CueBot: Cue-Controlled Response Generation for Assistive Interaction Usages

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

Conversational assistants are ubiquitous among the general population, however, these systems have not had an impact on people with disabilities, or speech and language disorders, for whom basic day-to-day communication and social interaction is a huge struggle. Language model technology can play a huge role in empowering these users and help them interact with others with less effort via interaction support. To enable this population, we build a system that can represent them in a social conversation and generate responses that can be controlled by the users using cues/keywords. For an ongoing conversation, this system can suggest responses that a user can choose. We also build models that can speed up this communication by suggesting relevant cues in the dialog response context. We introduce a keyword-loss to lexically constrain the model response output. We present automatic and human evaluation of our cue/keyword predictor and the controllable dialog system to show that our models perform significantly better than models without control. Our evaluation and user study shows that keyword-control on end-to-end response generation models is powerful and can enable and empower users with degenerative disorders to carry out their day-to-day communication.

1 Introduction

Conversational agents such as Google Home and Alexa have become almost an integral part of homes and used by people of all ages to carry out tasks such as setting reminders, playing music and accessing information. There are also agents that can simply engage in chit-chat conversations, however, these open domain conversational agents have mostly been research explorations (Ram et al., 2018). Large-scale pre-training has attained significant performance gains across many tasks within Language Modeling (Devlin et al., 2019; Radford and Narasimhan, 2018), including intent prediction (Castellucci et al., 2019; Chen et al., 2019b) and dialogue state tracking (Heck et al., 2020). These pretrained language models have demonstrated surprising generality in open domain dialog tasks, with models like DialoGPT (Zhang et al., 2020b), Meena (Adiwardana et al., 2020) and BlenderBot (Roller et al., 2020) achieving performance competitive with humans in certain settings. With the availability of these models, novel products and applications are emerging (Bommasani et al., 2021) such as Communication Systems (eg. email response completion (Chen et al., 2019a)), Creativity Tools (story writing assistance (Roemmele and Gordon, 2018; Roemmele, 2021)), Human-AI collaboration for Software Engineering (Chen et al., 2021), biosciences (protein structure prediction (Rives et al., 2020) and several others.

One such accessibility 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. For example, Amyotrophic Lateral Sclerosis (ALS) is a progressive, degenerative, neurological disorder, where people lose their muscle movement, voice and the ability to carry out a normal day-to-day conversation. It takes huge effort and time for these patients to use existing systems [2] to communicate sentences character by character using various data input mechanisms available to them (gaze, fingers, muscle movements). Henceforth, we will use the term 'user' for such patients with disabilities, for whom our system is intended to support.

Our goal is to empower these users to communicate faster by having an intelligent agent be their voice and reduce the silence gap in the conversation resulting from users slower keystroke inputs.

[*] These authors contributed equally to this work

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

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

Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), pages 66 - 79
May 27, 2022 ©2022 Association for Computational Linguistics

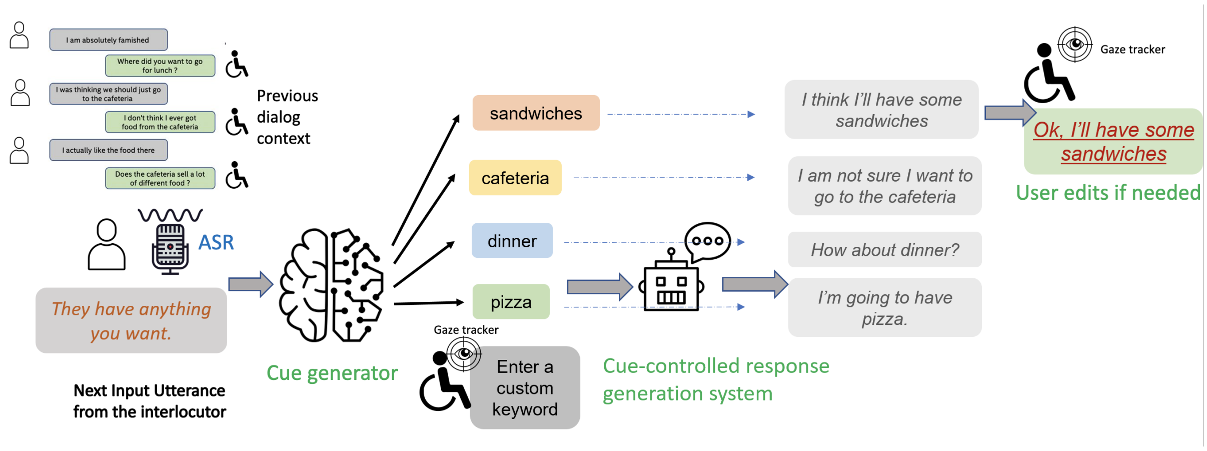

I am absolutely famished
Where did you want to go for lunch ?
I was thinking we should just go to the cafeteria
I don't think I ever got food from the cafeteria
I actually like the food there
Does the cafeteria sell a lot of different food ?
Previous dialog context
ASR
They have anything you want.
Next Input Utterance from the interlocutor
Cue generator
sandwiches
cafeteria
dinner
pizza
Gaze tracker
Enter a custom keyword
Cue-controlled response generation system
I think I'll have some sandwiches
I am not sure I want to go to the cafeteria
How about dinner?
I'm going to have pizza.
Gaze tracker
Ok, I'll have some sandwiches
User edits if needed

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 could be edited.

