Knowledge Enhanced Multi Turn Dialogue System: Past, present and future trends

Archana Patil
Department of Computer Science and Engineering
COEP Technological University
Maharashtra, India

Shashikant Ghumbre
Department of Computer Engineering
GCOE&R, Avasari
Maharashtra, India

Vahida Attar
Department of Computer Science and Engineering
COEP Technological University
Maharashtra, India

Abstract

Humans communicate with each other using sign, speech or written way of communication. Recently dialogue systems are gaining a lot of importance as they can be applied in a wide variety of applications. Humans take help of their general knowledge while conversing. Making machines learn how to converse as humans is difficult, also machines lack general knowledge, emotions which are important features of human nature. Various researchers have tried to add external knowledge to machines which can be used for generating responses which will be in coherent with the given input. Aim of this paper is to consolidate different ways of implementing dialogue system, ways of extracting internal knowledge from input and external relevant knowledge from externally given knowledge fact input, various response generation methods for dialogue system. We also consolidate various recent datasets, metrics available for evaluation of dialogue system. Finally, we propose few research directions which will help researchers to pursue their research in this direction.

Keywords: Dialogue system; Knowledge enhanced; Multi turn; Natural language generation; Artificial intelligence

I) Introduction

Given structured data, unstructured data or their combination as input the process that generates natural language as output is called as Natural Language Generation (NLG). Reiter and Dale [1] define NLG as the area within computational linguistics and artificial intelligence that focuses on building computerized systems capable of producing comprehensible texts in human languages from some underlying linguistic and non-linguistic representation of information. NLG can take input in a variety of ways like text, data, image, and video but the output will be in the text. As per Santhanam et al. [115] NLG finds it application in Machine Translation where input is given in one language and it is converted into another language as output, Dialogue System which is built to provide communication between humans and machines, Story generation where the goal is to generate clear and short text for long videos, Poetry Generation whose aim is to produce a text which can be used as poetry. Text Summarization which gives summary of the input where input can span from a one document to many documents. All these tasks are challenging as they all are knowledge-intensive and have to deal with several levels of language, from lexical to semantics.

According to Reiter et al [2], the three-stage pipelined architecture for NLG consists of text planner, sentence planner and linguistic realization. Text Planner incorporates content determination which determines what information is to be conveyed in the response and discourse planning helps in proper structuring of message so that response understanding becomes easy for user. Sentence planner combines lexicalization which selects proper words to be conveyed in response, referring expression generation and aggregation are also used which helps for generating contextually correct response. Linguistic realizer involves syntactic, morphological, and orthographic processing so that system produces correct text output.

With the recent surges of deep learning technologies, understanding and generating natural language using deep learning model have achieved remarkable performance. NLG is used in the dialogue system which is a communication between human and machine and can be categorized as Task-oriented dialogue systems and open-domain dialogue systems. Task-oriented dialogue agents are created for a specific task and are programmed to have brief conversations with the user to gather the information that will aid in the completion of the task. Open Domain Dialogue Systems are systems meant for lengthy talks, rather than focusing on a specific activity like ordering plane tickets, they are put up to emulate the unstructured conversational characteristics that are present in human conversation. Generating responses for the dialogue system using the input text provided sometimes produces bland responses. Hence various researchers have tried to improve the system to generate relevant and coherent responses. Researchers have also explored various verticals for including emotions, personality and knowledge externally to improve systems performance. Knowledge can be categorized as internal hidden knowledge in input or externally given knowledge. Plenty of external knowledge in a variety of formats is available. Hence, diverse methods for finding internal and external knowledge, and incorporating the appropriate information from the internal and external knowledge source into the dialogue system have been proposed for response generation. Adding external knowledge to the dialogue system is one of the ways for improving performance of a system and these systems are called as knowledge enhanced dialogue system. There are various ways of incorporating external knowledge into the system and can be broadly categorized as: knowledge base, knowledge graph or grounded text. Main contribution of this paper is to consolidate: (i) Different methods for building dialogue system (ii) Methods for extracting hidden internal knowledge from input (iii) Methods for extracting external relevant knowledge from external knowledge fact (iv) Methods for response generation (v) Various corpora's available for Dialogue System. (vi) Different metrics for evaluation of Dialogue System.

The organization of the paper is as follows: literature survey in (II) which includes various methods for building dialogue system are given in (II-A), different ways of extracting internal knowledge are given in (II-B), external knowledge extraction methods in (II-C), (II-D) gives different methods for response generation of dialogue system, various corpora's available for dialogue system are listed in (II-E) along with (II-F) listing different metrics used for evaluation of dialogue system. Then we throw a light on various Research directions in (III) for this domain which will help researchers to pursue their research in this field and then finally conclude the paper.

II) Literature Survey

Communication between humans and machines can be for the purpose of some task completion or for entertainment. Conventionally dialogue systems were designed using rule based approaches [8]. Racter [10] is a dialogue system designed which creates prose. Rule based systems sometimes suffer with issues like general and non-committal response. Diversity of generated response by dialogue system has been taken care with the boom of machine learning models, but adding characteristics of human nature like understanding, factuality, informativeness and emotions were open challenge until the evolution of neural network models. With the development of neural network-based model researchers have used seq2seq model and its variants for building dialogue system. Contextual response generation in dialogue system was done using various methods like attention mechanism, hierarchical encoder-decoder, transformer model etc. Seq2seq model does not satisfy criteria for having consistent, semantic and long conversation with user, also it required large amount of training data. Such chatbots might frequently respond with phrases like "I do not know" or "I see." These off-topic replies provide a safe response to various queries, but they are often uninteresting and lack substantial information. Consequently, such responses can swiftly bring the conversation to a close, significantly diminishing the overall user experience. [5] used a conditional Recurrent Neural Network Language Model which uses the contextual information of dialogue for response generation for handling multi-turn conversation. Attention mechanism was proposed in [6] which further improved the performance of the system as it gave importance to different words in input and hence important words were focused to generate response for the given input. Further the development of Dialogue System is supported by advancement in Deep Learning especially transformer based pre-trained Large Language Models.

