Problems with Evaluating Chatbots

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Chatbots are difficult to evaluate. Task-based evaluations are out, since chatbots are focused on dialog rather than goal-based tasks. Automatic evaluation metrics like the BLEU (Papineni et al., 2002) score from machine translation are poor fits for dialog. As a result, the current best practice is to analyze and compare dialog systems are using human judgments. However, model results are extremely hard to replicate, in part because model parameters and code are rarely published and in part due to differences in evaluation procedures. We survey and categorize evaluation methods for human assessments in recent non-task driven neural dialog papers, propose a framework for evaluation, establish baselines for sequence-to-sequence models on several datasets, and compare against state-of-the-art models.

1 Introduction

Although there has been a tremendous amount of research in neural dialog agents recently, there are no standard experimental design or evaluation methods. To address this problem we developed an evaluation framework, which we call the *Standard Evaluation Tool for Chatbots* (SETC).[1]

While there are some promising directions for automatic evaluation of chatbots, human judgements are the gold standard for comparing and analyzing dialog systems. However, the evaluation criteria and methods used are highly variable across research projects. Published results are nearly impossible to replicate, in part because it is rare for both the prompts used in evaluation and the model parameters or code to be published. We propose to establish standard baseline neural models on a standard set of human evaluation tasks.

Reproducibility and model assessment for dialog systems is extremely challenging, as many small variations in the training setup or evaluation technique can result in large differences in perceived model performance. There are three essential aspects to a neural conversational agent: the dataset, the method, and the evaluation. Most papers focus on new methods; however, insufficient attention is paid to standardizing dataset splits into train, development, test, and evaluation sets for rigorous model evaluation. Most existing work provides insufficient detail to actually reproduce the work or directly compare results. Due to the unconstrained nature of the task, human evaluation is still the gold standard for comparing models, further complicating comparisons.

Recent papers tend to evaluate their methodological improvements to sequence-to-sequence (Seq2Seq) baseline (Sutskever et al., 2014), as first proposed by Vinyals and Le (2015) with the the neural conversation model (NCM), rather than compare against each other. However, none of these supposed Seq2Seq baselines seem to have successfully reproduced the original NCM results. Indeed, we found no model, neither among those we trained nor those available online, that matched the performance of NCM during human evaluation. Cakechat, an open-source system released by a chatbot-based startup, achieved the best human evaluation performance. However, human evaluators overwhelmingly preferred NCM to CakeChat. We found that on the NCM's publicly released evaluation set, where NCM was preferred 55% of the time compared with 15% for Cakechat.

[1] `https://github.com/jsedoc/SETC/blob/master/README.md`

A lack of standardization in framework, training procedure, and evaluation strategy makes most existing results in non-task driven dialog unreproducible. We believe that the field requires a standard procedure which includes publication of code, model parameters, dataset splits, evaluation instructions and evaluation code for both human and automatic metrics.

The impact of small changes to training results is well known in machine learning and other deep learning fields. These issues are exacerbated for non task-driven dialog systems as there is rarely a single "correct" response, this leads to more local minima.

Possible variations at train time include:

- Random seed for weight initialization and batch ordering
- Data preprocessing, including tokenization and filtering
- Number of epochs trained for
- Model architecture and loss function

For human evaluation, we have seen the following variations in the literature:

- How the prompts were chosen (randomly from a test set, cherry-picked, etc.)
- Scalar (how good was this response?) vs. ordinal (rank the following responses) style questions
- Length of the prompts

This presents a large challenge in the field. To further complicate the issue, the standard method of evaluation is to use expensive human ratings from sources such as Amazon Mechanical Turk, CrowdFlower, and community annotators. Human judgments are often unreliable since for non-task driven chatbots, there is no clear objective. There are further tensions between local coherence assessments using standard evaluation sets, and human interactive evaluation.

This paper focuses on issues with current reproducibility and proposes a scientific framework for the community which requests openly available code, checkpoints, and standard splits of datasets. We make two major contributions: 1) We show that small variations present in baseline Seq2Seq models and human evaluation parameters can lead to large variation in assessed chatbot performance. 2) We provide a framework and code, SETC, for full transparency and reproducibility in evaluating chatbots. SETC allows researchers to submit their code, datasets [if not using standard ones], model parameters and hyper-parameters. We also provide open source evaluation code (including MTurk templates), model parameters, and evaluation results. Finally, we publish model parameters, hyper-parameters, and coefficients (i.e. checkpoints) in SETC, against which to validate methodological improvements.