The system needs to listen to an ongoing conversation (using automatic speech recognition(ASR)) and should use be able to use very limited user input and suggest responses that can be interactively chosen and edited for near real time social interactions. Such a system needs to be context-aware (contexts such as ongoing conversations, user's emotions, environment), personalized (language usage and style of the user and also be aware of users' interests/likes and dislikes) and most importantly, controllable. In this work, we focus on the controllability aspect and design the control mechanism via keywords in the system. We present the following contributions: i) **Minority Group Application:** We bring forth a novel usage for response generation systems, i.e., to represent users with disabilities and help them in their day-to-day communication needs. ii) **Minimal user intervention:** We present a human-controllable response generation using keywords/cues. We also build keyword/cue predictor models that further speed up communication time and evaluate these. iii) **Keyword Loss:** We introduce a keyword loss to our training objective that further helps in incorporating soft lexical constraints in the form of keywords/similar words in the generated responses, validated through automatic and human evaluation. We also present a user-study to understand the usefullness and the effectivness of our overall system.

Figure 1 shows the interaction flow of our system. An ASR system converts an ongoing conversation (between an interlocutor and a user with disabilities) to text, which is input to the cue/keyword generator that generates possible, relevant cues that the user might want to respond with. The user can choose one of these keywords or also enter his/her own keyword to control the system. This keyword is an input to the response generator model that can generate relevant responses based on the keyword. The user can then either use one of the suggested responses or edit a response with just a few keystrokes, thus drastically reducing communication time.

2 Related Work

Assistive Technologies: Various AI technologies have proven to be helpful for people with limited mobility, hearing capabilities and speech impairments. (Brady et al., 2013; Guo et al., 2020; MacLeod et al., 2017; Elakkiya, 2020; Mišeikis et al., 2020; Ozawa et al., 2020; Ramli et al., 2020; Shor et al., 2019). People with ALS need Augmentative and Alternative Communication (AAC) strategies to address and support daily communication, such as speech generation (Beukelman et al., 2011), eye-tracking tools (Gibbons and Beneteau, 2010) and Brain Computer Interaction (BCI) interfaces (Wolpaw et al., 2018). (Linse et al., 2018). Current systems use interfaces with inputs via eye-gaze, touch or BCI (Orhan et al., 2011) with some predictive text capability and some systems using simpler n-gram based language models (Verbally, 2021; TherapyBox, 2021) and do not exploit the potential of using response generation technology using deep learning based language models. There has also been some work on collecting AAC communication data for language modeling (Vertanen, 2013), (Vertanen and Kristensson, 2011). While this data could be used to support single-turn retrieval-based dialog systems, these do not support multi-turn dialog response generation. To the best of our knowledge, there aren't many research explorations for conversational technology based applications that exploit the latest language modeling techniques for people with ALS.

Controllable Generation: Controllability in text generation and dialog systems has emerged as an active research area. (Keskar et al., 2019) pretrain a conditional transformer model with different types of control codes. (Xu et al., 2020b) and (Xu et al., 2020a) presents a keyword controlled story and dialog generation respectively. While (Ghazvininejad et al., 2017; See et al., 2019) use post-processing techniques to control generated text, (Dathathri et al., 2020) present a plug-and-play architecture, where the base language model is untouched and small attribute models induce control, further extended in (Madotto et al., 2020). (Smith et al., 2020) and (Gupta et al., 2020) control generation using style and semantic exemplars. However, these controllable attributes are too broad and not suitable for our use-case. These techniques also require a lot of computational resources which is not feasible in real-time assistive applications.

Similarity-based Loss Function: For improving the generated response, some recent work has focused on addressing the loss functions during model training. (Kovaleva et al., 2018) use similarity-based losses to enhance the diversity and meaning in the generated sentence. (Sha, 2020) aims to lexically constrain the language generation at word level. In our work, we aim to compute the loss across the entire sentence to guide keyword generation.

3 Keyword and Response Modeling

3.1 Controllability using cues/keywords

In order to make response generation controllable with minimum user-intervention, we incorporate cues/keywords as input control to generate relevant responses to a given dialog context. We enable keyword-control by 1) providing automatically generated keywords as auxillary input to the model and 2) by introducing a novel keyword-based loss that encourages the model to generate sentences containing the keyword or words semantically similar to the keyword. In the working system, the keyword is either entered by the user or selected by the user from a set of provided options. To generate data to train such a model, given a conversation context and a response output, we automatically extract keywords from the responses using keyBERT (Grootendorst, 2020) and use the HuggingFace TransferTransfo model (Wolf et al., 2019) as our base architecture. We use the top 1-gram keyword for each dialog response, and use both single keywords and multiple keywords as inputs.

3.1.1 Keywords as context

For a given conversation context, we incorporate keywords into the TransferTransfo 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.

3.1.2 Keyword-based loss functions

We propose keyword-based loss functions that encourage the occurence of the input keyword(s) in the generated sentence. We introduce variations to this loss function to enable the generation of semantically similar word to the input keyword as well as incorporate multiple-keyword inputs as control to the model. With addition of this loss, the overall loss of the model is a combination of : language model loss L_m, next sentence prediction loss L_n (both part of the TransferTransfo architecture) and keyword loss L_k,

$$\text{Overall Loss}, L = \alpha L_m + \beta L_n + \gamma L_k \qquad (1)$$

where α, β and the γ are the hyper-parameters.

Keyword Loss: In order to encourage the generation of the cue/keyword in a sentence, we maximize the similarity between the keyword, kw, and one of the generated words (at some output position). From the probability distribution (generated logits), we compute the negative log of the probability of the keyword (p_i) at every timestep i=1 to T. We then take the minimum of these scores across the generated sentence as the loss w.r.t keyword K,

$$L_k = \min_{i=1}^{T}(-\log p_i(kw)), \qquad (2)$$

Keyword Loss with similar words: We incorporate embedding-based similarity scores into the keyword loss computation as shown in equation 3 in order to encourage generation of not just the keywords, but also semantically similar words in the sentence. Let $pool = kw \cup sim\_words(kw)$.