A. Methods for building a Dialogue system:

a. Convolution Neural Network(CNN):

Researchers have explored CNN for various applications in image processing [11], speech recognition [12] and Natural Language Processing [13], [14]. CNNs offer several advantages in dialogue systems, particularly in terms of efficiency and feature extraction for short text inputs [15]. However, CNN cannot handle data of varing length, long range dependencies and is inefficient to handle sequential data, hence less suitable for sequential data like dialogue system.

b. Recurrent Neural Network (RNN):

RNN is used in image classification, image captioning, speech and signal processing [16]. RNNs sequential structure makes it an excellent choice for modeling text sequences as it provides a way to represent the history conversation, in its recurrent connections, thus model's decision will depend on information which is produced in the past and also it can handle input of varying size. As per [19], RNN can model long term dependencies theoretically, but faces issue like vanishing gradient and exploding gradient problem [17] while doing it practically.

c. Long-Short Term Memory (LSTM):

LSTM [18], variant of RNN was designed using various gating mechanisms to analyze important data from input as well as past conversation in dialogue system needed to ensure the correctness of response generation. LSTM uses both short term memory and long term memory for encoding the input and uses gating mechanism to decide what is important to remember and other can be discarded. Encoder is used to encode input into hidden representation, while decoder uses words generated by itself in the past along with the hidden representation of input for generating the output word by word. Sequence-to-sequence (Seq2seq) architecture is widely used for natural language generation along with its use in dialogue systems because it can handle input-output with different lengths. In [20] author presented the Seq2Seq framework using LSTM based encoder-decoder framework.

d. Hierarchical Recurrent Encoder Decoder (HRED):

In [23] author proposed HRED model which was used for predicting suggestions for next query in web applications in the given session. Author demonstrated that system performance can be improvised by just not considering only recent input but all session history in the given session. Similar idea was mapped to dialogue system domain [24] which showed competing performance with the available state of art models.

e. Gated Recurrent Unit (GRU):

GRU was introduced in [21], GRUs use gating mechanisms to control the flow of information, allowing the model to retain and utilize long-term dependencies more effectively. Various attention mechanisms that can be used with sequence-to-sequence model for improving the performance of models for handling sequence based task are summarized in [22]. In [23] author proposed a hierarchical RNN models using GRUs for dialogue systems, capturing dependencies at multiple levels for more coherent multi-turn dialogues. GRUs plays a significant role in the evolution of dialogue systems, offering a balance between simplicity and effectiveness in handling sequential data [25]. Their integration with advanced techniques like attention mechanisms, hierarchical structures has further enhanced their capabilities.

f. Variational Autoencoder (VAE):

VAE are a powerful tool in the realm of generative models which combines the strength of both probabilistic modeling and deep learning. In [26] author demonstrated that conditioned on the dialogue history, Conditional Variational Autoencoders (C-VAE) can be used for generating contextually appropriate but diverse response. [27] Introduced a hierarchical Latent Variable Encoder-Decoder model that captured long-term dependencies in dialogues and generated coherent multi-turn responses. Variational autoencoder enhances response diversity, contextual understanding and handling of uncertainty. However, their implementation comes with challenges including complex training and potential issues with response quality.

g. Generative Adversarial Networks (GAN) and Deep Reinforcement Learning (DRL):

Response generation in traditional models considered only single input for generating response without considering its impact on future conversation. [28] explored the application of Reinforcement Learning(RL) for long term success of dialogue in Open Domain Dialogue System by improving user engagement, coherence and response quality of generated responses. Flat RL was also used for achieving specific goals like ordering food, flight booking or scheduling appointments [29]. [30] showed that using hierarchical RL can be used for achieving success in learning dialog policies for composite tasks completion like booking air ticket for travel, rent a car and book a hotel. [31] introduced Generative Adversarial Networks for distinguishing real images from fake images. This idea was picked and applied in Natural Language Processing domain to dialogue system by [32] for open domain dialogue generation, where the model jointly trained two systems, a generative model to produce human like response sequences, while the discriminator model was trained to distinguish between the machine generated and human-generated dialogues. Output of the discriminator were used as rewards for generative model, so that the system generated dialogues resembles human generated dialogues and are not dull and generic. In [33] author presented a novel approach to dialogue generation by transitioning from imitation learning to inverse reinforcement learning (IRL) which was used to understand the rewards that drive human conversational behavior, which are not explicitly provided in the dataset, thus improving the quality and diversity of generated dialogues resembling human nature.

h. Transformer based:

Complex recurrent or convolutional neural networks with an encoder-decoder are the foundation of the most popular sequence translation models. The top-performing models additionally use an attention mechanism to link the encoder and decoder. Paper [34] proposed an innovative sequence modeling architecture called transformer that consists of attention modules and feedforward neural networks. Its self-attention module encodes each word in a text sequence by considering the context, thereby generating richer semantic vector representations for each word. Due to its powerful semantic feature extraction capabilities and parallel processing efficiency, the Transformer has gained significant attention and achieved remarkable advancements across various NLP fields. Pre-trained language models like BERT [35], GPT [36], Text-to-text transfer transformer (T5)[78], BART [79] have marked new era for NLP research.

B. Methods for extracting hidden internal knowledge:

Language understanding means to extract hidden knowledge present in given input. It is important because system should well understand what the requirement of user is before generating response for the given input. There are various ways of extracting hidden knowledge in the given input message. People have attempted it using pattern matching, domain identification and intent prediction. ELIZA [8] used high ranked keywords in the input message. Researchers used domain identification, intent prediction using various machine learning algorithms like SVM classifier [40], n-gram classifier, Naive Bayes classifier and Maximum Entropy classifier [41] for utterance classification. [43] Show that incorporating hierarchical structure in intent improves the performance of system. To bridge the gap between seen and unseen intent and make model learn generalized intent, [42] proposed intent expansion framework, the utterances in training data were used for model training, and the model generates embeddings for both seen and unseen intents without model re-training to predict intents i.e. they used zero-shot learning. [44], [45] demonstrated that machines can be trained to generate personalized consistent responses by embedding personality in seq2seq network which was again demonstrated using transformer architecture in [46]. Machines can be trained to generate response depending upon emotional state of machine [47] and politeness was incorporated by [48]. As per [50] abstract meaning representation for dialogue can help better understanding of given input and hence can help improving performance of appropriate response generation. Different approaches like dependency parsing tree of the sentence [39], Transformers for NLU [35], NER using BERT [49][80] are explored. StructBERT was proposed by [53], which incorporates language structures into BERT for better language understanding. [54] proposed TinyBERT with smaller model size, faster inference without hampering the model accuracy. Supervised learning requires large amount of labeled data, to mitigate the issue clustering approach for language understanding [51]. OpenIE tool was explored [55] for extracting ontology from the contextual data for understanding requirement of user.