2 Related Work

2.1 Machine Translation Evaluation Metrics

Neural dialog models largely borrowed from machine translation (MT). However, MT evaluation methods such as BLEU (Papineni et al., 2002), TER, and other metrics, do not correlate well with human preferences (Liu et al., 2016; Lowe et al., 2017). Very recent work has started to create automatic metrics for evaluation. From MT shared tasks there are standard evaluation sets and workshops, which have yielded standardized results (Callison-Burch et al., 2007; Callison-Burch et al., 2011). Recent shared tasks in non-task driven dialog include WOCHAT,[2] NIPS Conversational Intelligence Challenge,[3] and Dialogue Breakdown Challenge.[4] Nonetheless, there is no standard evaluation framework and design.

2.2 Automatic Evaluation

The difficulties in using automatic evaluation techniques for dialog generation are well-documented in earlier work (Liu et al., 2016). Standard metrics borrowed from machine translation and text summarization only weakly correlate with human judgments. Automatic metrics for dialog evaluation is an extremely active area of research (Liu et al., 2016; Tao et al., 2017; Lowe et al., 2017; Novikova et al., 2017; Galley et al., 2015; Sugiyama et al.,). More recently, Tao et al. (2017) introduced RUBER, which is designed specifically for open-dialog systems. RUBER rewards predicted responses that are

[2] `http://workshop.colips.org/wochat/`

[3] `http://convai.io/`

[4] `https://dbd-challenge.github.io/dbdc3/`

both similar to their ground-truth responses and predicted by a neural net to be appropriate given the query.

2.3 Shared Tasks

New work on frameworks such as ParlAi (Miller et al., 2017) offers a promising direction for a more unified skeleton for consensus among the community; however, without a standard experimental setup, model comparison remains difficult if not impossible. Recently, new evaluation methods and metrics have developed from shared tasks workshops such as WOCHAT[5] and DBDC.[6] As stated by Emer Gilmartin, chatbots "need practical tweaks to make systems sound more conversational." [7] However, these "tweaks" make comparisons exceedingly difficult.

3 Designing Human Evaluation Experiments

While human evaluation remains the gold standard for dialog research, the design of human evaluation experiments is far from standard. In this section, we describe several factors which could influence the conclusions that are drawn.

We restrict our analysis to setups where the evaluator is shown a prompt and two possible responses, and then asked to select the better one (or specify a tie). Riesler (2018) found that relative rankings yield more discriminative results than absolute assessments when evaluating natural language generation. While some prior work has focused on evaluations where users actually converse with the agent and then rate the conversation (Zhang et al., 2018). We believe that one or more context utterances and the model predicts the next single-turn is a critical and as yet unsolved building-block to building agents capable of longer conversation.

3.1 Selection of Evaluation Set

The majority of papers we surveyed randomly select a sample of prompts from a test set for use in evaluation. The test set is either chosen from data withheld at training time or from a related dataset. Such a system has the advantage that the distribution of the test set closely matches that of the training set. However, in many cases there is a mismatch between the distribution of the train/test set and the actual types of conversation we would expect a user to have when chatting with the agent. This is very much true for models trained on OpenSubtitles (Li et al., 2016), where the goal is most definitely not to generate movie scripts. For models trained on Twitter, it is reasonable to form a test set out of Tweets only if the end goal is to create a chatbot that responds well to Tweets. Much existing work fails to clearly state what the goal of their chatbot is. We argue that this information is critical to deciding on the right prompts to use for human evaluation.

In this paper, we consider three possible test set sources. We first follow standard convention and choose random subsets of 200 query-response from the test sets of Twitter and OpenSubtitles. For Twitter, we performed slight manual filtering to remove extremely offensive Tweets. In our survey of previous work, we noticed that very few papers using a random evaluation set actually release their selected list of prompts. Our paper is expected to serve as a baseline for future work, and we therefore publicly release all our evaluation sets.

For our second evaluation set, we consider the list of 200 questions released by Vinyals and Le (2015) in their seminal work on neural conversational models using a standard Seq2Seq framework borrowed from machine translation. We show that despite our best efforts, we have not been able to reproduce the results of their paper, and we argue that *all* papers claiming to improve upon Vinyals and Le (2015)'s Seq2Seq baseline should show human evaluation results on the same set of queries.

