future.

4 Experiments

4.1 Dataset

We use the DailyDialog dataset[1] (Li et al., 2017b) in our experiments. With 90 000 utterances

[1] http://yanran.li/dailydialog.html

in 13 000 dialogues, it is comparable in size with the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011), but contains real-world high quality dialogues, instead of movie conversations, which are "not truthful representations of real-life conversations" (Danescu-Niculescu-Mizil and Lee, 2011). The vocabulary was set to 16384, covering most of the words in the corpus (roughly 19000).

4.2 Models

For dialogue modeling we use `transformer` (Vaswani et al., 2017), a novel encoder-decoder architecture. Compared to the standard recurrent neural network (RNN) based `seq2seq` models, it doesn't use recurrent connections and relies only on attention mechanisms (Bahdanau et al., 2015). Consequently, it can be trained much faster, and using less memory (training the `seq2seq` model of (Vinyals and Le, 2015) was not possible with the 8GB of GPU memory we had access to). We further justify the use of this model with the fact that it achieves state-of-the-art performance in NMT. Since the original `seq2seq` model was adopted from NMT (Cho et al., 2014) to dialogue modeling, it is natural to do the same with the `transformer` architecture. To justify its use for the dialogue task, we also train a `seq2seq` model (of limited size) on the same dataset for comparison. The `transformer` and `seq2seq` models contain 53M and 317M parameters, respectively. They are both large compared to the dataset thus they easily overfit it, as will be shown in Section 5. In the case of the `transformer` model we also experimented with different dropout (Srivastava et al., 2014) values our findings will also be presented in Section 5.

We trained randomly initialized word embeddings (of size 512) together with the model parameters. For the `transformer` model layer, attention, and relu dropout was set to 0.2, 0.1 and 0.1 respectively. At test time we used beam search with a beam size of 10 (Graves, 2012).

4.3 Filtered data

The 90 000 utterance pairs in the DailyDialog dataset contain about 72 000 unique utterances. We plot target entropies of source utterances in Figure 1, ranked from lowest to highest entropy, not showing the majority of utterances which have 0 entropy (i.e. they do not appear with more than one target). Source entropies of target utterances are very similar. In the following experiments we shall discard utterance pairs whose target and/or source entropy is greater than 1. This affects 5.64%, 6.98% and 12.24% of the data, for the TARGET-BASED, SOURCE-BASED and ST-BASED scenario, respectively.

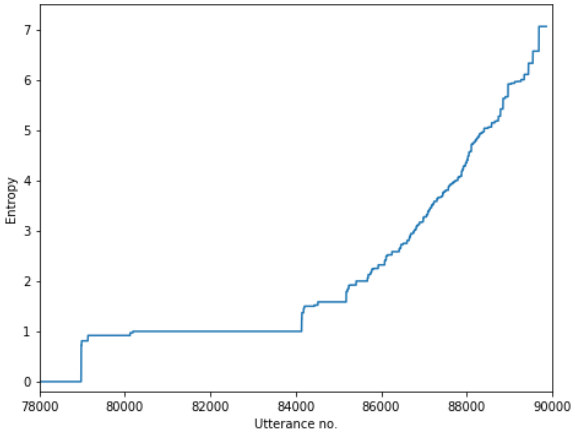

Figure 1: Source utterances by target entropy

Entropy is clearly proportional to utterance frequency (Figure 2), however we found that only 485 utterances overlap in the top 700 utterances (roughly what gets discarded) when ordered by both entropy and frequency, and those that are different in the frequency ordered list are long utterances, that we don't wish to filter out. Entropy offers a more fine-grained measure compared to frequency, and in the case of low frequency pairs, this is especially helpful. For example, all utterances that have a frequency of 3, are in the same category based on frequency, but their entropy can range from 0 to $\log_2 3 \approx 1.58$, which would be over our filtering threshold.

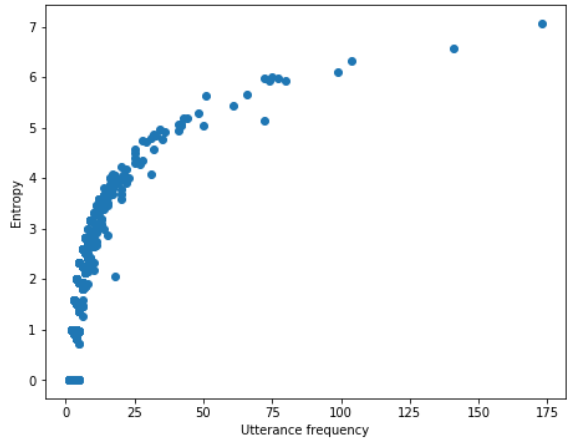

Figure 2: Target entropies of sources with respect to utterance frequency.