C. Methods for extracting external relevant knowledge from external knowledge fact:

External knowledge fact comes into variety of format like knowledge base, knowledge graph or unstructured text generally called grounded text. Text documents are unstructured form of data and are large in sizes, while same information can also be represented using knowledge base (KB) or knowledge graph (KG) in structured form. A KB represents the knowledge in the form of triplets <subject, relation, object> while KG represents information in the form of graph KG(Ve, Ed) consisting of all entities e ϵ Ve and edges d ϵ Ed which represents the relationship between entities in dataset. Mumbai is capital of Maharashtra in the form of triplet is represented as <Mumbai, capital_of, Maharashtra> and in graph form it can be represented as 'Mumbai' and 'Maharashtra' as nodes with an edge 'capital_of' between them which represents relationship. Researchers have explored external knowledge sources for response generation in dialogue system [56][57][58][59]. Various ways are experimented by researchers to extract relevant fact from these external knowledge resources using exact matching [60], N-gram matching [66], [71] pattern matching [64], jaccard similarity [63], tf-idf method [64], [71], cosine similarity [62], entity linking [67], SQL and CYPHER [68], SPARQL [61], Memory network [70], transformer [65]. Once the natural language understanding is done and relevant information from external knowledge is extracted, these are presented to decoder for appropriate response generation.

D. Methods for response generation:

Dialogue system uses various ways of generating response for the given input message from rule-based system to generative based system. In rule based system if user input contains specific pattern then response corresponding to the pattern is provided. ELIZA [8] used pattern matching and replacement methodology for response generation. In template based system, system has predefined templates with placeholders that can be filled with relevant information from the user's input or external knowledge fact. In retrieval based system, a predefined set of responses are stored, and the system retrieves the most appropriate response based on the input. This can be done using similarity metrics or neural models. Generative based system can use sequence-to-sequence model or transformer based model for appropriate response generation. Sequence-to-sequence based models require large corpus for training so that they can generate more appropriate responses, but they fail to generate creative responses. Transformer based models like GPT can also be used for generating responses as they are trained on diverse datasets which helps them to generate more contextually relevant, coherent and diverse responses. Hybrid approaches for response generation combine the advantages of both retrieval-based system and generative based system. Retrieval-based systems are for handling specific intents, while neural based models for more open-ended response generation. Reinforcement learning based model [30] [33] are also used to fine-tune response generation based on feedback from user. System learns to optimize response over time through a reward based system. Transfer learning based model [78][79] uses models which are pre-trained on large dataset for a related task and then fine-tune them for specific dialogue dataset which help to generate more contextually relevant and creative responses as the pre-trained model leverages knowledge from diverse sources. People have also explored dialogue system using pre-trained contextual embedding like BERT or ELMo to capture the context of the conversation and generate responses which are contextually relevant. To improve results obtained from models using pre-trained contextually embedding, researchers have also used contrastive learning embedding [84] which works on minimizing distance between similar objects and maximizes distance between dissimilar objects.

E. Various Corpora's available for Dialogue System:

Researchers have made various dialog corpora's available to carry work in this direction. Table I. gives details about various dialogue corpora's.

Dataset	Language	Dataset Statistics	Topic	Speaker	Dataset Feature
MultiWOZ [85]	English	Dialog - 10438, average number utterance per dialog - 14	Restaurant, hotel, attraction, taxi, train, hospital, police	Human to Human	Multiple domain and topics
MuTual [86]	Chinese	dialog - 6371, questions - 11323, context response pair - 8860, utterances per dialog - 4.73	Open Domain	Human written	Conversational reasoning (Chinese student for English listening)
BlendedSkillTalk [87]	English	dialog - 6,808, average utterances per dialog - 11.2	Personal, Knowledge, and Empathy	Human to Human	Conversational skills
MultiDoGO [88]	English	dialogs - 40,576 , utterances per dialog - 20.06	Software support, Media, Insurance, Finance, Fast food, Airline	Human to Human	Multi domain
Schema-Guided Dialogue Dataset[89]	English	dialogs - 16,142, utterances per dialog - 20.44	Weather, Travel, Services, Train, Restaurants, Rental cars, payments, messaging, movies, music, media, hotels, homes, flights, events, calendar, Buses, Alarm, Banks	Machine to Machine	Multi domain
CrossWOZ [90]	Chinese	dialogs - 5,012, utterances per dialog - 16.9	Hotel, restaurant, attraction, metro, and taxi	Human to Human	Multi domain
The Gutenberg Dialogue Dataset[91]	Multilingual	dialog – 2,526,877 utterances – 14,773,741	Fiction	Human written	Multi domain
Fusedchat [93]	english	ODD turns – 60,000, TOD turns – 5,000	Train, Attraction, Hotels, Restaurants, Police, Taxi, Hospital	Human creators	Task oriented dialogue and Open domain dialogue dataset

