# Evaluating Coherence in Dialogue Systems using Entailment

**Nouha Dziri*** **Ehsan Kamalloo*** **Kory W. Mathewson** **Osmar Zaiane**

Department of Computing Science
University of Alberta
{dziri,kamalloo,korym,zaiane}@cs.ualberta.ca

## Abstract

Evaluating open-domain dialogue systems is difficult due to the diversity of possible correct answers. Automatic metrics such as BLEU correlate weakly with human annotations, resulting in a significant bias across different models and datasets. Some researchers resort to human judgment experimentation for assessing response quality, which is expensive, time consuming, and not scalable. Moreover, judges tend to evaluate a small number of dialogues, meaning that minor differences in evaluation configuration may lead to dissimilar results. In this paper, we present interpretable metrics for evaluating topic coherence by making use of distributed sentence representations. Furthermore, we introduce calculable approximations of human judgment based on conversational coherence by adopting state-of-the-art entailment techniques. Results show that our metrics can be used as a surrogate for human judgment, making it easy to evaluate dialogue systems on large-scale datasets and allowing an unbiased estimate for the quality of the responses.

## 1 Introduction

Recently, we have witnessed a big success in the capability of computers to seemingly understand natural language text and to generate plausible responses to conversations (Serban et al., 2016; Xing et al., 2017; Sordoni et al., 2015; Li et al., 2016; Serban et al., 2017; Devlin et al., 2018; Radford et al., 2018). A challenging task of building dialogue systems lies in evaluating the quality of their responses. Typically, evaluating goal-oriented dialogue systems is done via human-generated judgment like a task completion test or user satisfaction score (Walker et al., 1997; Möller et al., 2006). However, the task of evaluating open-ended dialogue systems is not well defined as there is no clear explicit goal for conversations. Indeed, dialog systems are ultimately created to satisfy the user's need which can be associated with how entertaining and engaging the conversation was. It is unclear how to define a metric that can account comprehensibly for the semantic meaning of the responses. Moreover, grasping the underlying meaning of text has always been fraught with difficulties, which are essentially attributed to the complexities and ambiguities in natural language. Generally, a good dialogue can be described as an exchange of information that sustain coherence through a train of thoughts and a flow of topics. Therefore, a plausible way to evaluate open-ended dialogue systems is to measure the consistency of the responses. For example, a neural dialogue system can respond to the utterance "*Do you like animals?*" by "*Yes, I have three cats*", thereafter replies to "*How many cats do you have*" by "*I don't have cats.*". Here, we can notice that the dialogue system failed to provide a coherent answer and instead generated an inconsistent response.

In this work, we characterize the consistency of dialogue systems as a natural language inference (NLI) (Dagan et al., 2006) problem. In particular, NLI is focused on recognizing whether a hypothesis is inferred from a premise. In dialogue systems, we cast a generated response as the hypothesis and the conversation history as the premise, projecting thus the automatic evaluation into an NLI task. In other words, we propose directly calculable approximations of human evaluation grounded on conversational coherence and affordance by using state-of-the-art entailment techniques. For this purpose, we build a synthesized inference data from conversational corpora. The intuition behind this choice is motivated by the fact that utterances in a human conversation tend to follow a consistent and coherent flow where each utterance can be inferred from the previous interactions. We train the state-of-the-art infer-

*Equal Contribution



*Proceedings of NAACL-HLT 2019*, pages 3806–3812
Minneapolis, Minnesota, June 2 - June 7, 2019. 

ence models on our conversational inference data and then the learned models are used to evaluate the coherence in a given conversation. Finally, we fare our proposed evaluation method against existing automated metrics. The results highlight the capability of inference models to automatically evaluate dialogue coherence. The source code and the dataset are available at https://github.com/nouhadziri/DialogEntailment

## 2 Related Work

Evaluating open-ended dialogue systems has drawn the attention of several researchers in recent years. Unfortunately, word-overlapping metrics such as BLEU have been shown to correlate weakly with human evaluation, which in turn, introduces bias against certain models (Liu et al., 2016). Many studies have been proposed to improve the quality of automated metrics. In particular, Lowe et al. (Lowe et al., 2017) introduced an automatic evaluation system called ADEM which learns to score responses from an annotated dataset of human responses scores. However, such system is heavily biased towards the training data and struggles with generalization capabilities on unseen datasets. Further, collecting an annotated gold standard of human judgment is very expensive and thus, ADEM is less flexible and extensible. Venkatesh et al. (Venkatesh et al., 2018) introduced a framework for evaluating the quality of the conversations based on topical diversity, coherence, engagement and conversational depth and showed that these metrics conform with human evaluation. However, a big part of their metrics relies on human labels, which makes the evaluation system not scalable. Recently, Welleck et al. (Welleck et al., 2018) investigated the use of NLI models (e.g., ESIM (Chen et al., 2016) and InferSent (Conneau et al., 2017)) to measure consistency in dialogue systems. They built a Dialogue NLI dataset which consists of sentence pairs labeled as entailment, neutral, or contradiction. The utterances are derived from a two-agent persona-based dialogue dataset. To annotate the dataset, they used human annotation from Amazon Mechanical Turk. In this work, we propose a method that employs NLI approaches to detect coherence in dialogue systems. The proposed evaluation procedure does not require human labels, making progress towards scalable and autonomous evaluation systems.

## 3 Natural Language Inference

Reasoning about the semantic relationship between two utterances is a fundamental part of text understanding. In this setting, we consider inference about entailment as a useful testing bed for the evaluation of coherence in dialogue systems. The success of NLI models[1] allows us to frame automated dialogue evaluation as an entailment problem. More specifically, given a conversation history $H$ and a generated response $r$, the goal is to understand whether the premise-hypothesis pair $(H, r)$ is entailing, contradictory, or neutral.

### 3.1 Coherence in Dialogue Systems

The essence of neural response generation models is designed by maximizing the likelihood of the target response given source utterances. Therefore, a dialogue generation task can be formulated as a next utterance prediction problem. In particular, the model predicts a response $u_{i+1}$ given a conversation history $(u_1, ..., u_i)$. One key factor for a successful conversation is having coherence across multiple turns. A machine's response can be considered as incoherent when it contradicts directly its previous utterances or follows an illogical reasoning throughout the whole conversation. Inconsistency can be clearly identified when it corresponds to logical discrepancy between two facts. For example, when you indicate clearly during the conversation that you have cats but when you get asked "*How many cats do you have*", you answer by "*I don't have cats.*". Nevertheless, in general, inconsistency can be less explicitly recognizable as it may describe an error between what the person has said and what she/he truly believes given her/his personality and background information.
To detect dialogue incoherence, we consider two prominent models that have shown promising results in commonsense reasoning: the Enhanced Sequential Inference Model (ESIM) (Chen et al., 2016) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018):

**ESIM** (Chen et al., 2016): employs a Bi-LSTM model (Graves and Schmidhuber, 2005) to encode the premise and the hypothesis. Also, it explores the effectiveness of syntax for NLI by encoding syntactic parse trees of premise and hypothesis through Tree-LSTM (Zhu et al., 2015). Then, the

[1]Recent models have achieved high accuracy in Stanford NLI corpus (Bowman et al., 2015) (90.1%) and GLUE Benchmark (Wang et al., 2018) (86.7%)

input encoding part is followed by a matrix attention layer, a local inference layer, another BiLSTM inference composition layer, and finally a pooling operation before the output layer. We further boost ESIM with by incorporating contextualized word embeddings, namely ELMo (Peters et al., 2018), into the inference model.

**BERT** (Devlin et al., 2018): exploits a multi-layer Bidirectional Transformers model (Vaswani et al., 2017) to learn pre-trained universal representations of text using only a plain text corpus from Wikipedia. BERT has achieved state-of-the art results on various natural language understanding tasks and has been shown to handle strongly long-range dependencies in text. BERT can be fine-tuned to achieve several tasks by solely adding a small layer to the core model. In this work, we adopted BERT to the task of NLI.

Overall, the goal of the above models is to learn a function $G_{NLI}$ that predicts one of three categories (i.e., entailment, contradiction or neutral) given premise-hypothesis pairs.

## 4 Inference Corpus for Dialogues

To train the inference models, we build a synthesized dataset geared toward evaluating consistency in dialogue systems. To this end, the Persona-Chat conversational data (Zhang et al., 2018) is used to form the basis of our conversational inference data. The continuity of utterances in human conversation facilitates the use of entailment in the dialogue domain. Typically, when we interact with one another, we tend to reference information from previous utterances to engage with the interlocutor. This is why we build our synthetic inference dataset upon a dialogue corpus. The Persona-Chat corpus is a crowd-sourced dataset where two people converse with each other based on a set of randomly assigned persona. To build an inference corpus, we need to find three different labels (i.e., *entailment*, *contradiction*, and *neutral*). For this purpose, we map an appropriate and on topic response to the *entailment* label. Consequently, the *entailment* instances are derived from the utterances in the conversations. For *contradiction*, grammatically-impaired sentences are constructed by randomly choosing words from the conversation. We also added randomly drawn contradictory instances from the MultiNLI corpus (Williams et al., 2018) to account for meaningful inconsistencies. Finally, random utterances from other conversations or generic responses such as "*I don't know*" comprise the *neutral* instances. Following this approach, we build a corpus of 1.1M premise-hypothesis pairs, namely **InferConvAI**. Table 1 summarizes the statistics of InferConvAI.

| | **Train** | **Dev** | **Test** |
|---|---|---|---|
| #entailment | 218.2K | 12.2K | 1.4K |
| #neutral | 579.5K | 28.0K | 3.1K |
| #contradiction | 261.9K | 9.8K | 1.1K |
| **Total** | 1.1M | 50.2K | 5.6K |

Table 1: Distribution of labels in the InferConvAI corpus.

## 5 Experiments

In this section, we focus on the task of evaluating the next utterance given the conversation history. We used the following models to generate responses. These models were trained on the conversational datasets, using optimization, until convergence:

- Seq2Seq with attention mechanism (Bahdanau et al., 2015): predicts the next response given the previous utterance using an encoder-decoder model.
- HRED (Serban et al., 2016): extends the Seq2Seq model by adding a context-RNN layer that accounts for contextual information.
- TA-Seq2Seq (Xing et al., 2017): extends the Seq2Seq model by biasing the overall distribution towards leveraging topic words in the response.
- THRED (Dziri et al., 2018): builds upon TA-Seq2Seq model by levering topic words in the response in a multi-turn dialogue system.

The training was conducted on two datasets: OpenSubtitles (Tiedemann, 2012) and Reddit (Dziri et al., 2018). Due to lack of resources, we randomly sampled 6M dialogues as training data from each dataset, 700K dialogues as development data, and 40K dialogues as test data. Each dialogue corresponds to three turn exchanges. To evaluate accurately the quality of the generated responses, we recruited five native English speakers. The judges annotated 150 dialogues from Reddit

| Method | Reddit | OpenSubtitles |
|---|---|---|
| ESIM + ELMo | 0.526 | 0.455 |
| BERT | **0.553** | **0.498** |

Table 2: Accuracy of inference models on InferConvAI.

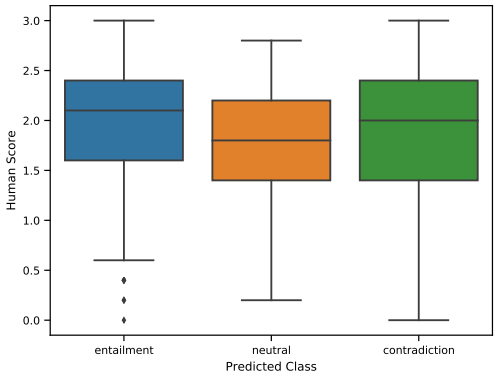


Figure 1: BERT predictions for each class vs. human scores. The labels in the horizontal axis are (from left to right): entailment, neutral, contradiction.

and 150 dialogues from OpenSubtitles. All subjects have informed consent as required from the Ethics Review Board at the University of Alberta. Due to lack of space, we will omit an exhaustive description of the human evaluation process and refer readers to (Dziri et al., 2018) as we conducted the same evaluation procedure.

### 5.1 NLI in Dialogues

In this section, we evaluate the performance of the state-of-the-art entailment models on predicting a score for the generated utterances. In particular, the conversation history $H$ is treated as a hypothesis, whereas the generated response $r$ acts as a premise. We pick two state-of-the-art NLI models (i.e., ESIM (Chen et al., 2016) and BERT (Devlin et al., 2018)). These models were trained on the InferConvAI dataset. During evaluation, we use our test dialogue corpus from Reddit and OpenSubtitles, in which the majority vote of the 4-scale human rating constitutes the labels. The results are illustrated in Table 2. Both models reach reasonable performance in this setting, while BERT outperforms ESIM. Note that this experiment examines the generalization capabilities of these inference models as the test datasets are drawn from an entirely different distribution than the training corpus. Figure 1 illustrates the performance of BERT

| Method | Pearson | |
|---|---|---|
| | Reddit | OpenSubtitles |
| $\text{SS}(H_{-2})_{\text{BERT}}$ | -0.204 | -0.290 |
| $\text{SS}(H_{-2})_{\text{ELMo}}$ | -0.146 | -0.365 |
| $\text{SS}(H_{-2})_{\text{USE}}$ | -0.248 | -0.314 |
| $\text{SS}(H_{-1})_{\text{BERT}}$ | -0.214 | -0.337 |
| $\text{SS}(H_{-1})_{\text{ELMo}}$ | -0.178 | **-0.404** |
| $\text{SS}(H_{-1})_{\text{USE}}$ | **-0.287** | -0.320 |
| $\text{A}_{\text{BERT}}$ | 0.135 | 0.131 |
| $\text{A}_{\text{ELMo}}$ | 0.085 | 0.162 |
| $\text{A}_{\text{word2vec}}$ | 0.037 | 0.196 |
| $\text{G}_{\text{BERT}}$ | 0.208 | 0.132 |
| $\text{G}_{\text{ELMo}}$ | 0.037 | 0.072 |
| $\text{G}_{\text{word2vec}}$ | -0.033 | 0.015 |
| $\text{E}_{\text{BERT}}$ | 0.162 | 0.144 |
| $\text{E}_{\text{ELMo}}$ | 0.035 | 0.116 |
| $\text{E}_{\text{word2vec}}$ | -0.065 | 0.118 |

Table 3: The Pearson Correlation between different metrics and human judgments with $p$-value $< 0.001$. The semantic similarity (SS) metric is measured with respect to the most recent utterance $H_{-1}$ and the most recent two utterances $H_{-2}$ in the conversation history. We adopt different embedding algorithms to compute the word vectors: ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), word2vec (Mikolov et al., 2013) and Universal Sentence Encoder (USE) (Cer et al., 2018).

for each class with respect to the human scores. The test utterances that are predicted as *entailment* tend to be rated higher than other utterances, exhibiting that the entailment models correlate quite well with what humans perceive as a coherent response. Another observation is that the inference models often classify acontextual and off-topic responses as *neutral* and the annotators typically dislike these types of responses. This contributes to the lower scores of *neutral*-detected responses compared to responses predicted as *contradiction*.

### 5.2 Automated Metrics

#### 5.2.1 Word-level Metrics

We consider as evaluation metrics baselines three textual similarity metrics (Liu et al., 2016) based on word embeddings: Average (A), Greedy (G), and Extrema (E). These word-level embedding metrics have been proven to correlate with human judgment marginally better than other world-overlap metrics (e.g., BLEU, ROUGE and METEOR) (Liu et al., 2016). One critical flaw of these embedding metrics is that they assume that

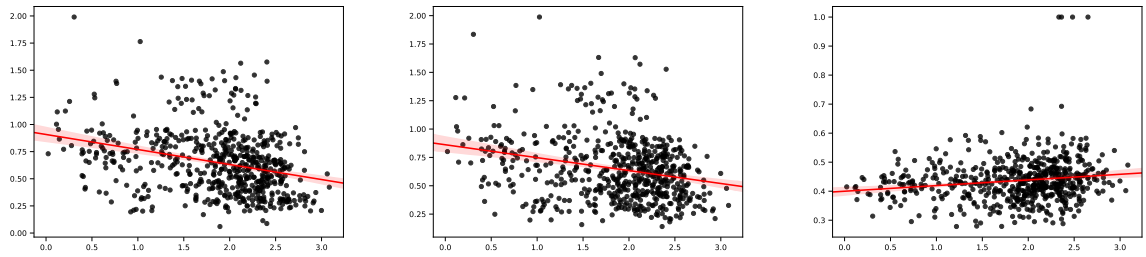

Figure 2: Scatter plots illustrating correlation between human judgment and the automated metrics on the Reddit test dataset. In order to better visualize the density of the points, we added stochastic noise generated by Gaussian distribution $\mathcal{N}(0, 0.1)$ to the human ratings (i.e., horizontal axis) at the cost of lowering correlation, as done in (Lowe et al., 2017). From left to right: $\text{SS}_{\text{USE}}$ w.r.t. the second most recent utterance ($H_{-2}$), $\text{SS}_{\text{USE}}$ w.r.t. the most recent utterance ($H_{-1}$), and $\text{Extrema}_{\text{BERT}}$

each word is independent of the other words in the sentence. Further, the sentence is treated as a bag-of-words, disregarding words order and dependencies that are known to be substantial for understanding the semantic of a sentence. The correlation of these metrics with human judgment is showcased in Table 3. We can notice that the three metrics A, G and E correlate weakly with human judgment in both datasets, demonstrating the need for a well-designed automated metric that provides an accurate evaluation of dialogues.

#### 5.2.2 Semantic Similarity

The Semantic Similarity (SS) metric was suggested by (Dziri et al., 2018). It measures the distance between the generated response and the utterances in the conversation history. The intuition of this metric revolves around capturing good and consistent responses by showing whether the machine-generated responses maintain the topic of the conversation. In this project, we measured SS with respect to two different utterances, the conversation history $H$ and the most recent utterance $H_{-1}$. The conversation history is formed by concatenating the two most recent utterances. We report the Pearson coefficient of this metric with human judgment in Table 3. The SS metric is expected to have a negative correlation as the higher human ratings correspond to the lower semantic distance. The results demonstrate that SS metrics correlate better than word-level metrics as they make use of word interactions to represent utterances. Moreover, the Universal Sentence Encoder (USE) (Cer et al., 2018) model performs better on Reddit, whereas the ELMo embeddings achieve higher correlation on OpenSubtitles. This arguably underlines that deep contextualized word representations can manage better complex characteristics of natural language (e.g., syntax and semantics). The SS metric, which requires no pretraining, reaches a Pearson correlation of -0.404 with respect to the most recent utterance on OpenSubtitles. Such correlation can be compared with a correlation of 0.436 achieved by ADEM (Lowe et al., 2017) which required large amounts of training data and computation. Moreover, in order to investigate whether the results in Table 3 are in line with human evaluation, we visualized the correlation between the human ratings and SS metric as scatter plots in Figure 2.

## 6 Conclusion

Evaluating dialogue systems has been heavily investigated, but researchers are still on the quest for a strong and reliable metric that highly conforms with human judgment. Existing automated metrics show poor correlation with human annotations. In this paper, we present a novel paradigm for evaluating the coherence of dialogue systems by using state-of-the-art entailment techniques. We aim at building a system that does not require human annotation, which in turn, can lead to a scalable evaluation approach. While our results illustrate that the proposed approach correlates reasonably with human judgment and provide an unbiased estimate for the response quality, we believe that there is still room for improvement. For instance, measuring the engagingness of the conversation would be helpful in improving evaluating different dialogue strategies.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly

learning to align and translate. *international conference on learning representations*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1116–1126.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sebastian Möller, Roman Englert, Klaus Engelbrecht, Verena Hafner, Anthony Jameson, Antti Oulasvirta, Alexander Raake, and Norbert Reithinger. 2006. Memo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In *Ninth International Conference on Spoken Language Processing*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

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. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

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. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*.

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

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1112–1122.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generatio. 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*.

Xiaodan Zhu, Parinaz Sobihani, and Hongyu Guo. 2015. Long short-term memory over recursive structures. In *International Conference on Machine Learning*, pages 1604–1612.