After noticing that high-entropy utterances are relatively short, we also examined the relationship between entropy and utterance length (Figure 3). Given the relationship between frequency and entropy it comes as no surprise that longer sentences have lower entropy, although this effect is less pronounced in the range affected by filtering.

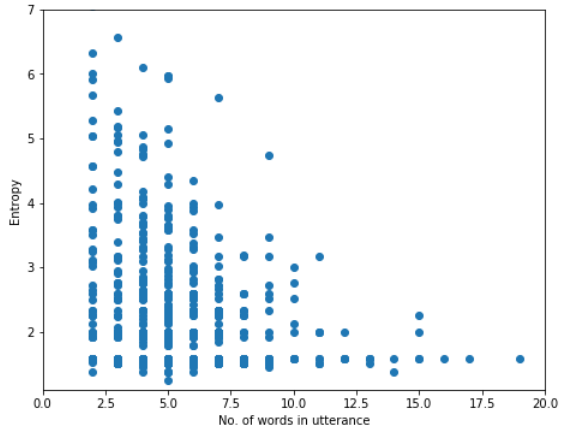

Figure 3: Target entropies of sources with respect to utterance length.

5 Results

5.1 Evaluation Principles

Due to the limited scope of automatic evaluation (Liu et al., 2016), our claims are supported by both qualitative and quantitative evaluation of our models' outputs. We first summarize the principles used for comparison, then present our findings. Because of space constraints we couldn't include the large samples of responses based on which we conducted the manual evaluation, but we invite readers to inspect the tables[2].

We shall evaluate models based on answers they give to a list of input utterances. When comparing responses given to the same question, we first consider their *coherence*, i.e. whether they could have been written by a human speaker. Models that return coherent answers to an input are further compared based on whether the answer is boring/generic (e.g. *I don't know*) or engaging/specific. Generic responses make sense in more dialogue histories, however because of this they do not add new information to the conversation. More engaging responses, i.e. those that can further a conversation, are preferred, but only if they are coherent with the source utterance. While these principles, and thus our judgements presented here, are quite subjective, we believe that the differences observed between various trainings are sufficiently pronounced, and our findings are also grounded by the quantitative analysis.

[2] https://anonfile.com/54YeAbf6b6/tables.pdf

Input	S2S	S2S-O	TRF	TRF-O
what is the color of water?	it's the best we can do it	leopard. it's famous for its quality only.	**it's red.**	**it's red.**

Table 1: Responses to an input utterance, by overfitted (O) and non-overfitted versions of the `seq2seq` (S2S) and `transformer` (TRF) models trained on unfiltered data. Best responses are boldfaced.

Four metrics are used to quantitatively evaluate our models in Section 5.5. In order to measure the information content of a models' responses, average utterance length $|U|$, word entropy H_w and utterance entropy H_u is computed (Serban et al., 2017b). The entropies are computed with respect to the maximum-likelihood unigram distribution of the training set. Thus: $H_w = -\sum_{w \in U} p(w) \log(p(w))$, and H_u is simply the product of utterance length and word entropy. Additionally the average jaccard similarity J between source utterances and model responses is computed, measuring their coherence and relevance.

5.2 `seq2seq`, `transformer`, and overfitting

We first compare overfitted and non-overfitted versions of the `seq2seq` (S2S) and `transformer` (TRF) models trained on unfiltered data. Overfitted model versions, are those that were trained further after the lowest value of validation loss was reached, until training loss converges. We used a list of 34 general input utterances chosen from the ones used in Vinyals and Le (2015), which we will call the NCM test inputs. We filtered this list by cutting well-known inputs to which each model learns to respond well (eg. *Hello!*, *How are you?*), and also removed inputs where any word was missing from the vocabulary.

For each question we determined the answers which we judged best, based on the principles outlined at the beginning of this section. A representative row of the complete table can be seen in Table 1. Generally, the `transformer` performed better than the `seq2seq` model, achieving 11 best responses (among the 4 models), compared to only 7 for `seq2seq`. It managed to output colours when asked about the colour of objects, while the

seq2seq's replies were irrelevant.