The Keyword loss L_k,

$$\begin{aligned} L_k &= sim(k, kw)\min_{i=1}^{T}(-\log p_i(k)), \\ where\ k &= \arg\min_{x\in pool}(\min_{i=1}^{T}(-\log p_i(x))) \end{aligned} \tag{3}$$

Keyword Loss with multiple inputs : Consider $k_1, k_2 ... k_N$ as the N multiple control inputs, where it is desirable that the generated output contains all of the keywords (or similar words). To enable this, we minimize the negative log probability for each keyword, k_j, across the entire sentence and add these scores as the total loss.

$$L_k = \sum_{j=1}^{N}\min_{i=1}^{T}(-\log p_i(k_j)) \tag{4}$$

3.2 Keyword Generation

While keyword-controlled responses reduce the interaction time significantly, we try to further improve the experience and minimize the latency by automatically suggesting keywords to the user. We build two types of models:

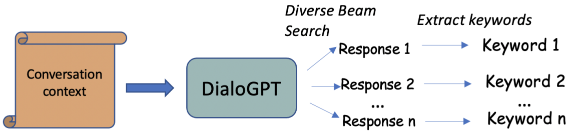
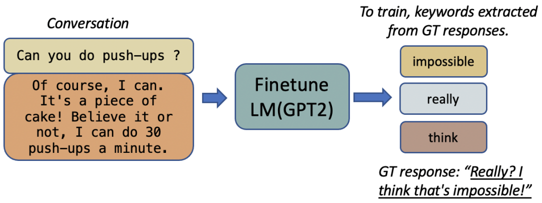


Figure 2: Extractive Keyword Predictor

1) Extractive keyword predictor: Figure 2 shows the extractive keyword predictor. Given a conversation context, we use DialoGPT(Zhang et al., 2020c) 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.

2) Generative keyword predictor: To train a generative keyword predictor, we finetune GPT2 using a conversation context as input and keywords from the ground truth response as output. Figure 3 shows this process. The model generates multiple keywords for a given context using diverse beam search and presents these as suggestions. Keywords are extracted from the DailyDialog dataset (Li et al., 2017) to create the data to train the keyword predictor.


Conversation
Finetune LM(GPT2)
To train, keywords extracted from GT responses.
impossible
really
think
GT response: "Really? I think that's impossible!"

Can you do push-ups ?

Of course, I can. It's a piece of cake! Believe it or not, I can do 30 push-ups a minute.

Figure 3: Generative Keyword Predictor

4 Experimental Setup

We initialize the TransferTransfo architecture weights of DialoGPT 'medium' model with 345M parameters. Language modeling and multiclass-classification coefficients, α and β are set to 1 as in the original model. We use a batch_size of 64 for training, nucleus sampling for generation with top_p set to 0.9 to fine-tune the model for 3 epochs. We run an ablation study to determine the effect of different ways of incorporating keywords using 5 main classes of models: i) No-keyword model (no_kw): Trained without any keyword information ii) Keyword-context ($kw\_context$): Trained with keyword as auxillary input + dialog context iii) Keyword-loss (kw_loss): Incorporates keyword loss + keyword as auxillary information. iv) Keyword sim-loss (kw_sim_loss): Incorporate similar words (embedding-based techniques such as Glove (Pennington et al., 2014) ($kw\_sim\_loss\_glove$) and wordnet-based ($kw\_sim\_loss\_wordnet$) similarity) for loss computation . We experiment with 2 variations, one using the similarity score, and the other using 1. v) Multiple-keyword-loss ($multi\_kw\_loss$): Incorporate multiple keywords into the input as well as into the loss computation.

4.1 Datasets

Although there are a few AAC datasets, (Vertanen and Kristensson, 2011), (Vertanen, 2013), they lack multi-turn dialogs, which is central to our task as well as our use-case. Hence we use the Dailydialog dataset (Li et al., 2017), which consists of 13,118 daily conversations involving various topics such as tourism, culture, education, etc., with the goal of exchanging ideas and information and enhancing social bonding. The dataset includes conversations around health, ordinary life and emotions among others, which allows it to serve as a staring point for building systems to support social communication for AAC applications. We use the test set, consisting of 6740 context-response pairs, to evaluate our models.

4.2 Automatic Evaluation

Given the well-discussed fact that word-overlap based metrics do not agree well with human judgment, we utilize learning based and embedding-based metrics to evaluate the generated response with the reference ground truth.

4.2.1 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 keyword predictors using 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 use these keywords to generate responses, and score the responses based on 'human-like' and coherence scores using DialogRPT (Gao et al., 2020), a model trained to predict human feedback dialogue responses.

4.2.2 Metrics for Evaluating Controllable Response Generation Model

Keyword Insertion Accuracy(KIA): The main goal of this work is to provide fine-grained control to the user and have the model induce a keyword or a similar word in the response. To objectively evaluate this, we define keyword-insertion accuracy, where we identify if the input word or a word that is similar, is a part of the generated sentence or not. We compute the accuracy of exact keyword insertion and we also compute accuracies of insertion of words containing similar meaning into the generated response. We use embedding-based cosine similarity metrics and heuristically use a threshold 0.7 to compute the accuracies.

Similarity-Based & Response Quality Metrics Since we intend to generate keyword-based responses, computing measures of similarity between the generated response and ground truth using metrics such as BLEURT, BERTScore (Zhang et al., 2020a) (Sellam et al., 2020), Sentence-BERT (Reimers and Gurevych, 2019) gives a good assessment for the model performance.

We evaluate turn level response quality aspects such as fluency and context coherence using language model based evaluation (GPT-2) and diversity using n-gram based evaluation (Pang et al., 2020) [3]. We also measure the perpelexity (PPL) by employing pretrained GPT-2 "medium".

4.3 Human Evaluation

We perform human evaluation via Amazon Mechanical Turk(AMT) to evaluate the keyword predictor models and controllable response generation models in 3 separate crowd-tasks.