Dataset	Language	Dataset Statistics	Topic	Speaker	Dataset Feature
ProsocialDialog [95]	English	dialogues -58k, utterances -331K, unique RoTs -160k, dialogue safety labels accompanied by free-form rationales - 497K	Handle problematic condition following social norms	Human-AI collaborative framework	multi-turn dialogue dataset
Knowledge Enriched Task Oriented Dialogue System - KETOD [92]	English	Dialogs- 5324 Utterances per dialog - 9.78	Weather, Travel, Train, Services, Restaurants, Rental cars, Movies, Music, Messaging, Media, Hotels, Homes, Flights, Events, Calendar, Buses	Human to Human	Multi domain and chit chat dialogue dataset
OpenViDial 2.0 [94]	English	Dialogue turns - 5.6M, Visual contexts in images -5.6M	Movies and TV script	Tools and humans	Open domain multi-modal dialogue dataset
SODA [96]	English	dialog -1.5M, utterances - 11M+	Open domain	Pre-trained Language Model	Open domain
KdConv [97]	Chinese	dialog - 4.5K, utterances -86K	film, music, and travel	Human to human	Multidomain
STC [98]	Chinese	Posts - 219,905, responses -4,308,211	Open topics	Social media (Weibo)	One post multiple responses
Ubuntu Dialog [99]	English	Dialogues -930,000, Turns per dialogue - 7.71, words per turn - 10.34	Ubuntu technical issues	Online chat log	Task-specific dialog
PersonalDialog [100]	Chinese	Dialogues - 20.83M, utterances - 56.26M, user profiles -8.47M	Open topics	Social media (Weibo)	Personalization, rich user profiles
Persona-Chat [105]	English	Dialogues - 10,981, Utterances - 164,356	Day to day life	Human to Human	Personalization
CMU DOG [101]	English	Dialogues - 4,112, utterances per dialog - 31.6	30 movies' Wikipedia page	Human to Human	Knowledge-grounded

Dataset	Language	Dataset Statistics	Topic	Speaker	Dataset Feature
Holl-E [102]	English	Dialogues - 9,071, utterances per dialogues -10.0, words per turn -15.3	921 movies	Human to Human	Knowledge-grounded
Wizard of Wikipedia [103]	English	Dialogues -22,311, utterances per dialogues -9.0	1,365 Wikipedia articles	Human to Human	Knowledge-grounded
Topical-Chat [104]	English	Dialogues -11,319, turns per dialog - 22, words per turn -19.8	politics, fashion, sports, general entertainment, science and technology, music, books	Human to Human	Knowledge-grounded
DailyDialog [106]	English	dialogs -13,118, turns per dialog -7.9, words per turn -14.6	Day to day life	Web	Emotion and intent
OpenDialKG [107]	English	Dialogues - 15,673, utterances -91,209	Movie, book, sports, music	Human to Human	Knowledge-grounded
DuConv [108]	Chinese	Dialogues - 29,858, turns per dialog - 9.1, words per turn - 10.6	Films and film stars	Human to Human	Knowledge grounded/Proactivity modeling
DyKgChat [109]	Chinese English	Dialogues- 1,247/3,092, utterances per dialogue-13.8/18.7, words per turn - 27.0/16.5	TV series	TV series	Knowledge-grounded
Empathetic Dialogues [110]	English	dialogs - 24,850, turns per dialog - 4.31, words per turn - 15.2	Daily life	Human to Human	Emotional/empathetic dialog modeling
Target-Guided Conversation [111]	English	dialogs - 8,939, utterances -101,935, keywords - 2,678	Daily life	Human to Human	Proactivity, behavior and strategy
Key-Value Retrieval dataset [113]	English	dialogues - 3031, Average number of turns - 5.25	Calendar, Weather, POI navigation	Human to Human	Multi domain
PERSUASION-FOR-GOOD [112]	English	dialogs - 1,017, turns per dialog- 10.43, words per utterance-19.36	Charity donation	Human to Human	Personalization, behavior and strategy

Dataset	Language	Dataset Statistics	Topic	Speaker	Dataset Feature
JMultiWOZ [114]	Japanese	dialogues - 4246, turns -61186	Tourist attractions, accommodation, restaurants, shopping facilities, taxis, and weather	Human to Human	Multi domain

Table I: Corpora's for Dialogue System

F. Various Metrics for evaluation of Dialogue System:

Human evaluation is best for dialogue system but it is very time consuming and not always feasible. Researchers have been using various metrics proposed for different NLP tasks like language translation, text summarization, dialogue system etc. Table II. shows the consolidation of various metrics for different dialogue system task on various datasets explored by various researchers in their studies. Few of the commonly used metrics are listed below:

a. Manual:

People are hired to evaluate the system manually. The disadvantage is that it is expensive, time-consuming and not reproducible. Humans judge the text produced based on qualities like relevance, fluency, appropriateness, informativeness, politeness, consistency, diversity, engagingness etc. generally on the scale of 5.

b. Automatic:

i. BLEU (Bilingual Evaluation Understudy) [72]:

BLEU is automatic evaluation metrics for machine translation text and is also used by researchers in dialogue system for evaluation. It is geometric average of the modified n-gram precisions, p_i computed using n-grams up to length N (generally N=4) and positive weights w_i equal weight for all n-grams i.e. ¼ and is calculated using (1).

$$\textit{Geometric Average Precision (N)} = \exp\Big(\sum_{i=1}^{N} w_i \log p_i\Big) = \prod_{i=1}^{N} {p_i}^{w_i} = ({p_1}^{1/4}) \,.\, ({p_2}^{1/4}) \,.\, ({p_3}^{1/4}) \,.\, ({p_4}^{1/4}) \tag{1}$$

where w_i = weight for n-gram precision of order i
p_i = n-gram modified precision score of order i
N = maximum n-gram into consideration

Let c be the length of the candidate translation and r be the effective reference corpus length. Brevity penalty BP is calculated using (2)

$$\text{Brevity Penalty} = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \le r \end{cases} \tag{2}$$

where c = *number of words in the generated/predicted sentence* and
r = *number of words in the gold/ target sentence*

$$\textit{BLEU(N)} = \text{Brevity Penalty} * \text{Geometric Average Precision Score (N)} \tag{3}$$

BLEU is calculated using (3). Drawback of BLUE is it fails to capture semantic similarity.

ii. ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation – N) [73]:

ROUGE-N is a ratio of number of overlapping n-grams between generated response and target responses and number of total n-grams in target responses. ROUGE calculated using (4) fails to capture semantic similarity.