Overfitted models performed at least as well (in human evaluation) as non-overfitted models, strengthening the points raised in Csáky (2017); Tandon et al. (2017): the loss function does not adequately represent the quality of a chatbot model. The overfitted seq2seq model achieved 9 best response (2 more than the non-overfitted version), and in the case of the transformer the overfitted version achieved 12 best responses. We noted that overfitted models tend to generate longer responses, which is generally good, but in some cases we obtain too specific and probably memorized responses to unrelated inputs (eg. *they have a really good dj and a good dance.* to the input *what is the purpose of dying*).

Since our models overfit quickly, we also experimented with dropout. With a high dropout rate (0.5) we can essentially force the validation loss to stay at its minimum for longer, before starting to overfit. However, the minimum does not go lower compared to low dropout (0.2) trainings, and the replies were generally the same even after training more with high dropout, further consolidating the observation that the validation loss minimum does not represent the best state of the model.

5.3 High Entropy Inputs

We evaluate the transformer model trained on unfiltered and filtered datasets (according to the 3 filtering types discussed in Section 3) on the 45 highest entropy source utterances. These are the most challenging utterances (eg. *yes; thank you; why?; sure; no; what's that?; here you are*), where dialogue models tend to fail, because of the high diversity observed in the dataset. The TARGET-BASED filtering variant is excluded from this evaluation, because as we will see in Section 5.4 it performs poorly on the NCM test set, and also according to the automatic metrics (Section 5.5).

Counter-intuitively there is clear improvement in the performance on these utterances which the filtered models didn't see during training. The ST-BASED training gives the best response in 23 cases, while the unfiltered training only in 12 cases. Solely filtering the target side, gives slightly worse results, achieving only 11 best responses, however its responses can be often selected as the second-best after the ST-BASED. A closer look at the ST-BASED replies shows two main enhancements. First, the model was able to generate more diverse responses, while also keeping them general enough (eg. *I have a bad headache.*, or *I'm glad to hear that.*), which is probably mostly due the source-side filtering. Second, where the unfiltered model often choose to output the same safe response (*thank you.*), the filtered model responds with engaging questions to further the conversation. This is clearly due to the target side filtering, since the model was forced to not learn to output generic responses. The conclusion is further reinforced by the SOURCE-BASED training, where the model answers with questions more frequently. However, the SOURCE-BASED training is still not diverse enough, combining the two methods seems the most advantageous. We also experimented with an overfitted variant of the ST-BASED training, which performed a lot worse, and was too specific in many cases (giving the best response only in 7 cases). Overall it appears that with our filtered dataset the model performs better at the validation loss minimum.

5.4 NCM test inputs

We also evaluate the transformer model trained on unfiltered and filtered datasets on the NCM test inputs. The ST-BASED and SOURCE-BASED trainings are on par with the unfiltered training (15, 16, 15 best responses, respectively), followed by the TARGET-BASED training (11 best responses). These results prove that the model is still capable to output good responses to the general NCM test inputs, even when trained on the filtered dataset. Filtering the source side alone gives worse results than filtering the target side alone, demonstrating that discarding generic responses adds more to conversational quality.

Finally, the overfitted version of the ST-BASED training performs slightly worse (getting best response in only 13 cases), somewhat alleviating the problems discussed in Section 5.2. As with the high entropy inputs, this indicates that filtering a dataset based on entropy, makes the learning problem more aligned with the loss function.

5.5 Quantitative Analysis

In Table 2 all metrics are computed based on responses given to a separate test set, containing 10% of the utterances from DailyDialog. Looking only at the unfiltered trainings we can see that the transformer performed better than the seq2seq model across all metrics. In contrast to the manual evaluations however, on automatic

Unfiltered trainings	$\|U\|$	H_w	H_u	J
TRF-BASE	4.93	0.493	2.43	0.091
TRF-BASE-O	9.82	0.797	7.83	0.099
S2S-BASED	4.35	0.462	2.01	0.089
S2S-BASE-O	7.09	0.628	4.45	0.098
Filtered trainings				
TRF-ST-BASED	6.31	0.586	3.70	0.099
TRF-ST-BASED-O	**10.42**	**0.838**	**8.73**	**0.101**
TRF-TARGET-BASED	5.25	0.525	2.76	0.096
TRF-SOURCE-BASED	***6.81***	***0.61***	***4.15***	***0.100***
Targets	14.10	1.031	14.54	0.105

Table 2: Quantitative metrics computed based on the test set. First, trainings which were trained on the normal (unfiltered) dataset are presented, and then trainings run on the filtered datasets. TRF refers to the `transformer` model, and S2S refers to the `seq2seq` model. The type of filtering is also noted (ST-BASED, SOURCE-BASED, TARGET-BASED), and the O notation means that it is an overfitted version of the model. Results of best non-overfitted models are in italic boldface, while best results overall are noted by simple boldface.

