Cue-bot: A Conversational Agent for Assistive Technology

Shachi H Kumar, Hsuan Su, Ramesh Manuvinakurike,
Maximilian C Pinaroc, Sai Prasad, Saurav Sahay and Lama Nachman
Intel Labs, Santa Clara, CA, USA
*{shachi.h.kumar, hsuan.su, ramesh.manuvinakurike,
maximilian.c.pinaroc, sai.prasad, saurav.sahay, lama.nachman }@intel.com*

Abstract

Intelligent conversational assistants have become an integral part of our lives for performing simple tasks. However, such agents, for example, Google bots, Alexa and others are yet to have any social impact on minority population, for example, for people with neurological disorders and people with speech, language and social communication disorders, sometimes with locked-in states where speaking or typing is a challenge. Language model technologies can be very powerful tools in enabling these users to carry out daily communication and social interactions. In this work, we present a system that users with varied levels of disabilties can use to interact with the world, supported by eye-tracking, mouse controls and an intelligent agent Cue-bot, that can represent the user in a conversation. The agent provides relevant controllable 'cues' to generate desirable responses quickly for an ongoing dialog context. In the context of usage of such systems for people with degenerative disorders, we present automatic and human evaluation of our cue/keyword predictor and the controllable dialog system and show that our models perform significantly better than models without control and can also reduce user effort (fewer keystrokes) and speed up communication (typing time) significantly.

1 Introduction

Conversational agents, especially systems such as Alexa and Google Home, have become commodity items in people's homes. Such systems have enabled carrying out one-shot tasks such as setting reminders, playing music and accessing information simpler for the general population. We also have other PC and cloud based chatbots that are designed to perform certain goals or tasks, or to just engage in a casual conversation/chat with a user. The latter class of open-domain conversational agents have not yet seen widespread adoption besides mostly research exploration projects for developing conversational agents (Ram et al., 2018).

Large language models are being developed today with end-to-end pre-training. Large-scale pre-training has attained significant performance gains across many tasks within NLP (Devlin et al., 2019; Radford and Narasimhan, 2018), including intent prediction (Castellucci et al., 2019; Chen et al., 2019) and dialogue state tracking (Heck et al., 2020). Open-domain chatbots are also being trained using generative language modeling objective of minimizing perplexity on next word prediction task using large conversational corpora and transformer based models. These models have demonstrated surprising generality, with models like DialoGPT (Zhang et al., 2020b), Meena (Adiwardana et al., 2020) and Blender (Roller et al., 2020) achieving response generation performance competitive with humans in certain settings. These improving systems still suffer from issues such as repeated responses, hallucinated facts, and lack of controllability, grounding and embodiment (See et al., 2019).

With the availability of these pre-trained language enabling models, novel products and applications are emerging in several domains (Bommasani et al., 2021). One such accessesibility application we are exploring is aimed towards leveraging language modeling technology to support minority group of people with certain disabilities [1] to communicate with others effectively. One such example is Amyotrophic Lateral Sclerosis (ALS) or Motor Neuron Disease(MND), a progressive, degenerative, neurological disorder where people lose their muscle movement, voice and the ability to carry out a normal day-to-day communication. There have been technologies and platforms, one such example is Assistive Context-Aware Toolkit (ACAT)[2],

[1] According to WHO, there are more than 1 Billion people with disabilities

[2] https://01.org/ACAT

*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
System Demonstrations*, pages 196 - 203

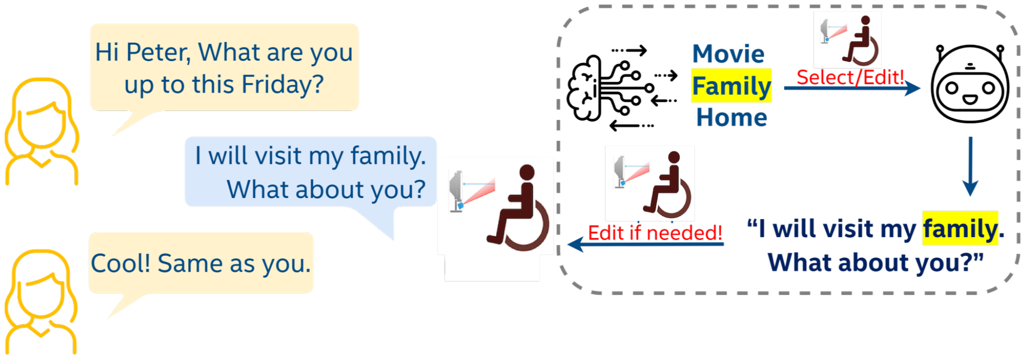

Hi Peter, What are you up to this Friday?
I will visit my family. What about you?
Cool! Same as you.
Movie
Family
Home
Select/Edit!
Edit if needed!
"I will visit my family. What about you?"

Figure 1: A dialog system for an assistive use-case can listen to a conversation and provide diverse cues to the user. These cues, provide human control to the dialog system that can generate relevant responses that can further be edited.

that enable these users to communicate, but it takes huge effort and time for these patients to communicate sentences character by character using various data input mechanisms that suit their situation such as gaze, fingers or muscle movements.

We want to enable full and faster communication and provide interaction support tools for people with such disabilities by having an intelligent agent be their voice and content assistent. The system should use very limited user input (e.g. gaze, single muscle movement, facial gesture, etc) and suggest cues and cue-based responses that can be interactively chosen and edited for near real time social interactions. The goal of such a system is to minimize the effort by minimizing the keystrokes input required for continued coherent interactions. Today's response generation systems suffer from several issues and are very hard to use as-is for our usage requirements. The system for our usage needs to be context-aware, personalized, should enable minimal user-intervention and most importantly, be assistive and controllable by the user. Fast response generation with response cues, editing and auto complete features can dramatically reduce the silence gap in the conversation resulting from users slower keystroke by keystroke input.

Our contributions in this work are, i) **Minority Group Application:** We bring forth a novel usage for open-domain chatbots/response generation systems, i.e., designing a reponse generation system that will represent users with communication disabilities and help them fulfill their day-to-day communication needs. ii) **Minimal user effort and intervention:** We show that the keyword predictor models can speed up communication time by suggesting cues to the user. We also present a technique for controllable response generation using these cues. We present human and automatic evaluation for this approach. iii) **Demo Interface:** We showcase a demo where a user can interact only using his/her eyes to control the interface, with minimum effort and time.

2 Motivation

To enable people with MND and other disabilities to communicate, Intel Labs developed ACAT, an open source platform that was originally developed for Professor Stephen Hawking. With ACAT, users have complete access to the capabitlies of their computers that they can control using various modalities and sensors such as proximity sensors, eye-gaze and further capabilities such as BCI-based controls are being developed. ACAT also includes word prediction, speech synthesis capabilities, this allows users to respond to ongoing conversations, and a range of tasks such as accessing emails, editing documents and browsing the web.

In ACAT, users can choose words that appear from the word predictors, or select letters to create words using the input modality that suits their condition. While this empowers users to communicate, this still involves a lot of effort in terms of the word/letter selections and involves a huge latency. With this work, we aim to reduce the user effort and intervention and also the time in generating a user response, by using the state of the art language modeling technology as will be described in the below sections. Our goal is to also integrate this work into the current ACAT system to enable the user to select entire responses based on input keywords, with minimum effort and intervention.

3 System Architecture

Figure 1 shows the interaction flow of the Cue-bot system. Consider an ongoing conversation between an interlocutor and the user. An Automatic Speech Recognition (ASR) system converts the interlocutor's current utterance to text. This, in addition to the dialog context, is input into the cue-generator model (described below), which outputs the possible cues or keywords for a potential response to the input that the user might want to respond with. In the figure, given the utterance "Hi Peter, what are you upto this Friday?", the cue-generator generates "movie", "family", "home" that the user can choose from. The interface also allows the user to enter his/her own cue or keyword, if the user needs, in case none of the suggested words are relevant. The user uses the Tobii eye-tracker and the OptiKey mouse controls to make a keyword selection or to type out a keyword of his/her choice.

Once the user chooses or enters a custom keyword, this information is sent to the Cue/Keyword-based Response Generation system (details below) that generates multiple responses relevant to the cue/keyword. The user can 1) choose one of these responses to use as his/her response to the interlocutor, 2) edit one of the responses or 3) type out the entire response if none of the suggested responses are relevant. These modules are described in detail below.

3.1 Models

The main software components of the system include the cue/keyword generator and the response generation models which are described in the sections below.

3.1.1 Cue/Keyword Generator

In order to minimize the keystrokes in the interaction and hence user effort, we build a model that can generate keywords that could aid in generating the user's response in the conversation. We present two types of keyword generators in this section - extractive and generative. To train the model, we obtain the data by extracting 'key' terms from the dataset. This data is generated automatically, hence enabling end-to-end automatic pipeline, without the need for any other additional data collection or labeling efforts. Given a conversation context and a response output, keywords are extracted from the response utterance and incorporated into the model. We use keyBERT (Grootendorst, 2020) to extract meaningful keywords from the responses. This technique uses BERT-embeddings and cosine similarity to find the sub-phrases in a document that are most similar to the document itself.

Extractive keyword predictor: Given a conversation context, we use DialoGPT(Zhang et al., 2020b) with diverse beam search(Vijayakumar et al., 2018) to generate multiple responses (we use 10 beams, 2 groups and diversity_penalty of 5.5). We then use keyBERT(Grootendorst, 2020) to extract keywords from the beam outputs and present these as keyword suggestions.

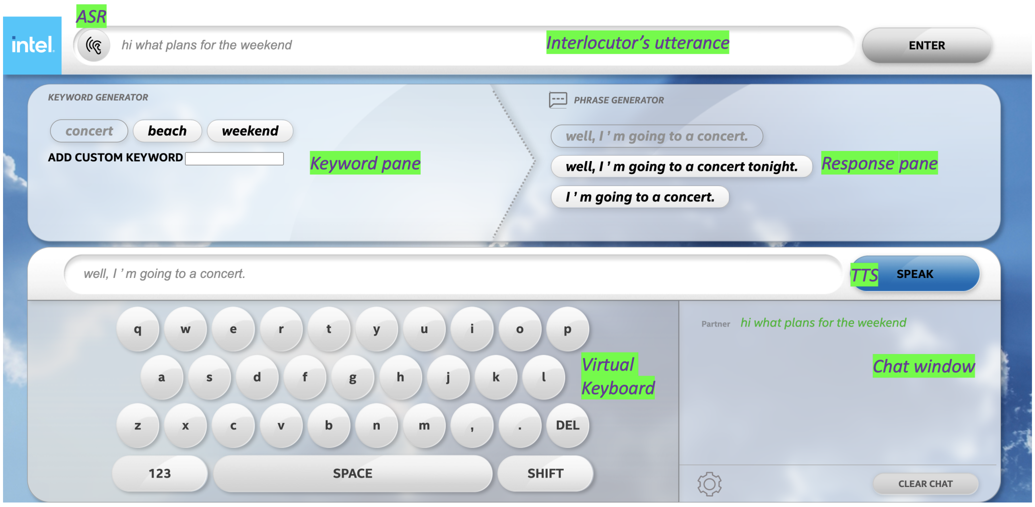
Generative keyword predictor: We fine-tune a large pretrained language model, GPT2, to generate keywords for a given context, and present these as suggestions. We use the training and validation dataset from DailyDialog (Li et al., 2017a) to build the keyword predictor. For evaluation of these models, we use the top keyword prediction. We further use diverse beam search (same configuration as above) and generate multiple keyword suggestions.

Cue/Keyword based Response Generation Given the conversation context, we enable fine-grained control over the responses generated by training the model with important keywords automatically generated (as described above). For a given conversation context, we incorporate keywords into the model by adding new keyword-specific-tokens, in addition to dialog-state/speaker tokens that represent speaker turns in the dialog. We further extend the dialog-state embeddings to add 'keyword-state-embeddings' with special keyword separator token to indicate the positions of the keyword tokens.

In this work, we modify the HuggingFace TransferTransfo model (Wolf et al., 2019) architecture, a model is similar to the Transformer based architecture from (Radford and Narasimhan, 2018) that uses autoregressive and discriminative fine-tuning by optimizing a combination of two loss functions : 1) language modeling loss and 2) next-utterance classification loss. We incorporate fine-grained keyword-based control as model inputs and fine-tune this model on the DailyDialog dataset with multi-task objective.

3.2 Other Components

Eye-Tracker To support users with severe neurological disabilities who are unable to move, speak or type, we enable interaction with the system



hi what plans for the weekend

well, I ' m going to a concert.

well, I ' m going to a concert tonight.

I ' m going to a concert.

well, I ' m going to a concert.

Partner hi what plans for the weekend

Figure 2: Cue-bot interface

through an additional input modality: eye-gaze tracking. We use the Tobii eye-tracker [3], (specifically the gaming version to lower the cost of the system), that works with Windows systems and can be mounted on the laptop or any other external monitor. This device needs to be calibrated to work for a user. This device has already been in use with the ACAT system supporting users with MND.

Mouse Control While the Tobii eye-tracker tracks the user's gaze, we need to translate the gaze to a mouse-click event. For this, we use OptiKey [4], which is an on-screen keyboard designed for users with MND, to interact with Windows systems. OptiKey can be integrated with the Tobii eye-tracker to allow users to control the system using eye-gaze only. We modify the OptiKey software to show the specific buttons needed to control the UI, such as left-click (single and double), right click, scroll up-down, finer mouse movement (in pixels).

Automatic Speech Recognition In a real system, where the user is communicating with an interlocutor, we need the cue-bot to be listening to the conversation in order to make relevant keyword and response suggestions to the user. To incorporate this in the web-interface as well, we integrate Google ASR that converts the interlocutor's speech to text that can be input into our models. This is enabled on a button-click on our user-interface as shown in Figure 2.

User Interface Design Figure 2 shows the user interface for this system. The top text area shows the placeholder for the interlocutor's utterance, that is obtained by converting speech-to-text using ASR. The interface is divided into two parts, the top area is further split into two panes 1) the left pane displays the generated keywords from the keyword predictor. The user can also add a custom keyword by clicking on the 'Add Custom Keyword' button. Once a keyword choice is made, 2) the right pane displays the generated responses from the keyword-based response generation model. The bottom area shows the virtual keyboard with buttons large enough to enable the gaze-tracker to detect gaze without ambiguities. Picking one of the generated responses from the right phrase pane, populates it into the textarea which can be edited by the user if needed. The 'Speak' button converts the user's response to speech. Finally, the chat window on the bottom-right keeps track of the ongoing conversation for the user's reference.

4 Experimental Setup

We initialize the TransferTransfo model with weights of DialoGPT 'medium' model with 345M parameters. We also use two candidates for the next utterance prediction task. We use a batch_size of 64 for training, nucleus sampling for generation with top_p set to 0.9. We fine-tune the model for 3 epochs. We compare the model trained without any information with a keyword-context model trained with keyword as auxillary input information.

[3] https://gaming.tobii.com/product/eye-tracker-5/
[4] https://github.com/OptiKey/OptiKey

4.1 Datasets

We use the Dailydialog dataset (Li et al., 2017b), which consists of 13,118 daily conversations involving various topics such as tourism, culture, and education among others. This dataset serves as a starting point for AAC applications as it contains suitable interactions for building applications to support social communication and daily life interactions. The training set has about 11,000 conversations, and the validation and test sets have 1000 conversations each. We use the test set, consisting of 6740 context-response pairs, to evaluate our models which will be discussed in the results section.

4.2 Automatic Evaluation

We use several automatic metrics to compute the performance of our models.

Metrics for Evaluating Keyword Predictor Models The keyword predictor model should be able to generate diverse keywords to present varied options for users to choose from. We evalute the extractive and generative models based on averaged cosine similarity between generated keywords as a measure of diversity; lower the similarity, higher the diversity. We hypothesize that meaningful keywords will result in generation of meaningful and context-relevant responses. Hence, we compute 'human-like' and coherence scores for the generated responses using DialogRPT (Gao et al., 2020), a model trained to predict human feedback dialogue responses.

Metrics for Evaluating Controllable Response Generation Model :

1) **Keyword Insertion Accuracy(KIA):** To evaluate the ability of the response generation model to induce a keyword into the response, thus enabling fine-grained control, we compute the keyword-insertion accuracies of the models.

2) **Similarity Based Metrics:** Because we intend to generate responses based on keywords, computing measures of similarity between the generated response and ground truth response (in the learnt embedding space) gives a good assesment for the model performance. We use BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2020a) , Sentence-BERT (Reimers and Gurevych, 2019) to compute similarity between generated response and ground truth.

3) **Response Quality Metrics:** For response quality aspects of fluency and context-coherence, we perform language model based evaluation. We also perform n-gram based diversity evaluation. We also measure the perpelexity (PPL) by employing a pre-trained GPT-2 "medium" model.

4.3 Human Evaluation

Keystrokes and Typing Time One of the main focus of this work is minimizing user effort, time and intervention. With this in mind, we evaluate and compare the number of keystrokes and typing time taken by a user with and without our models (keyword prediction+response generation models). Please note that in absence of our models, the user will need to type out the entire response character by character. We consider two scenarios, 1) user picks a suggested keyword *(#keystrokes=1)*, 2) user enters his/her own keyword *(#keystrokes=#characters entered)*. We also consider edited responses *(#keystrokes=1)* and non-edited responses *(#keystrokes=#edits)*.

Evaluation of the keyword-based response generation models We randomly pick 100 dialog contexts and present the context along with the keyword and pairs of responses from the models and ask 3 annotators to rate the responses based on the following criteria: 1) Fluency: how natural and fluent the responses are, 2) Generic: are the responses too generic given the dialog context?, 3) Context relevance: how relevant and coherent is a response to a given dialog context, 4) Keyword relevance: how relevant is a response to the input keyword? We present pairs of responses from the no-keyword and the keyword-based model, and provide 4 options for for each of the above criteria: A better than B, B better than A, Both and, Neither.

5 Results

5.1 Automatic Evaluation Results

Keyword Predictor Models: From Table 2, we can observe that the generative keyword predictor tends to generate more diverse keywords (lower similarity score), which is very important in our use-case. The responses generated by choosing the keywords from the generative predictor are more coherent and human-like.

Cue/Keyword controlled models: Table 1 shows the performance of the response generation models. From the table, the KIA for the no_kw model is negligible, given the one to many nature

	KIA	Similarity	BLEURT	BERT Score	Context	Diversity	Fluency	PPL↓
no_kw	0.083	0.271	-1.035	**0.868**/0.836/0.851	0.541	1.592	**0.407**	**39.098**
kw_context	**0.672**	**0.539**	**-0.607**	0.844/**0.853/0.868**	**0.568**	**1.789**	0.403	41.752

Table 1: Performance of the keyword-based response generation model

Kw Predictor	Coherence	Human-like	Diversity↓
Generative	**0.903**	**0.641**	**0.227**
Extractive	0.891	0.595	0.265

Table 2: Evaluation of keyword predictor models.

of open domain dialog. By guiding the model with cues or keywords, the KIA goes up to 67.2%. The cue/keyword based model outperforms the *no_kw* model in all of the similarity-based and response quality metrics, except perplexity where the *no_kw* model is lower.

5.2 Human Evaluation Results

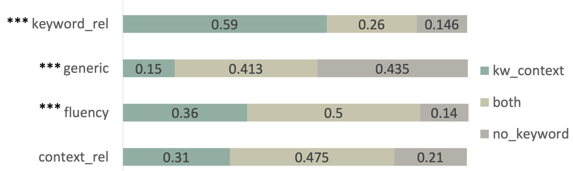


Figure 3: Results from human evaluation. (One-Sample Wilcoxon Signed Rank Test (mu=0) for the statistical tests.*** p<0.001, ** p<0.01, * p<0.05.)

Keystrokes & Typing time We compute both the interaction time and keystrokes to compare the keyword-based interaction models with typing out the entire sentences for 2 scenarios: 1) keyword picked from suggestion and 2) custom keyword entered. For 1) on average, using our models, it takes only 10% of the keystrokes taken to type out the entire sentence, and it takes 30% of the time to type the entire sentence, i.e., 70% of time is saved. For case 2) with our system, it takes about 35% of the keystrokes taken to type out the entire responses (with edits) and saves about 40% of the time to type the entire sentence.

Keyword-based response generation evaluation Figure 3 shows the scores for the response quality metrics for different model. From human ratings, we observe that the *kw_context* model outperforms the model without control, on all metrics significantly. The keyword-based model generates more fluent and relevant responses while at the same time, generating less generic responses compared to the *no_keyword* model.

6 Conclusion and Future Work

In this work, we present a system to support users with MND and other disabilities to carry out day to day social interactions with lesser effort, time and interventions. We use input modalities such as gaze tracking that allows users to control the entire interface only using their eyes. We build models that utilizes the ongoing conversations and suggests possible cues/keywords that the users can use, and generate relevant responses based on the selected keyword. We show through automatic and human evaluations that our models are better than the models without control and also save significant time and effort in interactions. For future work, we aim to integrate it with the ACAT toolkit that already supports MND users, to improve their quality of communication. We also aim to personalize the system by using user's data when available and also build a system that can continually learn through user interactions.

7 Ethics

CueBot aims to support users with disabilities and allow them to communicate while also enabling them to control the response generation. The system has been evaluated with automatic and human evaluation via AMT, where the AMT workers were fairly compensated (average >$15 per hour). Our tasks involved providing responses from humans and model which were rated by the AMT workers. We tried to mitigate any bias in the choices made by turkers by constantly shuffling the responses that we present. In our experiment we didn't collect any additional personal details (other than those collect by AMT by default) or identities from AMT workers', hence preserving their privacy. The next steps is to integrate this system with ACAT to enable user studies with ALS patients and further gain their feedback to improve the AI modules. Both the keyword suggestion and response generation modules use pre-trained language model DialoGPT (Zhang et al., 2020c) finetuned on DailyDialog dataset conversations. Given this, the responses generated could contain improper content or bias (from the large dataset these models are pre-trained on). This raises some important ethical questions that we intend to tackle as part of future work. In this current work we have not explored bias mitigation, which will also be a part of future work.

8 Acknowledgements

We thank the Torrey Frank (Intel Labs) and the team from French & Western for working with us in building the UI.

References

D. Adiwardana, Minh-Thang Luong, D. So, J. Hall, Noah Fiedel, R. Thoppilan, Z. Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, et al. 2021. On the opportunities and risks of foundation models.

Giuseppe Castellucci, Valentina Bellomaria, A. Favalli, and R. Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint bert-based model. *ArXiv*, abs/1907.02884.

Qian Chen, Zhu Zhuo, and W. Wang. 2019. Bert for joint intent classification and slot filling. *ArXiv*, abs/1902.10909.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking-training with large-scale human feedback data. In *EMNLP*.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

A. Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational ai: The science behind the alexa prize.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.