One limitation of the Vinyals and Le (2015) evaluation set is the lack of information on how prompts were selected. For dialog systems that have the goal of achieving engaging text-message-based chatting systems, we propose that the evaluation set from the DBDC SemEval task (Higashinaka et al., 2017) be

[5] http://workshop.colips.org/wochat/

[6] https://dbd-challenge.github.io/dbdc3/

[7] http://workshop.colips.org/wochat/@iwsds2017/documents/Report-IWSDS2017-Panel-SSWochat.pdf

used as a standard benchmark dataset. In that dataset, interactions between a human and three possible chatbots (TikTok, Iris, and CIC) were collected by showing a context represented by a short paragraph to a participant, and then asking participant user to converse with the chatbot. Participants knew that they were speaking with a chatbot, and the conversations reflect this. We believe that this dataset best reflects the kinds of text-based conversations that we would expect a user to have with an open-domain conversational agent.

3.2 Prompt Length

A large problem with using using Seq2Seq for dialog generation is that models, especially those trained on movie-based datasets such as OpenSubtitles, prefer to give short responses. Li et al. (2015) address this by rewarding longer responses. However, we noticed an even more fundamental problem: longer prompts tend to result in longer responses by the model.

4 Designing the Training Scheme

The model architecture and the loss function are the most obvious parameters to experiment with in order to improve earlier results. Indeed, such changes form the basis for most of the recent papers in dialog generation. Existing work has found that small variations in model architecture, such as changing the attention mechanism or reversing the input sequence can significantly affect results for machine translation (Britz et al., 2017).

In our experiments, we use the standard OpenNMT-py (Klein et al., 2017) Seq2Seq implementation with default parameters. We use two layers of LSTMs with 512 hidden neurons for the bidirectional encoder and the unidirectional decoder. We find that we can train models with highly variable performance by modifying only the weight initialization and the data preprocessing scheme. We note variations in performance that surpass the reported improvements upon a Seq2Seq baseline published in some existing work.

4.1 Random Seed

In order to assess the effect of local minima, we varied the seed values between 700, 701, and 702. This was done consistently across all experiment with and without pretrained embeddings and as well on both Twitter and OpenSubtitles datasets. For reproducibility, we verified consistent results with multiple training using the same seed.

4.2 Pre-trained Word Embeddings

We used the glove.42B.300d embeddings which are uncased and trained on 42 billion tokens of Common Crawl using GloVe. These embeddings are publicly available. [8] Experiments were run to test the effect of pretrained embeddings, which is similar to the work in machine translation of Qi et al. (2018). We did not work specifically use GloVe embeddings trained using a large Twitter, which may have improved results. We also repeat this for the OpenSubtitles dataset.

4.3 Data Preprocessing

For OpenSubtitles, we used the exact code from the ParlAi framework to generate the train, validation, and test sets. However, for Twitter data more processing was necessary. To our knowledge there is no standard data preprocessing script for our data. We are going to submit a patch to ParlAi in order to standardize the processing.

For the Twitter dataset we started with a version of the dataset which is built usin ParlAi and subsequently used the NLTK [9] TweetTokenized. We applied “garbage removal”, which is similar to the method from Ritter et al. (2010). We found that in the dataset there were many company based conversations and manually removed these (see Table 1 in the supplemental material for the complete list). Finally, we used ParlAi’s profanity filter code. In total roughly 35,000 out of 2.5 million tweets were removed. In our results, we showed that this process has no effect on human evaluation of model performance.