Task1: Collecting response for automatic and human-entered keywords We present a conversation context and keywords (from the extractive and generative keyword prediction models) to the turkers and ask them to come up with possible responses relevant to these keywords. To represent human-control in our analysis, the turkers are also asked to enter keywords of their choice, along with the corresponding responses. We use these in Task 2 to present to the turkers as human responses.

Task2: Overall system interaction and metrics: In the interaction flow, the user reads the conversation context, picks a keyword (From task 1) that he/she wants to respond with - which brings up a human response (from Task1) and a model response (kw_loss model). The user can use a response as is or edit or type a new response altogether. We analyse if the users tend to choose a model or a human response and also compute the word error rates (WER) for the corresponding edits.

Task3: Human Evaluation of controllable 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: relevance of a response to the keyword.

We present pairs of responses from models A and B and provide 4 options for each of the above criteria: A better than B, B better than A, Both and, Neither. We evaluate the pairs, $no\_keyword$ vs $kw\_context$, $no\_keyword$ vs kw_loss and

[3] https://github.com/alexzhou907/dialogue_evaluation

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

Table 1: Evaluation of keyword predictor models.

$kw\_context$ vs kw_loss. We compute the scores using a majority vote across 3 annotators.

5 Results and Discussion

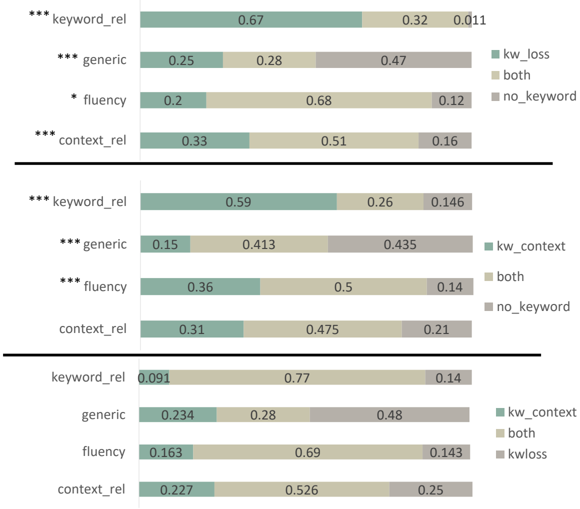


Figure 4: Results from human evaluation. (One-Sample Wilcoxon Signed Rank Test (mu=0) for the statistical tests.*** p<0.001, ** p<0.01, * p<0.05.)

5.1 Automatic Evaluation Results

Keyword Predictor Models: Table 1 shows that the generative keyword predictor tends to generate more diverse keywords (lower score of cosine similarity indicates higher diversity), which is very important in our use-case. The generated responses are also more coherent and human-like.

Cue/Keyword controlled models: We experiment the keyword-loss models with various values of γ ranging between 0 and 1 and see the best performance when γ=0.005. Henceforth, we use keyword-loss models with γ=0.005 for all our experiments.From Table 2, the KIA for the no_kw model is very low, given the one to many nature of open domain dialog. By guiding the model with keywords, the KIA goes up to 67.2% and this is improved to 69.4% in kw_loss model. All of the cue/keyword based models outperform the no_kw model in all of the similarity-based and response quality metrics, except perplexity where the no_kw model is the best. Adding keyword-loss greatly improves the context coherence and fluency as compared to adding keyword as context information alone. The context coherence is the highest when we use similarity-based keyword loss, which encourages generating sentences with words having similar meaning as the input word. The $kw\_simloss\_glove - 1$ and $kw\_simloss\_wordnet - 1$ models also show better performance as compared to the $kw\_context$ model. Table 2 also shows the results on using multiple keywords input. We observe that KIA improves with the kw_loss models, especially the glove-similarity based model.

5.2 Human Evaluation Results

We collect about 1000 responses for the keywords suggested by the two keyword predictors and also collect 1000 additional human keywords and corresponding responses from Task 1.

On analysing the response choice (human vs model generated) of the turkers in Task 2, we find that from 121 interactions, 34.7% of the interactions used model response, and 29.7% used human response. We also observe that 60 interactions result in edits of the response. Out of this, the WER for edits for a human response is 0.45 while WER for edits is lower when a model response is chosen, at 0.39. This further indicates that the model response is closer to what the user wants to convey.

Figure 4 shows the human ratings for response quality metrics for different models. We observe that the kw_loss and $kw\_context$ models outperform the model without control, on all metrics. The keyword-based models generate more fluent and relevant responses. We also observe that humans rate $kw\_context$ and kw_loss models as very comparable, with kw_loss models being more keyword and context relevant as also established by the automatic evaluations.

5.3 User Study

We perform a preliminary study with 7 users[4] by mimicking the disability scenario where the user can only interact with the system using eye-gaze as input. The user interface is controlled using a commercial eye-gaze tracker that works along with an open source mouse-control software, OptiKey[5], an on-screen keyboard designed for users

[4] pandemic, limited hardware availability among other socio-technical issues impedes the pace of the study

[5] https://github.com/OptiKey/OptiKey

	KIA	Similarity	BLEURT	BERT Score	Context	Diversity	Fluency	PPL$\downarrow$
Single Keyword								
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
kw_loss	**0.694**	**0.542**	-0.609	**0.885/0.852/0.868**	0.579	1.726	**0.407**	43.115
kw_sim_loss_glove-1	0.684	0.541	**-0.606**	0.884/0.852/0.868	**0.585**	1.729	0.405	42.544
kw_sim_loss_wordnet-1	0.686	0.540	-0.615	0.884/0.852/0.868	0.581	1.726	0.403	42.606
kw_sim_loss_glove	0.680	0.543	-0.610	0.885/0.852/0.868	0.570	1.741	0.403	42.362
kw_sim_loss_wordnet	0.672	0.541	-0.606	0.884/0.852/0.867	0.576	1.733	0.403	42.301
Multiple Keywords								
no_kw	0.041	0.271	-1.035	0.868/0.836/0.851	0.541	1.592	0.407	**39.098**
kw_context	0.293	0.607	**-0.499**	**0.895/0.857/0.875**	0.489	**1.396**	0.399	75.300
kw_loss	0.300	0.604	-0.524	0.894/0.856/0.874	**0.492**	1.354	0.412	83.971
kw_sim_loss_glove-1	**0.302**	**0.610**	-0.535	**0.895/0.857/0.875**	0.487	1.366	0.416	84.367