	ST-BASED		SRC-BASED		TGT-BASED	
	BASE	FILT	BASE	FILT	BASE	FILT
$\|U\|$	4.94	**6.29**	4.56	**6.32**	5.39	**5.57**
H_w	0.512	**0.598**	0.481	**0.604**	**0.551**	0.541
H_u	2.53	**3.76**	2.19	**3.82**	2.97	**3.01**
J	**0.121**	0.099	0.115	**0.119**	0.130	0.130

Table 3: Quantitative metrics computed from the responses of the base model and the 3 filtered trainings on 3 different test sets. The ST-BASED, SOURCE-BASED, and TARGET-BASED columns refer to test sets constructed by taking the difference between the unfiltered and the 3 filtered test sets respectively. BASE refers to the unfiltered `transformer` training, and FILT refers to the 3 different types of filtered trainings.

metrics all examined overfitted models performed much better than their non-overfitted counterparts.

The results of the filtered trainings are also presented in Table 2. It is clear that all types of filtering show significant improvement across all metrics. Interestingly, in contrast to the manual evaluations the SOURCE-BASED filtering achieves the best results, ST-BASED being the second best, and aligned with the manual evaluations TARGET-BASED is the last. Using SOURCE-BASED filtering alone, and thus filtering boring and generic responses is more important than TARGET-BASED filtering, and combining the two types is not beneficial according to these metrics. Also, the overfitted version of the ST-BASED training achieves the best performance, improving on the unfiltered training variant. Thus, while training on filtered datasets generally improves performance, a non-overfitted model still can't be competitive with an overfitted variant. However the performance gap between them gets somewhat smaller than in the case of unfiltered trainings. This also shows the limitation of these metrics that value diversity, since as seen in the manual evaluation, overfitted models tend to be too specific, by outputting learned responses.

We also constructed 3 smaller test sets, by taking the difference between the unfiltered test set and the filtered test sets (ST-BASED, SOURCE-BASED, and TARGET-BASED). With these test sets we want to show that the models trained on filtered data perform better even on the high entropy data that was filtered. Indeed in Table 3 we can see that the filtered trainings generally performed better, SOURCE-BASED being the only test set and training that performed better than the model trained on unfiltered data across all metrics.

6 Conclusion

We showed how with a simple entropy-based approach we can find generic and safe sources/targets that usual dialogue models have problems with. We compared the various trainings in an extensive qualitative and quantitative evaluation. The unfiltered and the filtered trainings were compared on two different objectively picked test sets, and on several automatic metrics used in the literature. We showed how the model trained on the filtered dataset outputs more engaging and interesting responses to inputs that it has never seen. Moreover, the `transformer` was shown to be at least as good for dialogue modelling as the `seq2seq`, and evaluating these models trained on unfiltered data at an overfitted point results in better conversational quality, while training on filtered data somewhat alleviates this issue.

For future work we wish to try clustering methods, that cluster based on "semantic distance", rather than syntactic similarity. Sentence similarity is an active research area, and many approaches exist such as word embedding averaging (Arora et al., 2016), and RNN based methods (Tai et al., 2015). We wish to use these sentence representations for clustering and filtering the dataset as described in Section 3. Furthermore, we want to test our filtering approach together with other approaches to the loss function problem discussed in Section 2, like dialogue history.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *arXiv:1601.03764v1*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR 2015)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, 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.

Richárd Csáky. 2017. Deep learning based chatbot models. Technical report, Budapest University of Technology and Economics, Budapest.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Kartik Goyal, Graham Neubig, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017. A continuous relaxation of beam search for end-to-end training of neural sequence models. *arXiv preprint arXiv:1708.00111*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *Representation Learning Workshop, ICML 2012*, Edinburgh, Scotland.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016b. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016c. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2017. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *arXiv preprint arXiv:1711.05715*.

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*.

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. *arXiv preprint arXiv:1711.02781*.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

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 of the association for Computational Linguistics*, pages 311–318, Philadelphia.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017a. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2200–2209.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

I. Sutskever, O. Vinyals, and Le. Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Shubhangi Tandon, Ryan Bauer, et al. 2017. A dual encoder sequence to sequence model for open-domain dialogue modeling. *arXiv preprint arXiv:1710.10520*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2017. Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. *arXiv preprint arXiv:1712.02250*.

Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*, pages 3351–3357.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.