[8] http://nlp.stanford.edu/data/glove.42B.300d.zip

[9] https://www.nltk.org/

Paper	SC	MP	Datasets	DP	# prompts in human eval	# annotators per prompt	Human Eval Technique	Prompt Selection
(Vinyals and Le, 2015)	no	no	OpenSubtitles, IT Helpdesk	yes / no [11]	200	4	Human judges were asked to pick between response by NCM or Cleverbot. Ties were permitted.	Deliberately selected, available at http://ai.stanford.edu/ quocle/QAresults.pdf
(Ghazvininejad et al., 2017)	no	no	Twitter grounded with Foursquare	no	500	7	Human judges were asked to pick between their model and seq2seq on two parameters: appropriateness to the topic, and informativeness. Ties were permitted.	Not available
(Zhou et al., 2017b)	yes	yes	STC conversations	yes	200	not listed	Human judges were asked to rate responses in terms of content (scale of 0 to 2) and emotion (eithe 0 or 1).	Randomly sampled from test set
(Zhou et al., 2017a)	no	no	Tencent Weibo	no	300	3	Human judges rate top-5 responses from each model on a scale of 1 to 3.	Randomly sampled, unclear if from separate test set
(Zhang et al., 2018)	yes	no	Persona Chat, OpenSubtitles	yes	100 dialogs	n/a	Human converses with the agent and rates it on fluency, engagingness, consistency, and profile detection.	n/a
(Li et al., 2017)	yes	no	OpenSubtitles	yes	200	5	Human judges were asked to rank 3 of their models. Ties were permitted.	Randomly sampled from test set
(Li et al., 2015)	yes	no	Twitter, OpenSubtitles	yes	1000	5	Human judges were asked to pick between response by their model and seq2seq baseline based on the relevance of the response. Ties were permitted.	Randomly sampled from test set
(Li et al., 2016)	no	no	Subset of OpenSubtitles	no	200-500	3	Human judges were asked to pick between response by their model and seq2seq baseline based on three parameters: single turn general quality, single-turn easy to answer, and 5-turn general quality. Ties were permitted.	Randomly sampled from test set
(Sordoni et al., 2015)	no	no	Twitter	yes	2114	5	Human judges were asked to pick between response by two model variants.	The entire validation set.
(Serban et al., 2016)	yes	no	Twitter, Ubuntu Dialogue	yes	not mentioned	not mentioned	Human judges were asked to pick between response by two model variants. Ties were permitted.	not mentioned

Table 1: Table showing the evaluation + code availability of several of the major recent papers in the field. SC stands for source code available, MP for model parameters available, and DP for dataset public.

5 Experiments

5.1 Even original paper Neural Conversational Model is NOT Reproducible!

Figure 2 shows that our Seq2Seq model trained on the exact same OpenSubtitles dataset as NCM cannot achieve the results reporting in the paper and the evaluation set. In fact, NCM perform as well as a human annotator [12] who responded to the same set of prompts.

5.2 Automatic Descriptive Metrics

In Table 2 for each of our models, we show its response diversity, defined as ”the number of distinct unigrams (div-1) and bigrams (div-2) in generated responses as a fraction of the total generated tokens.” (Li et al., 2015)

		Distinct-1	Distinct-2
DbdcLong	GT	0.162191	0.404056
	Twitter	0.262803	0.359838
	OpenSubtitles	0.149805	0.143969
	CakeChat	0.161865	0.421687
DbdcShort	GT	0.476834	0.449807
	Twitter	0.234426	0.255738
	OpenSubtitles	0.160684	0.194872
	CakeChat	0.183195	0.380772

Table 2: Distinct-1 and distinct-2 are measured as the total number of distinct bigrams/unigrams in the responses divided by the total number of tokens.

5.3 Results

In Table 3 the statistical significance of all of the various trials are shown. The relative percentages model preferences from human annotators using the NCM evaluation set are shown in Figure 2.