$$ROUGE\text{-}N = \frac{\#\text{ of overlapping n} - \text{gram between generated and target response}}{\#\text{ of n} - \text{gram in target response}} \tag{4}$$

iii. METEOR (Metrics for Evaluation of Translation with Explicit Ordering) [74]:

Consider the candidate response 'C' and the target response 'T', METEOR is the harmonic mean of precision and recall, where recall is being weighted nine times more than precision. METEOR calculated using (5), (6) and (7). It also uses chunk penalty. A chunk is the set of all unigrams that are consecutive in target response and the candidate response.

$$METEOR(C, T) = F - mean(C,T) * (1 - chunk\_penalty(C,T) \tag{5}$$

$$F - mean(C, T) = \frac{10(Precision(C,T) * Recall(C,T)}{Recall(C,T) + 9 * (Precision(C,T)} \tag{6}$$

$$chunk\_penalty(C, T) = 0.5 * (\frac{\#\, chunks(C,T)}{\#unigrams\_matched(C,T)}) \tag{7}$$

iv. Perplexity [75]:

Perplexity is one of the measures used to check how well a model predicts the required response. To predict the nth word, model uses earlier (n-1)th generated words, models with smaller perplexity are preferred. Perplexity of a model for the given data can be measured using (8) and (9).

$$\text{PPL} = \; = \sqrt[n]{\frac{1}{p(w_1 w_2 w_3 \ldots w_n)}} \tag{8}$$

$$= \sqrt[n]{\prod_{i=1}^{n} \frac{1}{p(w_i \; |w_1 w_2 w_3 \ldots w_{i-1})}} \tag{9}$$

v. ADEM (Automatic Dialogue Evaluation Model) [76]:

ADEM is a trained model which takes user input, context along with system response and produces a qualitative score between 1 to 5. This metric co-relates well with human judge. It captures semantic similarity beyond word overlap statistics, and also exploits both the context and the reference response to calculate its score for the model response.

III) Research directions

1. **Dialogue System for mental health [77]:**

Day to day life has become very hectic and stressful, there is a rising need of evaluating mental health of people and providing them mental support for making their life happy and easy. Chatbots can serve as a emerging solution for the same as this solution can be cost effective, scalable and made available for broader community.

2. **Proactive Dialogue System [83]:**

Lot of research has been done in dialogue system, but less explored area is proactive dialogue system where the target to be included in final result discussion is already known but chatbots has to drive the conversation slowly and smoothly towards the target.

3. **Dialogue System for toxity reduction and guiding pro-social behavior [81][82]:**

Today lot of pre-trained models are available for building chatbots. These pre-trained models are trained with huge amount of data freely available on the internet. Many times the response generated by these models can have toxic data. There is a need to design a system which can reduce toxic behavior if present in the response generate for dialogue system. Guiding user for pro-social behavior is another vertical which can be explored for human wellbeing.

4. **Group users Dialogue Systems:**

Generally chatbots are designed for one-to-one communication. Many times users communicate in groups. There is need of chatbot which can be used for group discussions or for various tasks like calendar setting for all members of family, meetings at work or even for class of student in some university.

5. **Recommendation Dialogue System ensuring ethics and privacy system [37]:**

Chatbots can be used to extract specific need and preference of user and accordingly recommend marriage councilors, doctor or any other requirement of user which will make a good vertical for future research. Conversation AI system built should be able to set trust among its users so that users share their personal and sensitive information during conversation with system is another area of upcoming research.

6. **Multilingual Chatbots:**

Most of the chatbots are generally designed for single language and majority English. Chatbots handling multiple language and avoiding language biasing is a need of hour so that facility can be made available to broader class of audience.

Conclusion

Thus we have consolidated review on knowledge enhanced dialogue system to facilitate improved text generation. We have reviewed different ways of implementing dialogue system, extracting internal knowledge and external knowledge, response generation methods, various dialogue corpora's available along with different metrics used for evaluation of different tasks of dialogue system and finally conclude the paper with upcoming research challenges which will facilitate various researchers aspiring to work in this domain.

	BLEU	Human Evaluatiom	METEOR	Perplexity	Distinct-N	Accuracy	Phoneme Recognition Rate	Precision	Recall	F-score	Dialogue-Act Match	Cosine distance	Dialogue Length	Diversity	Success rate	Average rewards	Avg. turns per dialogue	Classification error rate	Word error rate	Hits @1	correlations	Joint goal accuracy	Inform and success	ROUGE	Mean reciprocal rank
WMT’14 English to French dataset	[3] [6] [20] [21]			[6]																					
IT Helpdesk Troubleshooting dataset		[4]																							
OpenSubtitles dataset		[4][28]											[28]	[28]											
Twitter	[5]		[5]																						
WMT 2014 English-German	[6]			[6]																					
Chinese dataset from Baidu Tieba		[7]		[7]	[7]																				
ImageNet ,						[11]																			
CIFAR10 , C						[11]																			
IFAR100						[11]																			
TIMIT acoustic-phonetic corpus							[12]																		
Twitter dataset						[13]		[13]	[13]	[13]															
Movie Review dataset						[13]		[13]	[13]	[13]															
STC-SeFun dataset						[14]				[14]															
WMT 2015 English-German	[22]			[22]																					
Switchboard (SW) 1 Release 2 Corpus	[26]										[26]	[26]													
Ubuntu Dialogue Corpus		[27]																							
Twitter dialogue corpus		[27]																							
Frames		[30]													[30]	[30]	[30]								
The MovieTriples dataset	[33] [48]	[33] [48]		[48]	[33]														[48]						

	BLEU	Human Evaluatiom	METEOR	Perplexity	Distinct-N	Accuracy	Phoneme Recognition Rate	Precision	Recall	F-score	Dialogue-Act Match	Cosine distance	Dialogue Length	Diversity	Success rate	Average rewards	Avg. turns per dialogue	Classification error rate	Word error rate	Hits@1	correlations	Joint goal accuracy	Inform and success	ROUGE	Mean reciprocal rank
SMD	[39] [49] [56] [57] [58] [59] [68]	[58]	[49]			[49] [57]				[39] [49] [56] [57] [58] [59] [68]	[68]												[68]		
MultiWoz	[39] [55] [59] [68]									[39] [59] [68]	[68]											[55]	[55] [68]		
ATIS Corpus																		[41]	[41]						
Twitter Persona Dataset	[44]	[44]		[44]																					
Twitter Sordoni Dataset	[44]	[44]		[44]																					
Television Series Transcripts	[44]	[44]		[44]																					
Personal Dialog		[45]		[45]	[45]	[45]																			
PERSONA-CHAT dataset				[46]						[46]										[46]					
Emotional STC conversation dataset		[47]		[47]		[47]																			
Soccer dialogue dataset	[49] [58]	[58]	[49]			[49]				[49] [58]															
GLUE benchmark						[53] [54]				[53] [54]											[54]				
the SNLI Corpus						[53]				[53]															
SQuAD v1.1 QA dataset						[53]				[53]															
Camrest	[55] [59] [68] [70]	[70]								[59] [68] [70]	[68]											[55]	[55] [68]		
The bAbI dialog	[57] [68]					[57]				[57] [68]	[68]												[68]		
DSTC2	[57]					[57]				[57]															
Wizard of Wikipedia	[61]				[61]				[61]															[61]	
REDIAL	[61] [67]			[67]	[61] [67]				[61] [67]															[61] [67]	
Reddit dataset	[62] [66]	[62] [66]		[62]										[66]											

	BLEU	Human Evaluatiom	METEOR	Perplexity	Distinct-N	Accuracy	Phoneme Recognition Rate	Precision	Recall	F-score	Dialogue-Act Match	Cosine distance	Dialogue Length	Diversity	Success rate	Average rewards	Avg. turns per dialogue	Classification error rate	Word error rate	Hits @1	correlations	Joint goal accuracy	Inform and success	ROUGE	Mean reciprocal rank
DSTC9	[65]	[65]	[65]					[65]	[65]	[65]														[65]	[65]
OpenDialKg	[68]									[68]	[68]												[68]		
SODA		[69]																							
KETOD									[71]	[71]												[71]			
Empathetic Dialogues dataset	[82]	[82]		[82]																				[82]	
Prosocial Dialog	[82]	[82]		[82]																				[82]	

Table II: Metrics used for evaluation of Dialogue System

References:

1. Reiter E., & Dale R. 1997. Building natural-language generation systems. Natural Language Engineering, vol 3, pp 57-87
2. Reiter., & Dale R. 2000. Building Natural Language Generation Systems (1st ed.). Cambridge University Press, Cambridge, UK.
3. Sutskever, Ilya, OriolVinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems* 27 (2014).
4. Vinyals, Oriol, and Quoc Le. "A neural conversational model." *arXiv preprint arXiv:1506.05869* (2015).
5. Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, YangfengJi, Margaret Mitchell, Jian-YunNie, JianfengGao, and Bill Dolan. "A neural network approach to context-sensitive generation of conversational responses." *arXiv preprint arXiv:1506.06714* (2015).
6. Vaswani, Ashish, Noam Shazeer, NikiParmar, JakobUszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and IlliaPolosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
7. Xing, Chen, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. "Topic aware neural response generation." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1. 2017.
8. Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." Communications of the ACM 9, no. 1 (1966): 36-45.
9. Colby, Kenneth Mark, Sylvia Weber, and Franklin Dennis Hilf. "Artificial paranoia." Artificial intelligence 2, no. 1 (1971): 1-25.
10. Racter, The Policeman′s Beard Is Half Constructed: Computer Prose and Poetry by RACTER (William Chamberlain and Joan Hall), New York, Warner Books, 1984.
11. Sharma, Neha, Vibhor Jain, and Anju Mishra. "An analysis of convolutional neural networks for image classification." Procedia computer science 132 (2018): 377-384.
12. Palaz, Dimitri, and Ronan Collobert. "Analysis of CNN-based speech recognition system using raw speech as input." (2015).
13. Bertero, Dario, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. "Real-time speech emotion and sentiment recognition for interactive dialogue systems." In Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 1042-1047. 2016.
14. Wang, Yufan, Jiawei Huang, Tingting He, and Xinhui Tu. "Dialogue intent classification with character-CNN-BGRU networks." Multimedia Tools and Applications 79, no. 7 (2020): 4553-4572.

15. Bi, Wei, Jun Gao, Xiaojiang Liu, and Shuming Shi. "Fine-grained sentence functions for short-text conversation." arXiv preprint arXiv:1907.10302 (2019).

16. Giles, C. Lee, Gary M. Kuhn, and Ronald J. Williams. "Dynamic recurrent neural networks: Theory and applications." IEEE Transactions on Neural Networks 5, no. 2 (1994): 153-156.

17. Hochreiter, Sepp, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies." (2001).

18. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.

19. Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult." IEEE transactions on neural networks 5, no. 2 (1994): 157-166.

20. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).

21. Kyunghyun Cho, Bart Van Merrienboer, Caglar Gul- ¨ cehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078

22. Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

23. Sordoni, Alessandro, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion." In proceedings of the 24th ACM international on conference on information and knowledge management, pp. 553-562. 2015.

24. Serban, Iulian, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. "Building end-to-end dialogue systems using generative hierarchical neural network models." In Proceedings of the AAAI conference on artificial intelligence, vol. 30, no. 1. 2016.

25. Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

26. Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

27. Serban, Iulian, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. “A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues”. Proceedings of the AAAI Conference on Artificial Intelligence 31 (1). https://doi.org/10.1609/aaai.v31i1.10983.

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

29. Su, Pei-Hao, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. "Continuously learning neural dialogue management." arXiv preprint arXiv:1606.02689 (2016).

30. Peng, Baolin, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. "Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning." arXiv preprint arXiv:1704.03084 (2017).

31. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

32. Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. "Adversarial learning for neural dialogue generation." arXiv preprint arXiv:1701.06547 (2017).

33. Li, Ziming, Julia Kiseleva, and Maarten De Rijke. "Dialogue generation: From imitation learning to inverse reinforcement learning." In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 6722-6729. 2019.

34. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

35. 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.

36. Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

37. Gumusel, Ece, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. "User Privacy Harms and Risks in Conversational AI: A Proposed Framework." *arXiv preprint arXiv:2402.09716* (2024).

38. Wang, Shuhe, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. "Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts." *arXiv preprint arXiv:2109.12761* (2021).

39. Yang, Shiquan, Rui Zhang, and Sarah Erfani. "Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems." arXiv preprint arXiv:2010.01447 (2020).

40. Patrick Haffner, Gokhan Tur, and Jerry H Wright. 2003. Optimizing svms for complex call classification. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP). IEEE, volume 1, pages I– 632.

41. Ciprian Chelba, Monika Mahajan, and Alex Acero. 2003. Speech utterance classification. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP). IEEE, volume 1, pages I–280.

42. Yun-Nung Chen, Dilek Hakkani-T, Xiaodong He, et al. 2016a. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pages 6045–6049.

43. J. Schuurmans and F. Frasincar, "Intent Classification for Dialogue Utterances," in IEEE Intelligent Systems, vol. 35, no. 1, pp. 82-88, 1 Jan.-Feb. 2020, doi: 10.1109/MIS.2019.2954966.

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

45. Zheng, Yinhe, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. "Personalized dialogue generation with diversified traits." arXiv preprint arXiv:1901.09672 (2019).

46. Wolf, Thomas, Victor Sanh, Julien Chaumond, and Clement Delangue. "Transfertransfo: A transfer learning approach for neural network based conversational agents." arXiv preprint arXiv:1901.08149 (2019).

47. Zhou, Hao, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. "Emotional chatting machine: Emotional conversation generation with internal and external memory." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.

48. Niu, Tong, and Mohit Bansal. "Polite dialogue generation without parallel data." Transactions of the Association for Computational Linguistics 6 (2018): 373-389.

49. Chaudhuri, Debanjan, Md Rashad Al Hasan Rony, and Jens Lehmann. "Grounding dialogue systems via knowledge graph aware decoding with pre-trained transformers." In The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18, pp. 323-339. Springer International Publishing, 2021.

50. Bonial, Claire, Lucia Donatelli, Mitchell Abrams, Stephanie Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. "Dialogue-AMR: abstract meaning representation for dialogue." In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 684-695. 2020.

51. Padmasundari, A., Bangalore, S. (2018) Intent Discovery Through Unsupervised Semantic Text Clustering. Proc. Interspeech 2018, 606-610, doi: 10.21437/Interspeech.2018-2436

52. Xu, Hua, Hanlei Zhang, and Ting-En Lin. "Dialogue System." In Intent Recognition for Human-Machine Interactions, pp. 3-6. Singapore: Springer Nature Singapore, 2023.

53. Wang, Wei, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. "Structbert: Incorporating language structures into pre-training for deep language understanding." arXiv preprint arXiv:1908.04577 (2019).

54. Jiao, Xiaoqi, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. "Tinybert: Distilling bert for natural language understanding." arXiv preprint arXiv:1909.10351 (2019).

55. Zhi Chen, Yuncong Liu, Lu Chen, Su Zhu, Mengyue Wu, and Kai Yu. 2023. OPAL: Ontology-Aware Pretrained Language Model for End-to-End Task-Oriented Dialogue. Transactions of the Association for Computational Linguistics, 11:68–84

56. Eric, Mihail, and Christopher D. Manning. "Key-value retrieval networks for task-oriented dialogue." arXiv preprint arXiv:1705.05414 (2017).

57. Madotto, Andrea, Chien-Sheng Wu, and Pascale Fung. "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems." arXiv preprint arXiv:1804.08217 (2018).

58. Chaudhuri, Debanjan, Md Rashad Al Hasan Rony, Simon Jordan, and Jens Lehmann. "Using a KG-copy network for non-goal oriented dialogues." In The Semantic Web–ISWC 2019: 18th International Semantic Web

Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18, pp. 93-109. Springer International Publishing, 2019.

59. Wan, Fanqi, Weizhou Shen, Ke Yang, Xiaojun Quan, and Wei Bi. "Multi-grained knowledge retrieval for end-to-end task-oriented dialog." arXiv preprint arXiv:2305.10149 (2023).

60. Wang, Dingmin, Ziyao Chen, Wanwei He, Li Zhong, Yunzhe Tao, and Min Yang. "A template-guided hybrid pointer network for knowledge-basedtask-oriented dialogue systems." arXiv preprint arXiv:2106.05830 (2021).

61. Li, Yu, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. "Knowledge-grounded dialogue generation with a unified knowledge representation." arXiv preprint arXiv:2112.07924 (2021).

62. Santhanam, Sashank, Wei Ping, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. "Local knowledge powered conversational agents." arXiv preprint arXiv:2010.10150 (2020).

63. Cai, Deng, Yan Wang, Victoria Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. "Skeleton-to-response: Dialogue generation guided by retrieval memory." arXiv preprint arXiv:1809.05296 (2018).

64. Luo, Cheng, Dayiheng Liu, Chanjuan Li, Li Lu, and Jiancheng Lv. "Prediction, Selection, and Generation: Exploration of Knowledge-Driven Conversation System." arXiv preprint arXiv:2104.11454 (2021).

65. Jin, Di, Seokhwan Kim, and Dilek Hakkani-Tur. "Can i be of further assistance? using unstructured knowledge access to improve task-oriented conversational modeling." arXiv preprint arXiv:2106.09174 (2021).

66. Wu, Zeqiu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski et al. "A controllable model of grounded response generation." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp. 14085-14093. 2021.

67. Wang, Lingzhi, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. "RecInDial: A unified framework for conversational recommendation with pretrained language models." arXiv preprint arXiv:2110.07477 (2021).

68. Madotto, Andrea, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. "Learning knowledge bases with parameters for task-oriented dialogue systems." arXiv preprint arXiv:2009.13656 (2020).

69. Kim, Hyunwoo, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou et al. "Soda: Million-scale dialogue distillation with social commonsense contextualization." arXiv preprint arXiv:2212.10465 (2022).

70. Qin, Libo, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. "Entity-consistent end-to-end task-oriented dialogue system with kb retriever." arXiv preprint arXiv:1909.06762 (2019).

71. Chen, Zhiyu, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. "KETOD: Knowledge-enriched task-oriented dialogue." arXiv preprint arXiv:2205.05589 (2022).

72. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

73. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

74. Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65-72. 2005.

75. Jelinek Fred, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. "Perplexity—a measure of the difficulty of speech recognition tasks." *The Journal of the Acoustical Society of America* 62, no. S1 (1977): S63-S63.

76. 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), pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

77. Striegl, Julian, Sebastian Rottmann, and Claudia Loitsch. "Effectiveness and Acceptance of Conversational Agent-Based Psychotherapy for Depression and Anxiety Treatment: Methodological Literature Review." In *Intelligent Systems Conference*, pp. 188-203. Cham: Springer Nature Switzerland, 2024.

78. Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21, no. 140 (2020): 1-67.

79. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

80. Patil Archana, Shashikant Ghumbre, and Vahida Attar. "Named Entity Recognition over Dialog Dataset Using Pre-trained Transformers." In *International Conference on Data Management, Analytics & Innovation*, pp. 583-591. Singapore: Springer Nature Singapore, 2023.

81. Deng, Yang, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. "Towards Human-centered Proactive Conversational Agents." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 807-818. 2024.

82. Das, Souvik, and Rohini K. Srihari. "Improving Dialog Safety using Socially Aware Contrastive Learning." *arXiv preprint arXiv:2402.00446* (2024).

83. Deng, Yang, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. "A survey on proactive dialogue systems: Problems, methods, and prospects." *arXiv preprint arXiv:2305.02750* (2023).

84. Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." *arXiv preprint arXiv:2104.08821* (2021).

85. Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. “MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics. 2018

86. Cui, Leyang, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. "MuTual: A dataset for multi-turn dialogue reasoning." arXiv preprint arXiv:2004.04494 (2020).

87. Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents’ ability to blend skills. In Proceedings of ACL 2020, Online, July 5-10, 2020, pages 2021–2030.

88. Peskov, Denis, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. "Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data." 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)*, pp. 4526-4536. 2019.

89. Rastogi, Abhinav, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, pp. 8689-8696. 2020.

90. Zhu, Qi, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. "Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset." *Transactions of the Association for Computational Linguistics* 8 (2020): 281-295.

91. Csaky, Richard, and Gábor Recski. "The Gutenberg dialogue dataset." *arXiv preprint arXiv:2004.12752* (2020).

92. Chen, Zhiyu, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. "KETOD: Knowledge-enriched task-oriented dialogue." *arXiv preprint arXiv:2205.05589* (2022).

93. Young, Tom, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. "Fusing task-oriented and open-domain dialogues in conversational agents." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11622-11629. 2022.

94. Wang, Shuhe, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. "Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts." *arXiv preprint arXiv:2109.12761* (2021).

95. Kim, Hyunwoo, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. "Prosocialdialog: A prosocial backbone for conversational agents." *arXiv preprint arXiv:2205.12688* (2022).

96. Kim, Hyunwoo, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou et al. "Soda: Million-scale dialogue distillation with social commonsense contextualization." *arXiv preprint arXiv:2212.10465* (2022).

97. Zhou, Hao, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. "KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation." *arXiv preprint arXiv:2004.04100* (2020).

98. Shang, Lifeng, Zhengdong Lu, and Hang Li. "Neural responding machine for short-text conversation." *arXiv preprint arXiv:1503.02364* (2015).
99. Lowe, Ryan, Nissan Pow, Iulian Serban, and Joelle Pineau. "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems." *arXiv preprint arXiv:1506.08909* (2015).
100. Zheng, Yinhe, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. "Personalized dialogue generation with diversified traits." *arXiv preprint arXiv:1901.09672* (2019).
101. Zhou, Kangyan, Shrimai Prabhumoye, and Alan W. Black. "A dataset for document grounded conversations." *arXiv preprint arXiv:1809.07358* (2018).
102. Moghe, Nikita, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. "Towards exploiting background knowledge for building conversation systems." *arXiv preprint arXiv:1809.08205* (2018).
103. Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. "Wizard of wikipedia: Knowledge-powered conversational agents." *arXiv preprint arXiv:1811.01241* (2018).
104. Gopalakrishnan, Karthik, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. "Topical-chat: Towards knowledge-grounded open-domain conversations." *arXiv preprint arXiv:2308.11995* (2023).
105. Zhang, Saizheng. "Personalizing dialogue agents: I have a dog, do you have pets too." *arXiv preprint arXiv:1801.07243* (2018).
106. 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 (Volume 1: Long Papers), pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
107. Moon, Seungwhan, Pararth Shah, Anuj Kumar, and Rajen Subba. "Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs." In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 845-854. 2019.
108. Wu, Wenquan, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. "Proactive human-machine conversation with explicit conversation goals." *arXiv preprint arXiv:1906.05572* (2019).
109. Tuan, Yi-Lin, Yun-Nung Chen, and Hung-yi Lee. "Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs." *arXiv preprint arXiv:1910.00610* (2019).
110. Rashkin, Hannah. "Towards empathetic open-domain conversation models: A new benchmark and dataset." *arXiv preprint arXiv:1811.00207* (2018).
111. Tang, Jianheng, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. "Target-guided open-domain conversation." *arXiv preprint arXiv:1905.11553* (2019).
112. Wang, Xuewei, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. "Persuasion for good: Towards a personalized persuasive dialogue system for social good." *arXiv preprint arXiv:1906.06725* (2019).
113. Eric, Mihail, and Christopher D. Manning. "Key-value retrieval networks for task-oriented dialogue." *arXiv preprint arXiv:1705.05414* (2017).
114. Ohashi, Atsumoto, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. "JMultiWOZ: A Large-Scale Japanese Multi-Domain Task-Oriented Dialogue Dataset." *arXiv preprint arXiv:2403.17319* (2024).
115. Santhanam, Sashank, and Samira Shaikh. "A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions." *arXiv preprint arXiv:1906.00500* (2019).