Table 2: Performance of the various controllable models for single and multi-keyword inputs (γ = 0.005). Label "-1" indicates that we set $sim(k, kw) = 1$ in equation 3.

with Motor neurone disease(MND). The user interacts with a wizard-based interlocutor in a multi-turn dialog to complete open ended conversation goals. Users can pick two goals/tasks to complete (which the wizard is unaware of) out of sample tasks. After completion of the two tasks, the users are required to answer a survey with likert-scale questions where they rate the overall experience in the task. From the survey, we find that the users "felt that the provided tasks were meaningful and the keywords were very useful in carrying out the communication". The users also reported that they used the generated responses as it was 'very close to what they wanted to say' (one-tail t-test, mu= 0, mean=0.5, $p < 0.05$). Users appreciated that the study made them empathize with users for whom basic communication is a struggle. Some feedback from users: "*Typing a whole sentence character by character can be painful*", "*The keyword suggestion and response generation feature were quite useful as it cuts down significant efforts from user's side*", "*The responses were pretty good. keywords were sometimes not useful and not what I wanted to convey. I hoped that Spiderman would show up as a movie suggestion just when I entered spider(as it was hard to type) and it worked! that was good to see!*"

6 Conclusion

We present a novel usage for open domain conversational models - representing differently abled users and enabling them to communicate. In such a use-case, minimizing the need for user intervention is critical, hence the focus of this work has been to develop controllable response generation models that enable fine-grained human control in the form of keyword inputs from the user. We also introduce keyword-based loss functions that encourages the model to generate the keyword or similar words in the response. To further improve efficiency and time in interaction, we develop keyword predictors and evaluate them. We show with both automatic and human evaluation that our models outperform the baseline model with no control, at the same time maintaining the response quality. We are working with patients to collect feedback and plan to deploy our system as part of an open source tool to impact the quality of life of the patients and help the caregivers. Future research direction also involves improving the keyword predictors, and personalization of these controllable models (both speech and linguistic).

7 Ethics

CueBot aims to support users with neurological disorders in day-to-day communication while also enabling them to control the response generation. The system has been extensively evaluated using automatic metrics as well as human evaluation via AMT, where the AMT workers were fairly compensated (average >$15 per hour). One of the AMT tasks included rating of responses generated from our models and from humans. We tried to mitigate any bias that could arise in the choices made by turkers by constantly shuffling the responses that we presented. We did not collect any additional personal details (other than those collect by AMT by default) or identities from AMT workers' for any of our tasks, hence preserving their privacy. As next steps, we plan to use the feedback from our user study to improve the system, and integrate into ACAT to enable user studies with ALS patients and further gain their feedback to improve the AI modules. In the current system we use google ASR for the interlocuters speech, which raises some privacy concerns. To mitigate this, we plan to use a local ASR system rather than a cloud ASR so that the data is processed locally. To enable this, we need to evaluate the performance of local ASR systems against the cloud-based google ASR. Both the keyword suggestion and response generation modules use pre-trained language models such as GPT2 and DialoGPT finetuned on DailyDialog dataset conversations. Given this, the responses generated could possibly contain improper content or bias due to 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.

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.

David Beukelman, Susan Fager, and Amy Nordness. 2011. Communication support for people with als. *Neurology research international*, 2011:714693.

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

Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2117–2126, New York, NY, USA. Association for Computing Machinery.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019a. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, page 2287–2295, New York, NY, USA. Association for Computing Machinery.

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

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

R. Elakkiya. 2020. Machine learning based sign language recognition: a review and its research frontier. *Journal of Ambient Intelligence and Humanized Computing*.

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

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

Chris Gibbons and Erin Beneteau. 2010. Functional performance using eye control and single switch scanning by people with als. *Perspectives on Augmentative and Alternative Communication*, 19(3):64–69.

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

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in ai for people with disabilities sbg@a research roadmap. *SIGACCESS Access. Comput.*, (125).

Prakhar Gupta, Jeffrey P. Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *CoRR*, abs/2008.09075.

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.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Olga Kovaleva, Anna Rumshisky, and Alexey Romanov. 2018. Similarity-based reconstruction loss for meaning representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4875–4880, Brussels, Belgium. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. 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.

Katharina Linse, Elisa Aust, Markus Joos, and Andreas Hermann. 2018. Communication matters—pitfalls and promise of hightech communication devices in palliative care of severely physically disabled patients with amyotrophic lateral sclerosis. *Frontiers in Neurology*, 9:603.

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 5988–5999, New York, NY, USA. Association for Computing Machinery.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 2422–2433. Association for Computational Linguistics.

J. Mišeikis, P. Caroni, P. Duchamp, A. Gasser, R. Marko, N. Mišeikienė, F. Zwilling, C. de Castelbajac, L. Eicher, M. Früh, and H. Früh. 2020. Lio-a personal robot assistant for human-robot interaction and care applications. *IEEE Robotics and Automation Letters*, 5(4):5339–5346.

Umut Orhan, Deniz Erdogmus, Brian Roark, Shalini Purwar, Kenneth E. Hild II, Barry Oken, Hooman Nezamfar, and Melanie Fried-Oken. 2011. Fusion with language models improves spelling accuracy for erp-based brain computer interface spellers. In *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2011, Boston, MA, USA, August 30 - Sept. 3, 2011*, pages 5774–5777. IEEE.

Kuniaki Ozawa, Masayoshi Naito, Naoki Tanaka, and Shiryu Wada. 2020. A word communication system with caregiver assist for amyotrophic lateral sclerosis patients in completely and almost completely locked-in state.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3619–3629. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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.

Albara Ah Ramli, Rex Liu, Rahul Krishnamoorthy, I. B. Vishal, Xiaoxiao Wang, Ilias Tagkopoulos, and Xin Liu. 2020. Bwcnn: Blink to word, a real-time convolutional neural network approach. *Internet of Things - ICIOT 2020*, page 133–140.

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.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus.

2020. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*.

Melissa Roemmele. 2021. Inspiration through observation: Demonstrating the influence of automatically generated text on creative writing. *arXiv preprint arXiv:2107.04007*.

Melissa Roemmele and Andrew S Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pages 1–2.

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.

Lei Sha. 2020. Gradient-guided unsupervised lexically constrained text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8692–8703, Online. Association for Computational Linguistics.

Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, and et al. 2019. Personalizing asr for dysarthric and accented speech with limited data. *Interspeech 2019*.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2020. Controlling style in generated dialogue. *CoRR*, abs/2009.10855.

TherapyBox. 2021. Predictable: Text-to-speech aac app (accessed sept 2021).

Verbally. 2021. Verbally app (accessed sept 2021).

Keith Vertanen. 2013. A collection of conversational aac-like communications. In *The 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '13, Bellevue, WA, USA, October 21-23, 2013*, pages 31:1–31:2. ACM.

Keith Vertanen and Per Ola Kristensson. 2011. The imagination of crowds: Conversational AAC language modeling using crowdsourcing and large data sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 700–711. ACL.

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.

Jonathan Wolpaw, Richard Bedlack, Domenic Reda, Robert Ringer, Patricia Banks, Theresa Vaughan, Susan Heckman, Lynn Mccane, Charles Carmack, Stefan Winden, Dennis Mcfarland, Eric Sellers, Hairong Shi, Tamara Paine, Donald Higgins, Albert Lo, Huned Patwa, Katherine Hill, Grant Huang, and Robert Ruff. 2018. Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. *Neurology*, 91:10.1212/WNL.0000000000005812.

Heng-Da Xu, Xian-Ling Mao, Zewen Chi, Jing-Jing Zhu, Fanshu Sun, and Heyan Huang. 2020a. Generating informative dialogue responses with keywords-guided networks.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020b. MEGATRON-CNTRL: controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2831–2845. Association for Computational Linguistics.

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.

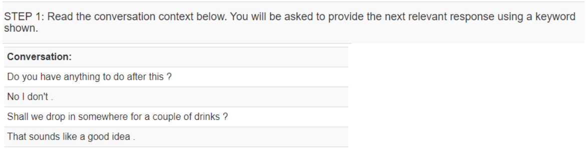
Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing

Liu, and Bill Dolan. 2020c. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Appendix

A Human Evaluation Setup Details

Human evaluation of our system is split into three tasks: task 1 for collecting keywords and corresponding responses from humans. Task 2 involved the crowd workers on Amazon Mechanical Turk interact with our system. We used the keyword suggestions from our extractive and generative keyword predictor models and also the human-generated keywords. We run our controlled response generation pipeline on these keywords to obtain relevant responses. In this task, we first present the turkers with the conversation context as shown in 5. We also present 9 keyword suggestions - 3 from the extractive keyword predictor, 3 from the generative keyword predictor and 3 keywords generated by humans (from task 1). Figure 6 shows this step. Choosing one of these keywords, brings up responses from the human responses generated from Task1, and our controllable response generation model. We use kw_loss model with γ=0.005 and diverse beam search to generate the responses. The users can choose one of the responses and further edit, or enter his/her own response in the box provided.



STEP 1: Read the conversation context below. You will be asked to provide the next relevant response using a keyword shown.

Conversation:

Do you have anything to do after this ?

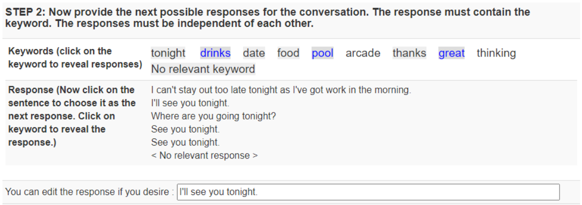
No I don't .

Shall we drop in somewhere for a couple of drinks ?

That sounds like a good idea .

Figure 5: Shows the step 1 for Task 2 on the MTurk study. Here the turkers are presented with the conversation context.

We then present a questionnaire to the turkers - asking them to answer on a likert scale, some questions about why they chose a particular keyword/responses. At the end, turkers are shown a virtual keyboard as you can see in Figure 7 and asked to type in the response that they chose/edited. Using their physical keyboard is disabled for this part of the task - this is to ensure that the turkers use the virtual keyboard and generate the given text. This data enables us to compare the time it took to complete a single interaction and the time it takes to actually type in the entire response (future work).



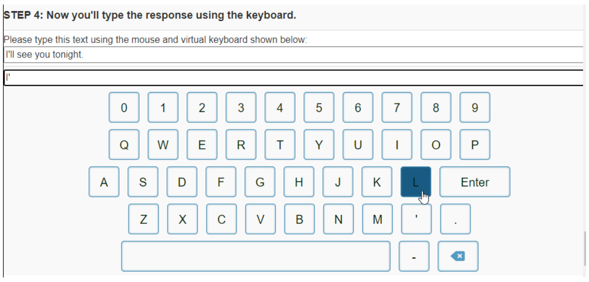
STEP 2: Now provide the next possible responses for the conversation. The response must contain the keyword. The responses must be independent of each other.

Keywords (click on the keyword to reveal responses) tonight drinks date food pool arcade thanks great thinking
No relevant keyword

Response (Now click on the sentence to choose it as the next response. Click on keyword to reveal the response.)
I can't stay out too late tonight as I've got work in the morning.
I'll see you tonight.
Where are you going tonight?
See you tonight.
See you tonight.
< No relevant response >

You can edit the response if you desire : I'll see you tonight.

Figure 6: Shows the step 2 for Task 2 on the MTurk study. Here the turkers are shown 9 keywords (generated from keyword predictor models and humans from task 1). Choosing one of them allows them to see the response generated from our models, and human-generated ground truth response, that can be chosen.



STEP 4: Now you'll type the response using the keyboard.

Please type this text using the mouse and virtual keyboard shown below.

I'll see you tonight.

I'

Figure 7: Shows the step 4 for Task 2 on the MTurk study. Step 3 is questionnaire with radio button options which is not shown above.

B Experiments

We present the effect of varying the γ coefficient in the keyword-based loss models. These results are presented in table 3. Please note that when $\gamma = 0$, the model is the $kw\_context$ model. We see from the table that increasing γ increases the KIA, which matches our intuition, and reaches close to 75% when $\gamma = 1$. However, we see that this is optimal when $\gamma = 0.005$. Similarity metrics such as BLEURT see a drop as we increase γ with the lowest at 1. Also, Response Quality deteriorate heavily with context coherence, diversity and fluency metrics. While the higher γ tries to increasingly encourage the model to generate the keyword in the sentence, this is at the cost of the overall quality of the response. Hence, in all of the experiments and results reported in the paper, we fix $\gamma = 0.005$, unless otherwise specified.

C Sample Model Outputs

In Table 4, we present the outputs from the various models - for a given context and keyword. We show the sample outputs from the no_kw, $kw\_context$, $kwloss\_0.005$, $kwloss\_sim\_loss\_glove$ models and the ground truth. We see that the keywords-based models are able to effectively induce the

	KWI Accuracy	Similarity	BLEURT	Context	Diversity	Fluency	PPL
coeff=0	0.672	0.539	**-0.607**	0.568	**1.789**	0.403	**41.752**
coeff=0.005	0.694	**0.542**	-0.609	0.579	1.726	**0.407**	43.115
coeff=0.01	0.681	0.538	-0.629	**0.581**	1.641	0.406	45.749
coeff=0.1	0.690	0.508	-0.846	0.519	0.888	0.397	92.567
coeff=1	**0.746**	0.527	-0.826	0.468	0.695	0.373	90.070

Table 3: Examining the effect of γ

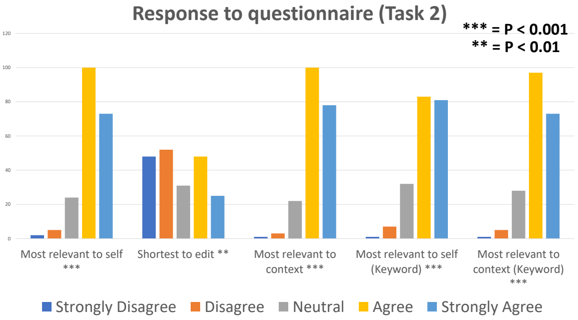


Figure 8: Shows the responses to the questionnaire in Task 2. (One-Sample Wilcoxon Signed Rank Test (mu=0)).

keywords into the generated sentence.

D Keyword Control with Multiple Inputs

Table 5 shows the results from our experiments with training the modesl with multiple keywords as control. We see that $kw\_sim\_loss\_wordnet - 1$ performs well on several metrics. We plan to look into these models further as part of future work.

E Human Evaluation Additional Results

Figure 8 shows some statistics on the responses to the questions asked to the user after the above interaction. The plot shows that most people agree/strongly agree that they picked the keyword/response because it seemed relevant to the context or it resonated with the response in their mind. The plot also shows that people did not choose a response because it was short to edit. This analysis shows that our procedure of suggesting keywords followed by relevant responses is the right strategy for building the controllable response generation system.

F User Study

F.1 User Interface

Figure 9 shows the user interface for our system. The top area shows the placeholder for the interlocutor's voice input which is converted to text for the model using ASR. The interface is divided into two parts, the top area is further split into two panes 1) the left pane brings up the keywords generated from the keyword predictor. Custom keywords can be added using 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.

F.2 Conversational Goals/Tasks

The users were presented with 4 tasks out of which they could choose any two(without communicating the information with the interlocutor), to carry out a conversation as natural as possible. The 4 tasks revolved around the assistive usecase and as below:

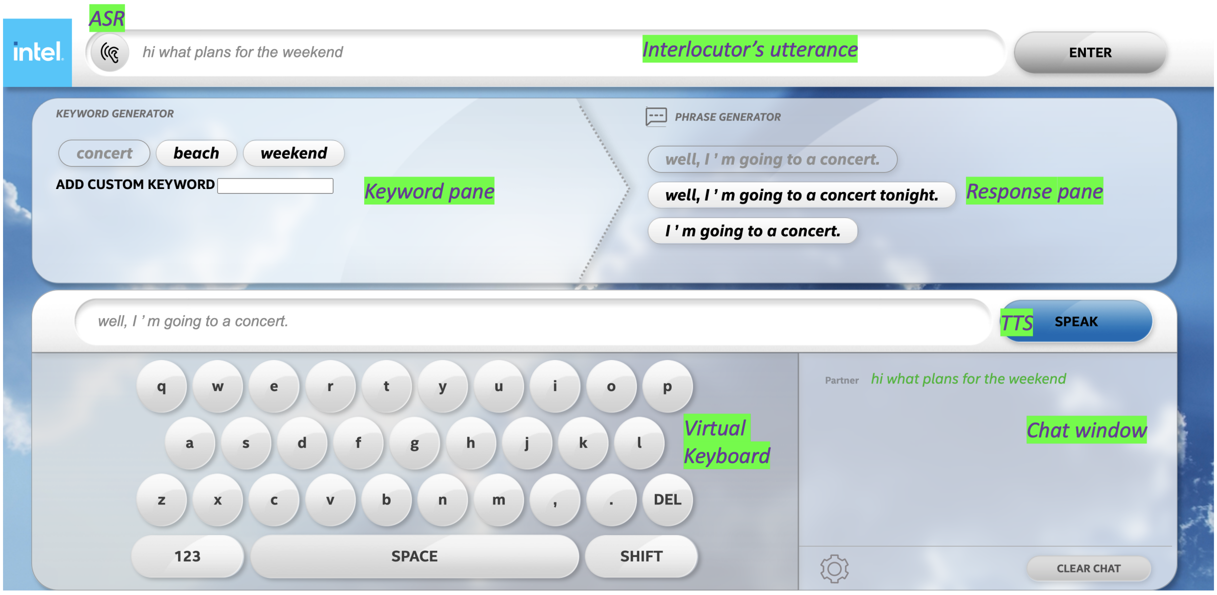
- Invite an interlocutor to a party hosted by the user or to watch a movie (details: when, where and what time).
- Request the interlocutor to get some groceries (details: what and from where)
- Convey to the interlocutor that the user has a bad headache and needs help in setting up a doctors appointment (details: when)
- Convey to the interlocutor that the user is hungry and wants something for lunch. (details: what)

A conversation is considered to be complete or successful when the user is able to convey the goal of the interaction to the interlocutor along with the details associated with the goal. Each session with

Context	**Speaker1**: John , I'd like you to meet Charles Brown , our new manager at the airport . **Speaker2**: Yes , certainly . How can I recognize him ? **Speaker1**: He is short , well-built , and he's got light blond hair . **Speaker2**: Is there anything else ?	**Speaker1**: Please tell me something about your insurance background . **Speaker2**: I majored in international banking at college so I took many insurance courses . And I worked at an insurance company for my graduation field work . **Speaker1**: Why do you want to be an insurance agent ? . **Speaker2**:I care about people and I know the importance of insurance . **Speaker1**: Do you have any work experience in promotion ?
Keyword	***mustache***	***salesman***
NoKeyword	*Yes, he is tall, but he is slim.*	*Yes, I have been promoted twice. I have been promoted twice as well.*
kwcontext	*No, he has a very nice mustache.*	*Yes, I have worked as a salesman for a few years.*
KwLoss_0.005	*He has a rather bushy facial beard and a moustache.*	*Yes. I have a sales job as a salesman in a car dealership.*
kwloss_sim_glove	*No, that's it. He has a rather light mustache, too.*	*Yes, I have been a salesman for two years.*
GroundTruth	*He has a mustache , if I remember correctly , and a light complexion .*	*Yes , I do . I have worked as a salesman for P and G Company Ltd .*

Context	**Speaker1**: Were you in a leadership position when you were a college student ? **Speaker2**: Yes . I was president of the Student Council of our university . **Speaker1**: Did you get any honors or awards at your university ? **Speaker2**: Yes . I was elected Excellent Party Member for four academic years . **Speaker1**: Were you involved in any club activities at your university ? **Speaker2**: Yes . I was a member of the Student Poetry Society . I like writing poems in my spare time	**Speaker1**: Do you have a free moment ? **Speaker2**: Sure . What do you need ? **Speaker1**: I'm having some trouble . How do you use this printer ? . **Speaker2**: Well , first you have to turn this switch on . **Speaker1**: Silly me
Keyword	***extracurricular***	***button***
NoKeyword	*What kinds of poems did you write?*	*It's not complicated.*
kwcontext	*Did you have any extracurricular activities?*	*Then push the button.*
KwLoss_0.005	*Were you involved in any extracurricular activities?*	*Now press this button.*
kwloss_sim_glove	*What was your extracurricular activities?*	*Then press the button.*
GroundTruth	*What extracurricular activities did you usually take part in at your college ?*	*Now press this button .*

Table 4: Sample conversation contexts and comparison of different model outputs



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 9: Cue-bot interface

Multiple Keywords	KIA	Similarity	BLEURT	BERT Score	Context	Diversity	Fluency	PPL↓
no_kw	0.041	0.271	-1.035	0.868/0.836/0.851	0.541	1.592	0.407	**39.098**
kw_context	0.293	0.607	**-0.499**	**0.895/0.857/0.875**	0.489	**1.396**	0.399	75.300
kw_loss	0.300	0.604	-0.524	0.894/0.856/0.874	**0.492**	1.354	0.412	83.971
kw_sim_loss_glove-1	**0.302**	**0.610**	-0.535	**0.895/0.857/0.875**	0.487	1.366	0.416	84.367
kw_sim_loss_wordnet-1	0.287	0.600	-0.525	0.894/0.856/0.874	0.488	1.351	**0.417**	80.403
kw_sim_loss_glove	0.293	0.598	-0.511	0.893/0.855/0.873	0.479	1.344	0.412	80.258
kw_sim_loss_wordnet	0.300	0.607	-0.518	0.894/0.856/0.875	0.483	1.364	0.416	79.888

Table 5: Performance of the various controllable models for multiple keyword input (γ = 0.005). Label "-1" indicates that we set $sim(k, kw) = 1$ in equation 3.

a user lasted between 60 minutes to 90 minutes. The first 30 minutes were spent in explaining the study to the user and helping the user familiarize with the gaze-tracker and the Opti-key mouse functions. Post user-study, a survey was sent to the users to get feedback about the experience with the system.