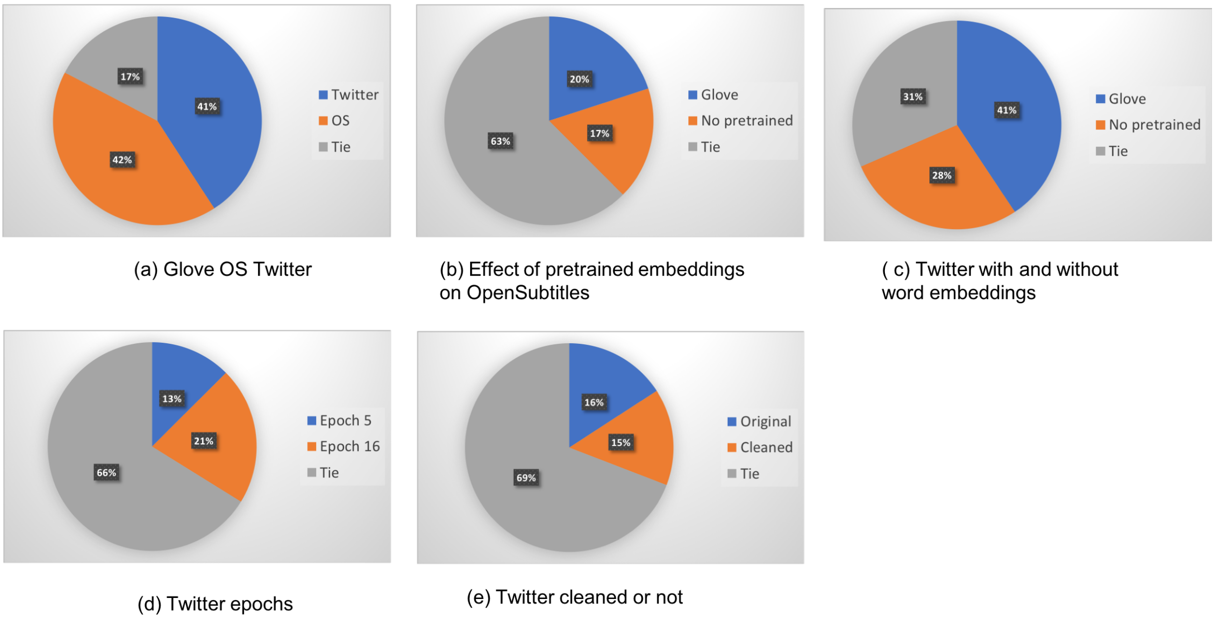
Training Dataset	**Variable**	**p-value**
OpenSubtitles	Seed	-
	# Epochs	***
	Word Embedding	-
	Filtering	***
Twitter	Seed	-
	# Epochs	***
	Word Embedding	***
	Data Filtering	-
NCM Eval Set	NCM > CC	***
	NCM >> Seq2Seq	***
	NCM Human	-
	Human >> CC	***
	Human >> Seq2Seq	***
	CC > Seq2Seq	***

Table 3: These are out of 200 prompts with 3 annotators for each win/loss/tie decision. P-value < 0.001 is signified by ***.

These experiments had the 200 prompts from the evaluation dataset, and 3 MTurk annotated 10 prompts per HIT. MTurkers were able to evaluate multiple times. The prompts presented were randomized, as well the ordering of the models was randomized.

[12] https://github.com/jsedoc/Chatbot_evaluation/blob/master/eval_data/neural_conv_model_human_responses.txt

As described in section 4, three different experimental conditions where varied over two datasets with three evaluation sets. After the models are trained then the actual workflow through SETC takes less than ten minutes. Figure 1 contains the human judgments from the different experimental variables changed.



(a) Glove OS Twitter

(b) Effect of pretrained embeddings on OpenSubtitles

(c) Twitter with and without word embeddings

(d) Twitter epochs

(e) Twitter cleaned or not

Figure 1: Results of the evaluation of Seq2Seq using the NCM evaluation set.

Next we compared different models on three evaluation sets. To assess the role of the length of the prompt, we split DBDC into small (less than 5 tokens) and large (more than 7 tokens). From Table 2 one can note that the distinct-1 and 2 are somewhat correlated to model judgments. This opens a possible area of exploration.

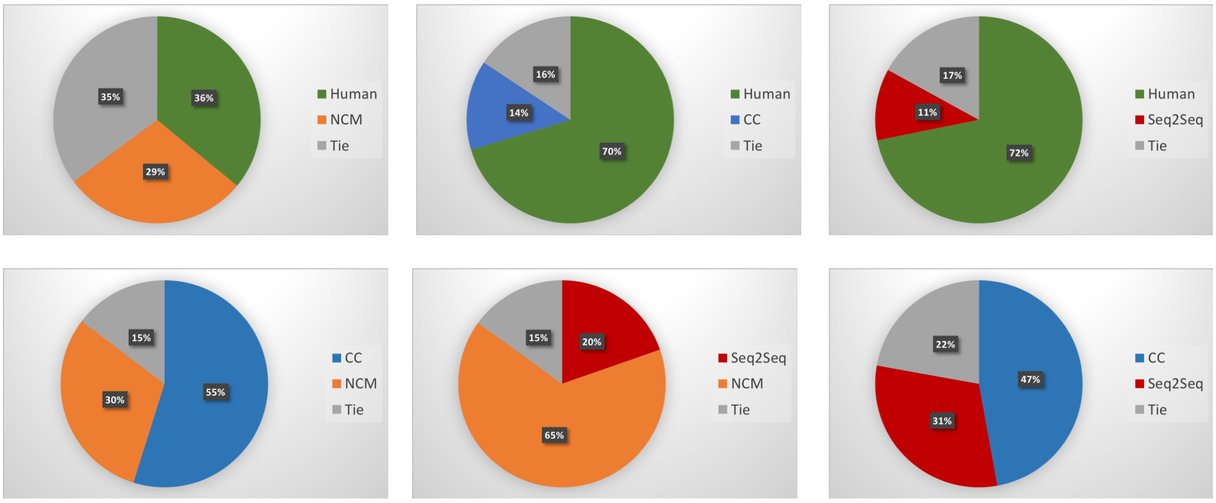


Figure 2: Results of the evaluation of neural conversation model (NCM), Cakechat (CC) and Seq2Seq models on the neural conversation model test set.

Having explored possible model iterations we settled on the Seq2Seq model using pre-trained GloVe embedding trained on Twitter data. We compared the NCM, Cakechat, human gold standard, and our Seq2Seq with results shown in Figure 2.

Our overall finding is that some choices help, while other are not statistically significant. In this paper we have shown that "standard" elements like random seed, tokenization, preinitialization using word embeddings, and training time make a significant impact on results.

“It matters that you do *something* that isn’t idiotic, but it doesn’t matter *which* non-stupid thing you do.” Furthermore, “it seems like the subjective stuff (count decimation, or whatever you’re calling it when you undersample frequent things) or ditching non-question marks– that’s where the system differentiation starts. ” – Amittai Axelrod
Thus, we encourage all chatbot researchers to standardize as much as possible, so that we can compare the things that matter.

6 Discussion

We compared a wide variety of different ways of preparing the data, seeding and building the model, and evaluating it. What effected evaluations – or didn’t – often depended on the corpus being used. For both corpora we evaluated, data preprocessing had negligible effect – as long as some tokenization was used, as opposed to just white-space splitting. Using pre-trained word embeddings helped for the Twitter data set, but had no measurable effect on the OpenSubtitles data set, most likely because it is five times larger. The selection of the random seed for training the neural nets had insignificant effect on the perceived quality of the chatbot responses on either data set, in spite of **surprisingly** and significantly affecting their perplexity on the validation dataset from OpenSubtitles. Perhaps unsurprisingly, training for 16 epochs (close to convergence) rather than 5 epochs improved response quality on both datasets.

For the OpenSubtitles corpus, we found that a couple other changes significantly affected response quality: subsampling frequent responses (e.g., ’yeah’, ’IDK’, ...) to only keep the square root of the original number helped, as did only keeping prompts that end in question marks.

7 Conclusion

We have found that a number of variations in the data preprocessing, model estimation, and evaluation procedures effect the performance of chatbots, and that documentation of such design choices in the literature is uneven. When deep learning models and the preprocessing of the data used to train them are not shared, it is impossible to replicate earlier work. Not only could we not replicate a widely cited baseline, but we found that raters preferred its answers to those generated by a human. It is, of course, that the chatbot is ’superhuman’, giving more realistic answers than those given by an actual human, but it is also possible that something has gone wrong in the process. Without better documentation it is impossible to tell. If the field of non-task driven chatbots is to make progress, it is important that we do a better job of clearly reporting what we have done, and making sure that comparisons are truly apples-to-apples.

Along with this paper, we release a web form with the key information that we believe every deep learning chatbot paper should release, and we release this data for a range of baseline models across different data sets. We hope that future researchers–and the entire field–will benefit from demonstrating how better designed data preprocessing or models improve on this performance.

References

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

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.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A Knowledge-Grounded Neural Conversation Model.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2017. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Dialog System Technology Challenge 6 (DSTC6) Workshop*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A Diversity-Promoting Objective Function for Neural Conversation Models.

Jiwei Li, Will Monroe, ALan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep Active Learning for Dialogue Generation. (4).

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data Distillation for Controlling Specificity in Dialogue Generation.

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.

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.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg.

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*, pages 311–318. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Orleans, USA, June.

Verena Riesler. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

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.

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.

Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. Automatic evaluation of chat-oriented dialogue systems using large-scale multi-references.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *Natural Language Dialog Systems and Intelligent Assistants*, 37:233–239.

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? pages 1–14.

Ganbin Zhou, Ping Luo, Rongyu Cao, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. 2017a. Tree-Structured Neural Machine for Linguistics-Aware Sentence Generation.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017b. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory.