Unstructured Text Enhanced Open-Domain Dialogue System: A Systematic Survey

LONGXUAN MA, MINGDA LI, WEI-NAN ZHANG, JIAPENG LI, and TING LIU, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

Incorporating external knowledge into dialogue generation has been proven to benefit the performance of an open-domain Dialogue System (DS), such as generating informative or stylized responses, controlling conversation topics. In this article, we study the open-domain DS that uses unstructured text as external knowledge sources (**U**nstructured **T**ext **E**nhanced **D**ialogue **S**ystem (**UTEDS**)). The existence of unstructured text entails distinctions between UTEDS and traditional data-driven DS and we aim at analyzing these differences. We first give the definition of the UTEDS related concepts, then summarize the recently released datasets and models. We categorize UTEDS into Retrieval and Generative models and introduce them from the perspective of model components. The retrieval models consist of Fusion, Matching, and Ranking modules, while the generative models comprise Dialogue and Knowledge Encoding, Knowledge Selection (KS), and Response Generation modules. We further summarize the evaluation methods utilized in UTEDS and analyze the current models' performance. At last, we discuss the future development trends of UTEDS, hoping to inspire new research in this field.

CCS Concepts: • **General and reference → Surveys and overviews;**

Additional Key Words and Phrases: Unstructured text, knowledge grounded, knowledge selection, open-domain dialogue

ACM Reference format:
Longxuan Ma, Mingda Li, Wei-Nan Zhang, Jiapeng Li, and Ting Liu. 2021. Unstructured Text Enhanced Open-Domain Dialogue System: A Systematic Survey. *ACM Trans. Inf. Syst.* 40, 1, Article 9 (August 2021), 44 pages. https://doi.org/10.1145/3464377

1 INTRODUCTION

Dialogue systems (**DS**) attract great attention because of its wide application prospects. Early DS such as Eliza [198], Parry [26], and Alice [47] attempted to imitate human behaviors in conversations and challenge different forms of the Turing Test [188]. They worked well but only in constrained environments, an open-domain DS remained an elusive task until recently [71].

One of the main challenges in open-domain DS is that the generated responses lack sufficient information [86]. Previous researchers proposed different methods to alleviate this issue.

This article is supported by the National Natural Science Foundation of China (No. 62076081, No. 61772153 and No. 61936010) and Science and Technology Innovation 2030 Major Project of China (No. 2020AAA0108605).
Author's address: L. Ma, M. Li, W.-N. Zhang (corresponding author), J, Li, and T. Liu, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, 92, Xidazhi Road, Nangang Qu, Harbin, Heilongjiang, China, 150001; emails: {lxma, mdli, wnzhang, jpli, tliu}@ir.hit.edu.cn.

1046-8188/2021/08-ART9 $15.00
https://doi.org/10.1145/3464377

Table 1. Classification According to the Properties of Text

Category	Examples
Factual description	Wikipedia articles [33, 141], News reports [57, 68], and Domain specific knowledge [175]
Fictional information	Stories [212], Novel [52], Persona information [227], and Virtual world description [189]. Movie script [186], Movie plots [246], TV dialogues [170], and Emotions [148].
Subjective comments	Reviews of goods [53] or movies [127], Reddit comments [223], and Interview [114]

Diversity enhancement strategies [51, 70, 201, 234] and large scale parameters [13, 145, 233] have been proven to be effective. In addition to the above methods, introducing external knowledge, such as structured **knowledge graph** (**KG**) [1, 136, 207, 208] or unstructured text [33, 53, 136, 218, 227, 246], into the dialogue generation attracts great attention. *External knowledge acts as an extra resource besides the context of dialogues and provides more diversified and grounded information for dialogue generation.*

Structured knowledge usually means KG,[1] which is a semantic map obtained after defining and extracting the named entities and the relations [217, 244] in unstructured text. It was first proposed to better understand the natural language used in search engine.[2] Although it has been widely applicated in **natural language processing** (**NLP**) [37, 69, 130, 194] and achieved great success, some researchers pointed out that complex reasoning [143, 230], unified representation framework [65, 66, 101], interpretability [210, 229], scalability [131, 216], knowledge aggregation [140], and knowledge automatic extraction [75] were still facing challenges. Besides the traditional KG, there are other forms of knowledge, to some extent, that can be seen as structured and used as external knowledge. For example, **Event Graph** [212] or **Event Logic Graph** [34] uses event descriptions (a sentence or a predicate phrase) as vertices and logical relations (causal, sequential, and so on) between events as edges; **Enhanced KG** [95, 245] takes a KG as its backbone and aligns unstructured text to related entities; Chen et al. [16] used **Table Text** [83] as external knowledge for a data-to-text generation task.

Unstructured text can be **Factual description**, **Fictional information**, or **Subjective comments**, as given in Table 1. Factual descriptions are statements of objective facts such as news reports [57], historical events [33], and domain specific knowledge [175]. Fictional information is usually artificially constructed text such as stories [212], movie script [186], or TV dialogues [170]. Subjective comments are usually reviews of things in the world that are recognized by more people on social networks [53, 127]. Generally, comments with a high degree of recognition have high authenticity.

Apart from the categories above, unstructured text can also be divided into **independent sentences** and **documents**. Independent sentences usually do not have logical relationships between themselves [53, 193, 227]. In contrast, a document contains multiple logically related sentences, which together constitute a description of the topic of the document [127, 141, 246]. But in the current stage, most works treat them in a similar way. Hence in this article, we use a unified term "unstructured text" to integrate them. But we do believe that there are more sophisticated semantic structures that can be extracted from documents in future work. Table 2 presents examples of independent sentences and a document in UTED tasks. The dialogue using document as external knowledge is also called **document-grounded conversation** (**DGC**) [246] or **background-based conversation** (**BBC**) [120]. Notably, the Conversational QA is also based on background documents. The differences between DGC/BBC and Conversational QA (such as CoQA [149], QuAC

[1]Following Ji et al. [75], we use the terms KG and knowledge base interchangeably in this article.
[2]https://www.blog.google/products/search/introducing-KG-things-not-strings.

Table 2. The Examples of Independent Sentences and a Document as External Knowledge in UTED Tasks

Sentences as external knowledge (Persona-Chat [227])
"I am an artist"; "*I have four children*"; "I recently got a cat"; "I enjoy walking ..."; "I love **watching Game of Thrones**"
Speaker 1: *My children* and I were just about to **watch Game of Thrones**. Speaker 2: Nice! How old are your children? Speaker 1: *I have four* that range in age from 10 to 21. You? Speaker 2: I do not have children at the moment.
Document as external knowledge (CMUDoG [246])
The Shape of Water is a 2017 American fantasy drama film directed by Guillermo del Toro ... It stars Sally Hawkins ... Set in Baltimore in 1962, the story follows a mute custodian at a high-security government laboratory who falls in love with a captured humanoid amphibian creature. *Rating Rotten Tomatoes: 92 % and average: 8.4/10* ... Critical Response: **one of del Toro's most stunningly successful works**, a powerful vision of a creative master feeling totally
Speaker 1: I thought **The Shape of Water was one of Del Toro's best works**. What about you? Speaker 2: Yes, his style really extended the story. Speaker 1: I agree. He has a way with fantasy elements that really helped this story be truly beautiful. Speaker 2: It has a *very high rating on rotten tomatoes*, too. I don't always expect that with movies in this genre.

The information used in the dialogue is marked with the same color and font in the external text.

[22], and MANTIS [137]) are the dialogue of DGC/BBC is more diversified (including chit-chat or recommendation) and not limited to QA.

Compared with structured knowledge such as the KG, unstructured text has the following advantages. **Firstly**, the unstructured text contains *diversified information*, such as commonsense knowledge [113], stylized information [52, 171], syntactic structures [39], and event logic [212]. **Secondly**, unstructured text is *ubiquitous*, *continuously updated* and *easy to obtain* due to the rapidly increased web-page content [151]. These advantages indicate that unstructured text owns a greater potential to serve as external knowledge in DS than structured knowledge. Hence, incorporating unstructured text into dialogue attracted the attention of academia (e.g., the Alexa Prize challenge [147], **Dialogue System Technology Challenges** (**DSTC**) [80, 223]) and industry [50, 79, 152, 175]. Thanks to the progressing of the semantic representation learning technology such as pre-trained models [32, 36, 84, 104, 139, 144, 145, 219], leveraging knowledge in unstructured text to build an informative and engaging dialogue agent has witnessed great improvement [33, 53, 64, 175]. In this article, we give a detailed investigation of the **Unstructured Text Enhanced Dialogue System (UTEDS)**.

Although UTEDS is considered as a promising research direction, the earlier survey papers in DS research [15, 31, 71, 160, 163, 214] did not cover this topic. Recent literature reviews in **natural language generation** (**NLG**) research noticed the emerging of UTEDS. Guo et al. [61] focused on the conditional-NLG technology and introduced knowledge-enhanced text generation. Santhanam and Shaikh [155] focused on the NLG techniques in DS and introduced several models incorporating world knowledge into dialogue generation. These work only partially reviewed some related UTED models. Most recently, Yu et al. [224] considered a variety of different knowledge-enhanced text generation tasks and answered two questions: how to acquire knowledge and how to incorporate different forms of knowledge to facilitate text generation. However, their investigation about free-form text grounded DS was also incomplete. In contrast, we focus on DS with unstructured text as external knowledge and give a detailed and systematic survey in this domain. We give definitions of the UTEDS related concepts, divide UTEDS into retrieval and generative models and summarize current models with a unified paradigm. We introduce the related datasets, specific architectures, methods, training objects, and evaluation metrics used in UTEDS. Then we summarize and analyze the experimental results of current models. In addition, we leverage the modules defined in this article to point out the future research trends of UTEDS.

Table 3. Comparison of UTED Datasets

Dataset (Language)	Know. Source	See Text	Dialogue Source	Domain	Labeled	Lead by
ARW [193]	Wikipedia	–	Reddit Comments	Philo&Liter	No	Both
GCD [53]	Fours./Twitter	–	Twitter	Comments	No	User
Persona-Chat [227]	AMT	Both*	ParlAI(AMT)	Persona	No	Both
X-Persona(**m.l.**) [96]	AMT	Both*	TranslationAPI	Persona	No	Both
M-Persona [116]	Reddit	Both	Reddit	Open	No	Both
ED# [148]	AMT	User	AMT	Open	No	User
CMUDoG [246]	Wikipedia	User/Both	AMT	Movie	No	Both
Holl-E [127]	Wikipedia	Both	AMT	Movie	Span & Flu.	Both
WoW [33]	Wikipedia	Both	ParlAI(AMT)	Open	Know.	User
CbR [141]	Wikipedia	All	Reddit	Open	No	All
T-Chat [57]	Multi Source	Both*	ParlAI(AMT)	Open	No	Both
LIGHT [189]	AMT	Both	ParlAI(AMT)	Game	No	Both
BST [164]	ParlAI(AMT)	Both*	ParlAI(AMT)	Open	Dialogue Mode	Bot
KOMODIS [49]	IMDB	Both*	AMT	Movie	Entities, and so on	User
Doc2Dial# [44]	Gov. Website	Bot	Doc2Dial [43]	Gov. service	Know., and so on	User
Interview [114]	Public Radio	Bot	Public Radio	Open	Dialogue acts	Bot
Kialo [157]	kialo.com	Both	kialo.com	Open	No	Both
PEC [243]	Reddit	Both	Reddit	Open	No	Both

*means the external knowledge of different interlocutors can be different. # means the dataset could be used as UTED. "–" means the text is crawled independently from the dialogue. "Know." is short for Knowledge. "Philo&Liter" means philosophy and literature. "Flu." means fluency. (**m.l.**) means the dataset is multi-lingual, while others are all English. "Fours." stands for Foursquare tips. "Gov." stands for Government.

As far as we know, we are the first to make a systematic review of the UTEDS. We believe that incorporating unstructured text information into dialogue generation is the future trend of the open-domain DS because a large amount of human knowledge is contained in these raw texts. The research of the UTEDS can assist machines in utilizing human knowledge stored on the internet and understanding natural language.

The structure of the rest of the article is as follows:

—In Section **2**, we review the UTED datasets.
—In Section **3**, we present the system components of Retrieval approaches.
—In Section **4**, we outline the system components of the Generative approaches.
—In Section **5**, we summarize the current evaluation metrics.
—In Section **6**, we analyze the current models' performance in UTED.
—In Section **7**, we put forward promising research directions.

2 DATASETS

Recently, a number of UTED datasets based on different domains have been released [33, 57, 127, 141, 148, 227, 246]. The background unstructured knowledge comes from multi-sources (e.g., Wikipedia, Newsreports) and the dialogues are either collected from crowd-sourcing platform (**Amazon Mechanical Turk** (**AMT**)) or crawled from websites such as Reddit.com.

Table 3 summarizes the characteristics of released UTED datasets from six aspects: the source of the external text, which side of the interlocutors can see the external text, source of conversation, the domain of dialogues, whether the knowledge texts in the dataset have been labeled, and which side leads the conversation. We give a brief introduction of these datasets based on the classifications we defined in Tabel 1.

(1) **Factual description:** Vougiouklis et al. [193] first proposed a dataset **Aligning Reddit comments with Wikipedia sentences** (ARW) and a task to generate context-sensitive and knowledge-driven dialogue responses. To provide the conversation model with relevant

long-form text on the fly as a source of external knowledge, Qin et al. [141] proposed CbR task where the dialogue utterance is accompanied with web-text. It is also used by the Sentence Generation Track of DSTC-7 [223]. Topical-Chat [57] dataset relies on multiple data sources, including Washington Post articles, Reddit fun facts, and Wikipedia articles about pre-selected entities, to enable interactions between two interlocutors with no explicit roles. The external knowledge provided to interlocutors could be the same or not, leading to more diverse conversations. Recently, Hedayatnia et al. [68] presented an amended version of the Topical-Chat dataset with dialogue action plan annotations on multiple attributes (knowledge, topic, and dialogue act). The Doc2Dial [44] uses documents from government websites as grounded knowledge to imitate dialogues in government service. The dataset was collected and processed with an end-to-end framework [43]. Although it was proposed as a task-oriented dataset, it is also suitable for the UTED tasks.

(2) **Fictional information:** The CMUDoG [246] is a dataset built with AMT where dialogue is generated based on the given background text (Wikipedia articles). The CMUDoG places no limits on the dialogues, except the two interlocutors, respectively, play an implicit recommender and recommended role. The Holl-E dataset [127] was proposed to address the lack of external knowledge in the DS, which uses Reddit Comments, IMDB, and Wikipedia introductions as the background. There are two versions of the test set in the Holl-E: one with a single golden reference and the other with multiple golden references. The dialogues in the Holl-E allow one speaker to freely organize the language, while the second speaker needs to copy or modify the existing information (add words before or after a span to ensure smooth dialogue). It is worth noting that the background document has different length settings (mixed, oracle-reduced, oracle, full-reduced, and full) in the dataset, please refer to the original article for more detail.

Zhang et al. [227] crowdsourced persona descriptions as external knowledge to construct a persona consistent dialogue (Persona-Chat). X-Persona dataset is proposed to create multi-lingual conversational benchmarks. In X-Persona, the training sets are automatically translated using translation APIs with several human-in-the-loop passes of mistake correction. In contrast, the validation and test sets are annotated by human experts to facilitate both automatic and human evaluations in multiple languages. In the **Empathetic-Dialogue** (**ED**) dataset, each conversation is grounded in a situation that is written by one participant following a given emotion label. There are 32 emotion labels. The person who wrote the situation description (Speaker) initiates a conversation to talk about it. The other conversation participant (Listener) becomes aware of the underlying situation through what the Speaker says and responds. The WoW dataset [33] is constructed with ParlAI [124] to fill the absence of a supervised learning benchmark task on knowledgeable open dialogue with clear grounding. Each conversation happens between a wizard who has access to knowledge (paragraphs and sentences of articles) about a specific topic and an apprentice who is just eager to learn from the wizard about the topic. **Learning in Interactive Games with Humans and Text** (**LIGHT**) [189] is a large-scale crowdsourced text adventure game, which is used as a research platform for studying grounded dialogue. Within that game world, the authors collected a large set of character-driven human–human interactions involving actions, emotes, and dialogues, with the aim of training models to engage humans in a similar fashion. In the BlendedSkillTalk (BST) [164] dataset, the authors provided responses from poly-encoder [72] models that have been trained toward specific skills (ConvAI2, ED, and WoW) as reference to workers in the conversation.

(3) **Subjective comments:** Ghazvininejad et al. [53] leveraged Foursquare tips from a different person as external knowledge for Twitter conversation, namely **Grounded Conversation Datasets** (**GCD**). Besides plots and articles, datasets such as the DSTC-7 track-2, CMUDoG, and Holl-E also used movie reviews or Reddit comments and external knowledge. The subjective comments often accompany by factual descriptions. For example, in the **Knowledgeable and**

Table 4. Statistics of UTED Datasets

Dataset	Dialogues	Turns/Dialogue	Words/Turn	Number.of.Text	Words/Text
ARW [193]	15,457*	5.0*	**57.7***	75,171*(sentences)	23.7*
GCD [53]	1,063,503	2.0	16.7	43,271,508 (facts)	17.6
Persona-Chat [227]	11,981	15.0	13.7*	1,155 (persona)	27.25*
X-Persona [96]	123,484*	14.7*	15.4*	555,185*(persona)	9.77*
M-Persona [116]	**700 M**	–	–	**5 M** (persona)	–
ED [148]	24,850	4.3	15.2	24,850 (emotion & situation)	19.8
CMU_DoG [246]	4,112	21.4	18.6	120 (documents)	908
Holl-E [127]	9,071	10.0	15.3	921 (documents)	727.8
WoW [33]	22,311	9.1	17.2*	1,356,509 (sentences)	30.7
CbR [141]	2.82M	**86.2**	18.7	32.7 k (documents)	**7,347.4**
T-Chat [57]	11,319	21.9	19.7	3,064* (documents)	–
LIGHT [189]	10,777	13.0	18.3	10,777 (background)	–
BST [164]	6,808	11.2*	–	6,808 (document)	–
KOMODIS [49]	7,519	13.8	14.4	13,818 (facts & opinions)	–
Doc2Dial [44]	4,470	15.6*	14	458 (documents)	947
Interview [114]	105,848	30.2	–	105,848 (situation)	–
Kialo [157]	241,882	2	–	18,255 (persona)	–
PEC [243]	355 K	2.3*	25.5*	250 K (persona)	10.9

*means the statistic is calculated by us. Bold font indicates the maximum value in the column.

Opinionated MOvie DIScussions (**KOMODIS**) dataset [49], every dialogue is constrained by a unique set of facts as well as a suitable set of opinions about the entities in the facts. The participants rated their partner in terms of Naturalness/Attentiveness/Consistency/Personality/Knowledgeability as labels. The CMUDoG [246] and the Holl-E [127] both include subjective comments. Interview [114] dataset is collected from National Public Radio, the dialogues are news interviews based on given situational context. Kialo [157] dataset is collected from www.kialo.com and uses stance-persona as external knowledge. M-persona [116] and PEC [243] construct persona descriptions from Reddit comments.

Table 4 illustrates the statistics of the above-mentioned UTED datasets with total dialogue numbers, average turns per dialogue, average word of each utterance, the total number of unstructured texts, and average words of each unstructured text. "Turn" represents one utterance of a speaker, two turns make a conversation. We can observe that, besides Words/Turn, there are big gaps between datasets in the rest metrics. In order to make better use of these data, the existing UTED models [4, 77, 92, 105, 119, 150, 173, 179, 232, 235, 237] adopted different structures and methods. In the following two sections, we categorize the UTEDS into Retrieval models and Generative models and analyze the current approaches from a component perspective. Although these models conducted experiments on UTED datasets with given texts, they can be easily generalized to online models with a decent **Information Retrieval** (**IR**) module collecting related knowledge [14, 151].

3 RETRIEVAL MODELS

Retrieval models produce a response by selecting from a candidate set. The advantage of the retrieval models is that it can give fluent responses, but the disadvantage is that the candidates are pre-defined, which causes the retrieval models hard to make full use of the dialogue context and external knowledge. Their structures are shown in Figure 1, the retrieval models are composed of three modules: **Fusion**, **Matching**, and **Ranking**.

Retrieval models take the external knowledge $\mathbf{K} = (K_1, K_2, \ldots, K_{|K|})$ with $|K|$ sentences, dialogue context $\mathbf{C} = (C_1, C_2, \ldots, C_{|C|})$ with $|C|$ utterances, and response candidates $\mathbf{R} = (R_1, R_2, \ldots, R_{|R|})$ with $|R|$ candidates as input, aims at selecting a best response R_i, $(i \in \{1, 2, \ldots, |R|\})$ by computing

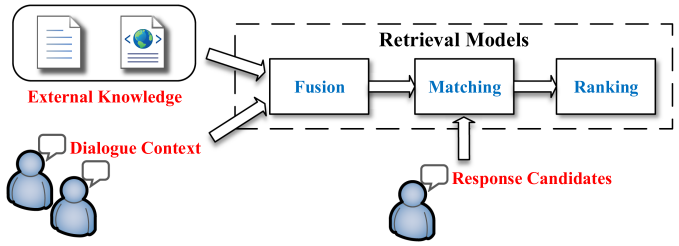

Fig. 1. The general system component of the Retrieval models in UTEDS.

matching scores over the interaction results of **K**, **C**, and **R**. The Retrieval models first encode **K**, **C**, and **R** to get the hidden representations h_K, h_C, and h_R. The encoders[3] for each input can share parameters [58] or not [107, 173, 235]. After obtaining the hidden representations, the **fusion** layer performs interactions among h_K and h_C, then **matching** layer matches each candidate with the fusion results. Lastly, the **Ranking** layer ranks all the candidates with the matching results.

3.1 Fusion

The Fusion layer fuses the information of **K** and **C** for the matching layer. Following Guo et al. [62] who classified deep matching models into interaction-focused and representation-focused, we categorize the fusion methods into **interaction-based** and **representation-based**.[4]

Interaction-based method obtains multiple representation matrices from h_K and h_C. The interaction can be different forms of attention mechanisms [5, 181, 190]. For example, Zhao et al. [235] conducted experiments on Persona-Chat and CMU_DoG datasets. In their **Document-grounded Matching Network** (**DGMN**), document-aware context matrices h_C^K and context-aware document matrices h_K^C are interaction results of the encoding results h_K and h_C. Gu et al. [58] also employed the interaction-based fusion to get the knowledge-filtered context and context-filtered knowledge.

Representation-based method focuses on learning a better semantic representation such as a fusion vector for computing a simple similarity with response candidates. Zhang et al. [227] proposed a Profile-Memory model learning to integrate weighted knowledge into context representation. The augmented context representation was used to match the response candidates. Sun et al. [173] adopted a **History-Adaption Knowledge Incorporation Mechanism** (**HAKIM**) which took the ith turn history representation h_C^i and knowledge representation h_K^i as input and output the updated knowledge representation h_K^{i+1} and knowledge-aware history representation $h_{C_K}^i$. For the $(i+1)$-th turn context, h_C^{i+1} and h_K^{i+1} were taken as input until all the context representations were updated. Then they employed a hierarchical encoder to encode all knowledge-aware context representations into a history vector.

3.2 Matching

The Matching layer pairs the fusion results with each candidate and extracts matching information from the pairs. It can be categorized into **shallow matching** and **deep matching**.

Shallow matching method performs a single interaction after fusion. For example, Lowe et al. [107] proposed a **Knowledge Enhanced** (**KE**) model with the Ubuntu dialogue Corpus as dialogue and the Ubuntu Manpages as external knowledge. They extended the dual-encoder model [108] by separately encoding the **K**, **C**, and $\mathbf{R}_i$ into hidden layer representations h_K, h_C, and h_{R_i}.

[3]For example, HAKIM [173] used BIGRU + self-attention, FIRE [58] used BILSTM to compute the sentence representations.
[4]Notably, the definition of ours is not exactly the same as Guo et al. [62], please refer to their article for more details.

Table 5. Module Comparison between Different Retrieval Models

Models	Fusion	Matching	Ranking	Task	Fusion	Matching	Ranking
TF-IDF baseline [227]	–	Shallow	Point	TMN [33]	–	Shallow	Point
Starspace [227]	–	Shallow	Pair	Transformer [116]	R-b	Shallow	Point
KV-Profile Memory [227]	R-b	Shallow	Point	DGMN [235]	I-b	Shallow	Point
KE [107]	R-b	Shallow	Point	HAKIM [173]	R-b	Shallow	Point
Retrieval-TMN [33]	R-b	Shallow	Point	FIRE [58]	I-b	Deep	Point

"–" means no fusion operation. "R-b" stands for Representation-based. "I-b" means Interaction-based. "Point/Pair" are short for Point/Pair-wise, respectively.

Then they used $\sigma\left(h_k{}^T M_k h_{R_i} + h_c{}^T M_c h_{R_i}\right)$ to compute the matching score of R_i. Where M_k and M_c were learnable parameters. Zhao et al. [235] also employed the shallow matching except h_{R_i} was interacted with three items: context h_C, knowledge-aware context h_C^K, and context-aware knowledge h_K^C. The three interaction matrices were aggregated into a vector with **Multilayer perceptron** (**MLP**) to calculate the final matching score. Yang et al. [218] proposed a **Deep Matching Networks** (**DMN**) for information-seeking conversation. The external knowledge was QA pairs extracted from a Knowledge Base using response candidates. The DMN calculated the matching information between each dialogue turn and the candidate response and used a **Gated Recurrent Unit** (**GRU**) Cho et al. [21] or MLP to aggregate all the sequence matching information into a vector. The vector was finally utilized to compute the score of this candidate.

Deep matching method performs a iteratively referring after fusion. For example, Gu et al. [58] proposed a **Filtering before Iteratively REferring** (**FIRE**) Model. They first employed the interaction-based fusion to get the knowledge-filtered context h_C^K and context-filtered knowledge h_K^C. On one side, they interacted h_R with h_C^K to get first layer refering results $h_R^{K_C,1}$ and $h_{K_C}^{R,1}$. Then $h_R^{K_C,1}$ and $h_{K_C}^{R,1}$ were interacted to get the second layer refering results $h_R^{K_C,2}$ and $h_{K_C}^{R,2}$. On the other side, h_K^C was interacted with h_R in the same way to obtain first layer refering results $h_R^{C_K,1}$ and $h_{C_K}^{R,1}$. $h_R^{K_C,1}$, $h_{K_C}^{R,1}$, $h_R^{C_K,1}$, and $h_{C_K}^{R,1}$ were used to calculate the first layer matching score g^1. In their experiments, the iteration layer was set to 3 and the final score was the average of three matching scores.

3.3 Ranking

There are three commonly used training methods for ranking: Point-wise, Pair-wise, and List-wise. Only the first two are employed by UTED models at present. **Point-wise** calculates the scores of all candidate responses independently [58, 227]. It usually employs Cross-Entropy Loss as an objective function. One disadvantage is that the relative relationship between different candidates is not considered. **Pair-wise** considers the relative relationship between different candidates. Models using Pair-wise ranking normally adopt hinge loss function [218, 227] to distinguish a positive candidate with a negative one and learn to score higher for the ground-truth response. One disadvantage is that the distribution of all candidates is not modeled. In contrast, List-wise directly models the distribution of all response candidates and employs **Kullback–Leibler** (**KL**) Divergence to optimize the gap between the generated distribution and the ground truth one. In Table 5, we summarize the current Retrieval-based models with the Fusion/Matching/Ranking methods.

4 GENERATIVE MODELS

Generative models generate novel sentences that are more natural and variable by conditioning on the dialogue history and external knowledge. Their structures are shown in Figure 2, the generative models consist of **Dialogue and Knowledge Encoding**, **Knowledge Selection** (**KS**), and **Response Generation**.

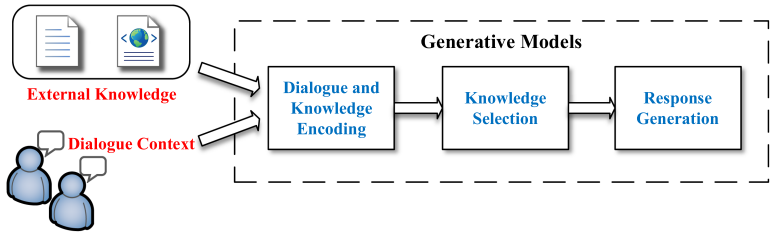

Fig. 2. The general system component of the Generative models in UTEDS.

The UTED generation task can be defined as follows: given unstructured text $\mathbf{K} = (K_1, K_2, \ldots, K_{|K|})$ with $|K|$ sentences and dialogue context $\mathbf{C} = (C_1, C_2, \ldots, C_{|C|})$ with $|C|$ utterances, the task is to generate an r tokens response $\mathbf{R} = (R^1, R^2, \ldots, R^r)$ with probability: $P(\mathbf{R}|\mathbf{C}, \mathbf{K}; \Theta) = \prod_{i=1}^{r} P(R^i|\mathbf{C}, \mathbf{K}, \mathbf{R}^{<i}; \Theta)$, where $\mathbf{R}^{<i} = (R^1, R^2, \ldots, R^{i-1})$ and Θ are the model's parameters. The **Dialogue and Knowledge Encoding** module encodes $\mathbf{K}$, $\mathbf{C}$ to get hidden representations h_K and h_C. After obtaining the hidden representations, the **KS** layer selects useful knowledge from the memory constructed by h_K and h_C, then **Response Generation** layer produces the response word by word with selected knowledge. There is also work trying to combine the advantages of retrieval and generative methods [204], where the authors first retrieved a response from candidates and then used the generation model to transfer the retrieval sentence into a more natural one.

4.1 Dialogue and Knowledge Encoding

Dialogue and knowledge encoding aims at learning good representations from raw text to capture important information for subsequent modules. Researchers utilized different encoding strategies to achieve this purpose. We categorize the current encoding methods into four classes: (1) without pre-training; (2) with pre-trained word embedding; (3) with contextualized word embedding; and (4) with explicit knowledge expansion.

4.1.1 Without Pre-Training. Some models used randomly initialized word embeddings and learned them from scratch during training. For example, on Holl-E task, Zhang et al. [232] used GRU [23] and Meng et al. [119] employed LSTM [174] as encoder; on CMU_DoG, Li et al. [92] used an incremental Transformer [190] encoder.

4.1.2 With Pre-Trained Word Embedding. The traditional word embeddings (Word2Vec [123], GloVe [138], and so on) were obtained by pre-training on large corpus. These embeddings contained semantic information and could help the downstream NLP tasks. In UTED, they were utilized to initialize the dialogue context and the knowledge, and then input into contextualized encoder [150, 227, 235]. Some work chose to train new word embeddings with these methods. For example, to learn the representations of internet slangs and spoken English in target corpus, Ye et al. [222] trained a 100 dimension word embeddings via GLoVe from conversations and facts. However, the traditional word embeddings are fixed and context-independent, they could not resolve the **out-of-vocabulary** (**OOV**) problem and the ambiguity of words in different contexts. So the contextualized word embedding was introduced.

4.1.3 With Contextualized Word Embedding. To obtain contextualized word embedding, some **Pre-Trained Models** (**PTMs**) (ELMo [139], GPT-1 [144], BERT [32], and so on) were introduced. These PTMs were first trained on a large corpus, then fine-tuned on specific tasks. The contextualized embedding has been proven to be better for the downstream NLP tasks [120, 142] than traditional word embedding.

Moghe et al. [128] used ELMo and BERT to obtain pre-trained words embedding with task corpus. Golovanov et al. [54] investigated the different architecture of UTEDS and proposed a Multi-Input (dialogue history, facts, and previously decoded tokens) model where a GPT-1 was duplicated to form an encoder–decoder structure. Golovanov et al. [55] used the Transformer-enhanced GPT-1 [144], and fine-tuned it on Persona-Chat dataset. Pre-trained GPT-1 was also utilized by Tuan et al. [187] as the knowledge base and performed KS in it. Li et al. [89] proposed **zero-resource knowledge-grounded conversation** (**ZRKGC**) model with a pre-trained UNILM [36]. AdapterBot [111] used DialoGPT [233] as a shared encoding backbone for several independent downstream adapters. Each adapter learned to converse with certain skill.

Some researchers proposed their own PTMs with Transformer structure. Dinan et al. [33] introduced **Transformer Memory Network** (**TMN**), a Transformer framework pre-trained with Reddit [116] and SQuAD [146] then fine-tuned on WoW. Fan et al. [41] introduced a Transformer **Sequence-to-Sequence (Seq-to-Seq)** model pre-trained on Reddit comments.

4.1.4 With Explicit Knowledge Expansion. The pre-trained word embedding has been proven to contain semantic knowledge (grammatical [121, 182], syntactical [183], and so on) and world knowledge [140], these knowledge can be considered as implicit. However, to obtain more explicitly semantic features for downstream tasks, researchers have adopted many **explicit knowledge expansion** methods, such as **named-entity recognition** (**NER**) [141], **part-of-speech** (**POS**) [185], syntactic tree [39], and hand-rule features.[5]

Zhao et al. [236] defined label knowledge (conversation topics, structured knowledge facts in KGs, and so on) as additional knowledge that enhanced language representations in different aspects. They used a label knowledge detection module to get the additional label representation, then divided the type of knowledge into two specific categories (labeled and unstructured) during fine-tuning and handled them in different ways. Wu et al. [209] proposed **controllable grounded response generation** (**CGRG**). First, they used rule-based extraction to retrieve the co-occurring keywords/phrases in the dialogue context and the external knowledge, then only used the knowledge sentences containing these keywords/phrases as external knowledge. They used these co-occurrence relationships to simplify the attention operations in the GPT-2 [145], thereby controlling effective information injection and generation. PEE [213] adopted a VAE-based topic model to conduct external persona information mining. Moghe et al. [128] used GCN to learn structured information. They tested three different graph-based structures (dependency, entity co-reference, and entity co-occurrence graphs). Majumder et al. [113] first trained a Commonsense Transformers Framework [11], then generated expansions sentence for each external knowledge facts.

4.2 Knowledge Selection (KS)

As the core component of UTEDS, **KS** module aims at extracting semantically and logically related information from the encoded text representation based on given dialogue history. Most UTED systems followed the *Memory Network Framework* [125, 203] with an attention mechanism [5, 190] to dynamically read appropriate knowledge from the constructed memory. In terms of the existence of a sampling mechanism that explicitly selects the most relevant text fragments from given background knowledge, we divide KS into **implicit (soft) selection** and **explicit (hard) selection**.

4.2.1 Implicit Selection. Early UTED system with **implicit knowledge selection** [53, 127, 129, 227, 246] usually employed the attentional Seq-to-Seq memory network which encoded the context and unstructured text, respectively, into a vector or a sequence of vectors as model memory and

[5]Matching features based on the original word, lower case, lemma, sequence feature, etc.

used the decoder hidden state as a query to attentively read the memory (or concatenate if the encoder output is a single vector) [178]. However, this naive structure depended solely on the attention mechanism to conduct KS, which was too simple to efficiently link the context with related knowledge and extract salient information.

To address this problem, some researchers employed some matching operations between context and knowledge before constructing the memory, we call this **early interaction**. Many of them borrowed idea from the **Machine Reading Comprehension** (**MRC**) task, such as Meng et al. [119] from match-lstm [195], Qin et al. [141], Tian et al. [185] from SAN [103], and Arora et al. [4], Zhang et al. [232] from BiDAF [159]. Instead of predicting a span, they took advantage of the matching techniques in the MRC model, such as cross attention and matching matrix, to generate a document-length memory. They posited that early interaction between context and background knowledge could enable the model to better utilize the context information to pre-select relevant knowledge and integrate the most salient ones together. Following a similar idea, Li et al. [92] proposed Incremental Transformer which employed a Transformer structure [190] with stacked self-attention and cross-attention layers to incrementally encode context and document into a compound memory. Based on ITDD, Ma et al. [110] considered the relationships between current turn and history turns, they designed a **Compare Aggregate Transformer** (**CAT**) to reduce the noise introduced by history turns. Ren et al. [150] further claimed that a single token lacks a global perspective, hence they utilized the matching matrix and implemented an "m-size unfold&sum" operation to select continuous spans from the background and form a topic transition vector which is used to direct KS in the decoding process.

Besides early interaction, different attention mechanisms were also investigated. Zheng and Zhou [241] proposed an **Enhanced Transformer Decoder** (**TED**) that attentively read from context and knowledge representations. Xu et al. [213] proposed a multi-hop memory retrieval mechanism which stacked several attention layers together, the output of the former layer was used as the query for subsequent. Wang et al. [197] investigated three different approaches (Concatenate, Alternate, and Interleave) to combining context and knowledge encodings into a Transformer type decoder. They showed that the Interleave mechanism consistently outperformed concatenate and alternate with extensive experiments. Notably, the attention mechanism used in Zheng et al. [242] employed the Alternate way.

4.2.2 Explicit Selection. Attention mechanisms have been found to operate poorly over long sequences, as the mechanism was blurry and difficult to make fine-grained decisions [42]. The significant model performance decline with longer sequences observed in Zhou et al. [246] also verified this proposition. As a consequence, some scoring and sampling mechanism was proposed to select fragments (usually a sentence) from original background text and subsequently fed the corresponding encodings into the decoder in a similar way as the implicit selection. We define this process as the **explicit selection** mechanism. Another difference between Implicit and Explicit is that models with the latter can measure the accuracy of KS.

Similar to the attention mechanism, scoring aims at matching dialogue context with each pre-segmented text piece, respectively, and generate a preference distribution over them. Dinan et al. [33] and Lian et al. [93], respectively, encoded knowledge sentence and utterance sentences into a sentence embedding vector and applied dot product attention [109] to derive the preference distribution. Other simple scoring methods included but were not limited to MLP attention [5] in Bao et al. [6], **Term Frequency Inverse Document Frequency** (**TF-IDF**) similarity in Gopalakrishnan et al. [57] and K-Nearest-Neighbors in Fan et al. [41]. Ahn et al. [2] classified the above methods as Reduce-Match strategy since they firstly aggregated the utterances into a single vector and then matched it with a knowledge vector. They argued that the coarse information condense

mechanism made fine-grained matching difficult, and they further proposed the Match-Reduce strategy which firstly implemented fine-grained token-level matching operations to obtain matching features (like matching matrices) and then utilized convolution and pooling operations to aggregate these features into a scalar score. Instead of computing one single preference distribution, Zhao et al. [238] regarded KS as a sequence prediction process that used LSTM to sequentially select knowledge sentences.

Besides dialogue context, some researchers attempted to utilize other **supplementary** information to facilitate explicit KS. Kim et al. [77] and Meng et al. [120] posited that tracking the knowledge usage in dialogue history would enable the model to capture the topic flow between multi-turns and reduce knowledge repetition. Kim et al. [77] employed a separate GRU to encode formerly selected knowledge and concatenated the output hidden state with context embedding to conduct scoring. Similarly, Meng et al. [120] explicitly modeled knowledge tracking distribution and used the knowledge sampled from tracking distribution to facilitate the derivation of knowledge shifting distribution, which was exactly the preference distribution over sentences. Zheng et al. [239] argued that the difference between the knowledge sentence selected at different dialogue turns usually provides potential clues to KS. They compared knowledge candidates with the previously selected sentence and used the comparison results as guidance for the final selection. In addition, Majumder et al. [113] used Commonsense Transformers Framework [11] to generate expansions for each persona sentence along the nine relation type that ATOMIC [156] provided. They posited that different utterances had a different preference over the nine relation type, hence they encoded the relation type as additional information to assist scoring. Liu et al. [105] even abandoned the scoring mechanism, instead, they expanded the background text with an augmented graph and aligned each knowledge sentence with a graph vertex. Reinforcement Learning was used to train the reasoning policy over the augmented graph and the final absorbed state (vertex) with its corresponding sentence was selected as the knowledge fragment.

Compared with implicit selection, the training process of the explicit selection model deserves more attention since the sampling operation over categorical distribution is non-differentiable which causes that the back-propagation with simple **Negative Log-Likelihood** (**NLL**) generation loss can not update the parameters of preference distribution. When the training dataset contains the ground truth knowledge label, an auxiliary knowledge loss (i.e., the cross-entropy loss between the predicted preference distribution and the true knowledge distribution) can be applied to train the scoring module [33, 77, 120]. However, if golden knowledge is not provided, other solutions are required. The most naive one is to construct a pseudo-knowledge label with some keyword matching techniques like TF-IDF cosine similarity [2], BM-25 [240], unigram F1 [238], and Jaccard similarity [150]. Moreover, Bao et al. [6] designed two reward functions, informativeness reward and coherence reward, on generated sentence and used policy gradient to train the explicit KS module with encoder and decoder parameters pre-trained and fixed. Notably, Gumbel-Softmax [74] is a common technique for training categorical variables. But Lian et al. [93] claimed that most of the existing KS scoring method which based on semantic similarity was problematic since in a dialogue setting the knowledge might not be semantically related to given utterance but logically. Simple semantic similarity matching would produce a biased prior distribution over the predicted preference distribution which made efficient training difficult without a true knowledge label. To remedy this problem, researchers introduced a Teacher–Student training mechanism that defined a response-anticipated scoring module and generated a posterior preference distribution which was easier trained with Gumbel-Softmax because the golden response was exposed to the model. The posterior preference distribution acted like a soft label that was used to train the original scoring module via a KL divergence loss during training. Similar idea was applied by Majumder et al. [113] and Tian et al. [185] as well.

4.3 Response Generation

The Response Generation module takes the results of previous modules (Knowledge and Dialogue Encoding and KS) and generates the final dialogue response. In this section, we summarize the Architectures, Methods, and Loss Functions used in generative models. The Losses are calculated with the decoding results and are closely related to the Architectures and Methods. Therefore, we introduce them together in this section for the convenience of reading.

4.3.1 Architectures and Methods. The Architectures and Methods utilized in UTED include:

(1) Sequence-to-Sequence (Seq-to-Seq) architecture [21, 174] takes a sequence of words and generates a sequence word by word. Most of the methods used in UTED are based on Seq-to-Seq architecture. RNN-based [119, 150] and Transformer-based [41, 241] Seq-to-Seq architectures were widely used in UTED. Song et al. [165] introduced a Generate, Delete, and Rewrite framework for persona consistent dialogue generation. They adopted Seq-to-Seq Transformer architecture and integrated a matching model to delete the inconsistent words.

(2) Reinforcement Learning (RL) agents make decisions and take actions serially through interaction with the environment, and obtain reward to train action strategy. Li et al. [87] first employed RL for dialogue generation, where a simple Seq-to-Seq model was used as the generation model. Three rewards were defined and combined together to boost diverse response generation. One advantage of RL is the method can leverage non-differentiable rewards. However, the model performance is sensitive to the design of the rewards and is hard to converge.

In UTED, Bao et al. [6] introduced the **Generation-Evaluation** (**GE**) framework for UTED with the objective of letting both participants know more about each other. For the sake of rational knowledge utilization and coherent conversation flow, a dialogue strategy that controlled KS was instantiated and continuously adapted via reinforcement learning. Under the deployed KS strategy, two dialogue robots introduced themselves to each other based on their background information and responded appropriately to the utterances of both parties. The dialogues they generated and the corresponding background information were used for the strategy evaluation module to measure informativeness and coherence. These assessments were integrated into a compound reward, which served as a reinforcement signal to guide the continuous evolution of dialogue strategies. Xu et al. [212] presented a novel **event graph grounded RL framework** (**EGRL**). The EGRL first constructed an event graph where vertices represented events (most simply verb phrases) and edges indicated relations (temporal order, causal relation, and so on) between the events, then used RL based multi-policy method to conduct high-level content planning. Liu et al. [102] proposed a Transmitter-Receiver based framework to explicitly model the understanding between interlocutors, where Transmitter (GPT) was responsible for dialogue generation, and Receiver (BERT) was responsible for personality understanding. The latter was similar to a discriminator, which determined whether the dialogue generated by the former fulfilled the requirements. A weighted sum of Language Style, Discourse Coherence, and Mutual Persona Perception was calculated as a reward for RL.

(3) Conditional Variation AutoEncoder (CVAE) is first used for Conditional Image Generation [215]. Zhao et al. [234] first proposed Dialogue-CVAE that learned a latent variable to capture discourse-level variations and generated diverse responses by drawing samples from the learned distribution.

In UTED, Song et al. [166] designed a memory-augmented CVAE architecture (Persona-CVAE) that used a standard memory network to encode persona to a vector representation. In order to capture the relations among dialogue contexts, facts, and responses, Ye et al. [222] introduced a CVAE model that applied three attention variants to model interactions: context-only attention, parallel attention, and context-guided fact attention.

The CVAE structure is known to suffer the vanishing latent variable problem [234] that the model tends to minimize the KL-divergence between prior and posterior to 0, which cause the model can not learn the posterior information. Some methods such as annealing loss trick [222] or bag-of-words loss [234] were applied to alleviate this problem. Lin et al. [98] pointed out that the autoregressive computation of the RNN limited the training efficiency of CVAE, they proposed Variational Transformers models to address the problem.

(4) Generative Adversarial Networks (GAN) [56] consists of two alternating training modules: Generator and Discriminator. The training goal of the generator is to generate data similar to the real data in the training set. The Discriminator is usually a binary classification model to determine whether the input is a real training sample or a sample generated by the generator. The training goal is to distinguish the real data and the generated one. In NLP, Adversarial training is usually adopted in Reinforcement Learning [88] framework, the output of the discriminator is used as a reward to train the generative model. The GAN is known to be remarkably difficult to train [3], especially in NLP [29]. However, according to de Masson d'Autume et al. [29], GANs do not suffer from exposure bias[6] since the model learns to sample during training.

In UTED, based on adversarial training in dialogue generation [88], Mor et al. [129] proposed an adversarial approach (Adv) for persona-based dialogue. They used a Seq-to-Seq model as a generator and GRU+MLP as a discriminator. But unlike Li et al. [88], the Adv model was strictly adversarial and did not employ pre-training nor auxiliary rewards. Song et al. [167] proposed a **Reinforcement Learning based Consistent Dialogue Generation approach** (**RCDG**) to exploit the advantages of the **natural language inference** (**NLI**) technique to address the inconsistent persona problem. The generator was a PTM (BERT) and the evaluator consisted of two components: an adversarially trained naturalness module and an NLI based consistency module. Pan et al. [134] argued that backward dependency of mutual information was crucial to the variational information maximization lower bound. They proposed **Adversarial Mutual Information** (**AMI**) to identify the joint interaction between source and target. In this framework, forward and backward networks could iteratively upgrade or downgrade the instances generated by each other by comparing real and synthetic data distributions. The model used adversarial training to maximize mutual information and minimize data reconstruction space simultaneously.

(5) Meta-Learning aims at training the initial parameters of the model so that these parameters can be quickly iterated to new tasks. It is naturally suitable for few-shot or cross-domain tasks. The previous methods learned an iterative function or a learning rule [192], while **Model-Agnostic Meta-Learning** (**MAML**) proposed by Finn et al. [46] maximize the sensitivity of the loss function on new tasks. The loss function could be greatly reduced when the parameters were only slightly changed.

In UTED, Madotto et al. [112] extended MAML to **Persona-Agnostic Meta-Learning** (**PAML**). The PAML trained an initial set of parameters that could quickly be adapted to a new persona from a few samples. Song et al. [168] proposed the **Customize Model-Agnostic meta-learning** (**CMAML**) algorithm, which could customize a unique model for each dialogue task. In CMAML, each dialogue model consisted of a shared module, a strobe module, and a private module. The first two modules were shared by all tasks, and the third private module had a unique network structure and parameters to capture the characteristics of the corresponding tasks. They also introduced two steps Customized Model Training method (Private Network Pruning and Joint Meta-learning).

(6) Copy Mechanism [59, 60] uses decoder hidden state to learn whether to copy words or fragments directly from the input or to generate words in the vocabulary. It was widely used in generative models to deal with OOV problem in NLP [158].

[6] A distributional shift between training data used for learning and model data required for generation.

In UTED, copy mechanism also played an important role since the decoder needed to copy information from the dialogue context [237], external knowledge [77, 150, 232], or both [95, 221].

The RefNet [119] model used a decoder switcher to switch between normal generation+copy and a semantic unit copy. The limitation was that the semantic unit copy needs knowledge labeled data such as Holl-E for training. Deepcopy [221] designed a decoder copying from multiple knowledge sentences and the dialogue context. It used a feed-forward network to calculate attention score over the context and each knowledge sentence with the decoders hidden state. They also introduced a Hierarchical Pointer Network where the decoder hidden state was used to attend over token level representations and the overall fact level representations.

4.3.2 Loss Function. Training objects used in generative UTED models include:

(1) Language model loss: Language Models (**LM**) commonly adopt **Cross-Entropy** (**CE**) loss where the final hidden state in the decoder is fed into an output layer with Softmax to obtain next token probabilities. These probabilities are then scored using NLL loss where the gold next tokens are taken as labels. For instance, some approach in UTED employed a single **Maximum Likelihood Estimation** (**MLE**) loss [197]; some others used multiple losses including LM loss [99, 206, 236, 237].

(2) Knowledge selection loss: In explicit KS models, if true labels were given, the knowledge selection loss, which computed the CE loss between gold knowledge distribution and predicted knowledge distribution, was always computed [33, 77]. Training with knowledge selection loss could provide the model with a straightforward signal to direct KS. If true labels were not given, some word overlap techniques [2, 150] were employed to generate pseudo-knowledge labels. Xu et al. [213] designed **Persona Exploration and Exploitation** (**PEE**) framework. The former was based on the VAE structure, the latter used the key-value memory structure and a mutually reinforcing multi-hop memory retrieval mechanism. Apart from the NLL, the PEE model introduced two rule-based persona-oriented loss functions: **Persona-oriented Matching** (**P-Match**) loss and **Persona-oriented Bag-of-Words** (**P-BoWs**) loss which, respectively, supervised persona selection in encoder and decoder.

(3) Priori and posterior loss: Some losses are introduced to minimizing the gap between training and inference when Teacher–Student or CVAE module is employed. Arora et al. [4] proposed a Teacher–Student model called RAM with MLE and **mean square loss** (**MSE**). The MSE was used to diminished the gap between the student model and the teacher model. The Persona-CVAE [166] model was trained with the sum of the following losses: the variational lower bound of the conditional log-likelihood, the CE loss of persona memory module and type distribution, and the BoW loss [234]. Lian et al. [93] introduced **Posterior Knowledge Selection** (**PostKS**) model and used the NLL, BOW, and KL-Divergence Loss. The posterior knowledge distribution was used as a pseudo-label for KS. GLKS [150] used MLE loss, Distant Supervision loss and **Maximum Causal Entropy** (**MCE**) loss. DukeNet [120] proposed a two phases training scheme: warm-up training phase and dual interaction training phase. In the first phase, MLE losses (knowledge tracking and shifting loss) were used along with the generation loss. In the second phase, KL-loss was used to reduce the impact of inaccurate reward estimation and the overall loss was the combination of MLE losses and KL-loss. Decoupling [187] and COMPAC [113] used NLL for generation and KL-Divergence for minimizing the KS distribution between the prior and posterior. To train a stronger prior KS module, Chen et al. [17] used the response in BOW format as the posterior information and an NLL loss over the vocabulary in a Teacher–Student framework. Meanwhile, a "fix" operation was introduced to ensure that the prior knowledge distribution did not affect the posterior selection.

(4) Mutual information loss: Mutual information between dialogue context and response can be used to improve diversity [86]. In UTED, Zemlyanskiy and Sha [225] introduced a DiscoveryScore Metric, which was based on maximizing the mutual information between the current dialogue and the revealed knowledge. Du and Black [38] designed an iterative training process and integration method based on Boosting, combining this method with different training and decoding paradigms as a basic model, including mutual information-based decoding and reward-enhanced maximum likelihood learning. The reward enhancement was to add exponential level feedback similar to the evaluation metric in MLE. ZRKGC [89] used Mutual Information Loss between response and grounding rate of latent knowledge. Besides, marginal log-likelihood with Generalized EM method [9] was used and Knowledge Selection Loss was adopted.

(5) Other losses: When different labels are given, auxiliary loss [114, 157] can be leveraged in a multi-task learning schema. L-PCFG [39] used both syntactic tree structures and lexicalized grammar and was trained by minimizing the NLL of both rules and words. RefNet [119] used generation loss, reference loss, and switcher loss (switching between Generation decoder and Reference Decoder). In a similar way, Tanaka and Lee [178] proposed a convergent and divergent decoding strategy that could generate informative and diverse responses considering not only given inputs (context and facts) but also inputs-related topics. They used NLL, copy loss, and switcher loss (selecting convergent or divergent decoding strategy). Li et al. [90] argued that models trained with **maximum likelihood estimation** (**MLE**): (i) rely too much on copying from the context; (ii) contain repetitions within utterances; (iii) overuse frequent words; and (iv) at a deeper level, contain logical flaws. To address these problems, they used the idea of Unlikelihood [201] technique to construct four additional losses besides MLE in the UTED task.

In Table 6, we summarize the current generation-based models using the categories we introduced in this chapter.

5 EVALUATION METRICS

At present, the evaluation methods of the UTEDS can be divided into two categories: auto evaluation and human evaluation. **Auto Evaluation** can be calculated by computer and is often based on accuracy[33], word overlaps [135], and word embedding similarity [100] between the generated response and the ground-truth response. Most recently, model-based [228] auto evaluation is introduced. **Human Evaluation** needs humans to evaluate the quality of responses generated by models. Auto evaluation and human evaluation are normally used together to reflect the models' performance.

5.1 Auto Evaluation

In this section, we introduce the automatic evaluation metrics currently used in the UTEDS.

5.1.1 Retrieval-Based. The ranking of candidate answers is the core of the retrieval DS. Zhao et al. [235] used R@N which evaluates whether the correct candidate is retrieved in the top N results. Dinan et al. [33] adopted Hit@1 (also called Recall@1) which calculated the accuracy of selecting the right knowledge.

5.1.2 Generative-based. **Probability-based: *Perplexity (PPL)* *[8]*** is usually employed to measure the probability of the occurrence of a sentence in LM. While in the DS, the PPL measures how well the model predicts a response, a lower perplexity score indicates better generation performance. Other than target words, some researchers [114] used Byte Pair Encoder (BPE) [41, 57] to compute perplexity. The BPE perplexity can better address the OOV problem. *Entropy-n [231]* reflects how evenly the empirical n-gram distribution is for a given sentence.

Table 6. Module Comparison between Different Generative Models

Models (Papers)	Knowledge and Dialogue Encoding	Knowledge Selection	Response Generation
KGNCM [53]	w/o pre-training (GRU)	Implicit (EarI.)	Seq-to-Seq
CaKe [232], GLKS [150]	w/o pre-training (GRU)	Implicit (EarI.)	Seq-to-Seq, Copy
Persona-CVAE [166]	w/o pre-training (GRU)	Implicit	Seq-to-Seq, CVAE
KIC [95]	w/o pre-training (LSTM)	Implicit	Seq-to-Seq, Copy
RefNet [119]	w/o pre-training (LSTM)	Implicit (EarI.)	Seq-to-Seq, Copy
L-PCFG [39]	w/o pre-training (LSTM)	Implicit	Seq-to-Seq, Syntactic Tree
Deepcopy [221]	w/o pre-training (LSTM)	Implicit	Seq-to-Seq, Copy
TED [241], GDR [165], Interleave [197]	w/o pre-training (Transformer)	Implicit	Seq-to-Seq
ITDD [92], CAT [110]	w/o pre-training (Transformer)	Implicit (EarI.)	Seq-to-Seq
CG+CVAE [222]	re-trained GloVe (GRU)	Implicit	Seq-to-Seq, CVAE
CDD [178]	GloVe (GRU)	Implicit	Seq-to-Seq, Copy
Adv [129]	GloVe (GRU)	Implicit	Seq-to-Seq, GAN
RAM [185]	GloVe (LSTM)	Implicit (EarI.)	Seq-to-Seq
CMAML [168]	GloVe (LSTM)	Implicit	Seq-to-Seq, Meta-Learning
CBS [177]	GloVe (LSTM)	Implicit	Seq-to-Seq, Copy
PAML [112]	GloVe (Transformer)	Implicit	Seq-to-Seq, Meta-Learning
DRD [237]	PTM (Transformer)	Implicit	Seq-to-Seq, Copy
KIF [41], Unlikelihood [90]	PTM (Transformer)	Implicit	Seq-to-Seq
Multi-Input [54], LIC [55], MKST [236], $\mathcal{P}^2$ Bot [102], Decoupling [187]	PTM (GPT-1)	Implicit	Seq-to-Seq
TransferTransfo [206]	PTM (GPT-1)	Implicit	Seq-to-Seq, Copy
SSS [128]	PTM (ELMo/BERT)	Implicit	Seq-to-Seq, Copy
RCDG [167]	PTM (BERT)	Implicit	Seq-to-Seq, GAN
Adapter-Bot [111]	PTM (DialoGPT)	Implicit	Seq-to-Seq
CGRG [209]	PTM (GPT-2)	Implicit	Seq-to-Seq
ZRKGC [89]	PTM (UNILM)	Implicit	Seq-to-Seq
GE [6]	w/o pre-training (GRU)	Explicit	Seq-to-Seq, RL
Match-Reduce [2]	w/o pre-training (Transformer)	Explicit	Seq-to-Seq
PostKS [93]	GloVe (GRU)	Explicit	Seq-to-Seq
DiffKS [239]	GloVe (GRU)	Explicit (Supp.)	Seq-to-Seq, Copy
AKGCM [105]	GloVe (LSTM)	Explicit (Supp.)	Seq-to-Seq, RL, Copy
RR [204]	GloVe (LSTM)	Explicit	Seq-to-Seq
TMN [33], TF [57]	PTM (Transformer)	Explicit	Seq-to-Seq
KnowledGPT [238]	PTM (GPT-2/BERT)	Explicit	Seq-to-Seq
SKT [77], DukeNet [120], PIPM+KDBTS [17]	PTM (BERT)	Explicit (Supp.)	Seq-to-Seq, Copy
PD-NRG [68]	PTM (GPT-1)	Explicit	Seq-to-Seq

"PTM" means Pre-trained Model. "Supp." means KS with Supplementary information. "EarI." stands for KS with Early Interaction.

Word overlap with ground truth: *BLEU-n [135]* measures the n-gram overlap between the generated response and the golden response. *BLEU* calculates the geometric average of the accuracy of BLEU-n. *ROUGE-n [94]* is based on the calculation of the recall rate of the common subsequence of generating response and the real one. *METEOR [82]* further considers the alignment between the generated and the real responses to improve BLEU. WordNet is adopted to calculate the matching relationship among specific sequence matching, synonym, root, interpretation, and so on. *NIST [35]* is an improvement of BLEU by summing up each weighted co-occurrence n-gram segments, then dividing it by the total number of n-gram segments. *Distinct-n [86]* measures the diversity of reply by calculating the proportion of distinct n-grams in the total number of n-grams to evaluate the diversity of generated responses. *Word-level F1 [146]* treats the prediction and ground truth as bags of tokens and measures the average overlap between the prediction and ground truth answer. ***Exact Match*** **(EM)** [211] requires the answers to have an exact string match with human-annotated answer spans. Wang et al. [197] proposed *context use percentage* to measure

how well the model utilizes the relevant information from the context. Li et al. [90] measured the fraction of generated n-grams that appear in the original context as *context repetition*.

Embedding-based: *Greedy Matching [153]* is an embedding-based metric that greedily matches each word in the generated sequence to a reference word based on the cosine similarity of their embeddings. The final score is then an average over all the words in the generated sequence. *Embedding Average [205]* computes a sentence embedding for both the generated sequence and the ground-truth response by taking an average of word embeddings. The score is then a cosine similarity of the average embedding for both the generated and reference sequence. *Vector Extrema [48]* follows a similar setup to Embedding Average, where the score is the cosine similarity between sentence embeddings. Rather than taking an average over word embeddings, this method identifies the maximum value for each dimension of the word embedding. Taking the maximum is motivated by the idea that common words will be de-emphasized as they will be closer to the origin. Vector Extrema showed some advantages on dialogue tasks [100]. *Skip-Thought* [81] uses a recurrent neural network to produce a sentence-level embedding for the generated and reference sequences. Cosine similarity is then computed between the two embeddings. Yavuz et al. [221] employed *CIDEr* [191] to measure the cosine similarity of generated sentence and reference based on TF-IDF vector. To train the diversity of responses, Du and Black [39] used a *K-means clustering* on the sentence embeddings of responses and calculated average Squared Euclidean Distance between members of each cluster.

Model-based: *BERTScore [228]* uses a pre-trained BERT model to greedily match each word in a reference response with one word in the generated sequence. By doing so, it computes the recall of the generated sequence. BERTScore was shown to have a strong system-level and segment-level correlation with human judgment on several machine translation and captioning tasks. In order to approximate manually labeled ratings, Mehri and Eskénazi [118] proposed ***UnSupervised and Reference-free (USR)*** evaluation metric. From the perspective of turn level, The USR used five sub-evaluation metrics (Understandable/Natural/Maintaining Context/Interesting/Using Knowledge) and a general evaluation metric (Overall Quality). A regression model was used on top of a **Masked Language Modelling** (**MLM**) and trained to reproduce the overall score from each of the specific quality scores rated by annotators. The *MLM* was a fine-tuned RoBERTa [104] model. The likelihood of a response estimated by the fine-tuned RoBERTa model is used as an automatic metric for evaluating the understandability and naturalness of responses.

Knowledge based: Qin et al. [141] calculated F1 with non-stop word tokens in the response that are present in the document but not present in the context. *Knowledge R/P/F1* [33] measures the F1 overlap of the models output with the external knowledge. Song et al. [166] proposed *Persona Coverage* to evaluate how well persona information was leveraged. Ma et al. [110] proposed ***Knowledge-Utilization (KU)*** to measure how many N-grams in external knowledge are used in responses. They also introduced ***Quality of Knowledge Utilization (QKU)*** to measure the quality of the N-grams. ***Point-wise mutual information (PMI)*** *[24]* aims at evaluating how smoothly the generated responses are related to the context and knowledge. Madotto et al. [112] and Song et al. [165] trained NLI models to measure the persona consistency between persona sentences and generated responses, their metrics were named *C score* and *Ent*, respectively.

Other: *Average length* of the generated response is often used [141], a response with more words is usually considered to contain more information.

5.2 Human Evaluation

Similar to the Ranking method of the retrieval models, we can classify human evaluations into 2 categories: **Point-wise Grading** and **Pair-wise Comparison**. In Point-wise Grading, workers need to rate each model independently. In Pair-wise Comparison, workers are required to compare

Table 7. Models with Different Manual Evaluation Methods

Models	DataSet	Categories
TMN [33]	WoW	Point, Dialog
KIC [95]	WoW	Point, Turn
KIF [41]	WoW	Pair, Dialog [91]
DiffKS [239]	WoW	Pair, Turn / Point, Dialog
PostKS [93]	WoW, Persona-Chat	Point, Dialog
MKST [236]	WoW, Persona-Chat	Point, Turn
AKGCM [105]	WoW, Holl-E	Pair, Turn
SKT [77]	WoW, Holl-E	Point, Dialog [33]
DukeNet [120], PIPM+KDBTS [17]	WoW, Holl-E	Point, Turn
Unlikelihood [90]	WoW, ConvAI2	Pair, Turn
Match-Reduce [2], KnowledGPT [238]	WoW, CMU_DoG	Point, Turn
ZRKGC [89]	WoW, CMU_DoG, and Topical-Chat	Point, Turn
ITDD [92], CAT [110]	CMU_DoG	Point, Turn
RefNet [119], GLKS [150]	Holl-E	Point, Turn
TF [57]	Topical-Chat	Point, Turn
PD-NRG [68]	Topical-Chat	Pair, Turn
CMR [141]	CbR	Pair, Turn
RAM [185], CGRG [209]	CbR	Point, Turn
LIC [55], PAML [112], CMAML [168], GDR [165], Per-CVAE [166], PEE [213], RCDG [167]	Persona-Chat	Point, Turn
DiscoveryScore [225], $\mathcal{P}^2$ Bot [102], RAML [132]+Boosting [38]	Persona-Chat	Point, Dialog
TF+AMI [134], COMPAC [113]	Persona-Chat	Pair, Turn
GE [6]	Persona-Chat	Pair, Dialog
RR [204]	Persona-Chat	Point and Pair, Dialog
TransferTransfo [206], Deepcopy [221]	ConvAI2	Point, Turn
CG+CVAE [222], CBS [177], CDD [178], and Moel [97]	DSTC-7	Point, Turn
	ED	Point, Turn / Pair, Dialog

"Point/Pair/Turn/Dialog" are short for "Point-wise Grading/Pair-wise Comparison/Turn-level/Dialogue-level", respectively.

different models and choose the preferred one. Meanwhile, according to different scoring objects, we can further divide human evaluations into two other categories: **Turn-level** and **Dialogue-level**. Turn-level evaluation scores the quality of the response and is an offline procedure. Dialogue-level evaluation assesses the quality of multi-turn responses or the entire conversation through interaction with the model. The same person usually plays the role of both the user (one who interacts with the system) and the evaluator [45]. For dialogue quality, Appropriateness [244], Coherence [89], Engagingness [120], Fluency [95], Interest [57], and Naturalness [119] are often used in human evaluations. For the utilization of external knowledge, Informativeness [95], and Relevance [236] are usually adopted. In Table 7, We summarize human evaluation metrics in current UTED models.

There are some other interesting human evaluation metrics recently introduced. Mehri and Eskénazi [117] introduced 18 metrics from the perspective of each turn or whole dialogue. KIF [41] employed Acute-Eval dialogue evaluation system [91] which comparing two full dialogues.

6 ANALYSIS

In Tables 8, 9, 10, and 11, we present some auto evaluation results of the current UTED models from original papers. We did not compare human evaluations in different articles because they lacked a unified standard. It should be pointed out that these experimental results can only be used as a reference due to the differences in data processing scheme [77] and experimental settings. For

Table 8. Experimental Results of Retrieval Models

Models	Dataset	R@1/2/5 %	Models	Dataset	R@1/2/5 %	F1 %
TF-IDF [107]	ARW	41.0/54.5/70.8	IR baseline [33]	WoW	17.8/– –/– –	12.7
KE [107]	ARW	41.3/55.4/82.4	BoW MN [33]	WoW	71.3/– –/– –	15.6
TF-IDF [227]	Persona-Chat	41.0/– –/– –	TMN [33]	WoW	87.4/– –/– –	15.4
Starspace [227]	Persona-Chat	49.1/– –/– –	DGMN [235]	CMUDoG	65.6/78.3/91.2	
KV-Profile Mem [227]	Persona-Chat	51.1/– –/– –	FIRE [58]	CMUDoG	81.8/90.8/97.4	
HAKIM [173]	Persona-Chat	57.6/72.9/89.9	HAKIM [173]	CMUDoG	82.7/93.8/99.5	
DGMN [235]	Persona-Chat	67.6/80.2/92.9	LSTM [116]*	M-Persona	66.3/79.5/90.6	
FIRE [58]	Persona-Chat	81.6/91.2/97.8	Transformer [116]*	M-Persona	74.4/85.6/94.2	

For WoW, we use the Test Seen and Predicted Knowledge version. For Persona-Chat, we use the original persona version. F1 is the unigram overlap of the model's prediction with the golden response. *means the results are R@1/3/10.

Table 9. The Generative Models' Performance on Holl-E/CMUDoG/Topical-Chat/CbR Datasets

Model	Dataset	F1%	BLEU(BLEU-1/2/3/4)%	ROUGE(1/2/L)%	Distinct(1/2)	PPL	R1%
GTTP [127]	Holl-E		– –(– –/– –/– –/ 8.73)	23.20/ 9.91/17.35			
SSS(BERT) [128]	Holl-E		– –(– –/– –/– –/22.78)	40.09/27.83/35.20			
CaKe [232]	Holl-E		26.02(– –/– –/– –/– –)	42.82/30.37/37.48			
RefNet [119]	Holl-E	40.18	27.00(– –/– –/– –/– –)	42.87/30.73/37.11			
GLKS [150]	Holl-E			43.75/31.54/38.69			
SKT [77]	Holl-E*	29.8				48.9	29.2
PIPM+KDBTS [17]	Holl-E*	30.8				39.2	30.6
DukeNet [120]	Holl-E*		– –(– –/– –/– –/19.15)	36.53/23.02/31.46			30.3
DiffKS(Dis) [239]	Holl-E*		– –(– –/29.9 /– –/25.9)	– –/26.4 /– –			33.5
AKGCM [105]	Holl-E*		– –(– –/– –/– –/30.84)	– –/29.29/34.72			42.04
ITDD [92]	DoG		– –(– –/– –/– –/ 0.95)			15.11	
CAT [110]	DoG*		– –(– –/– –/– –/ 1.22)	– –/– –/11.22		15.2	
DialogT [179]	DoG		– –(– –/– –/– –/ 1.28)			50.3	
DRD [237]	DoG	10.7	– –(15.0/ 5.70/ 2.50/ 1.20)			54.4	
ZRKGC [89]	DoG	12.2	– –(16.1/ 5.2 / 2.1 / 0.9)			53.8	
KnowledGPT [238]	DoG	13.5				20.6	
Reduce-Match [2]	DoG*	13.6	0.7(– –/– –/– –/– –)			52.4	27.7
ZRKGC [89]	T-Chat	16.1	– –(22.3/ 8.0 / 3.7 / 1.9)			42.8	
TF [57]	T-Chat	22			0.80/0.81#	43.6	
PD-NRG [68]	T-Chat*	22.3	– –(– –/– –/– –/ 2.0)	– –/– –/10.8	0.022/0.181#	12.62	
CMR [141]	CbR	0.38	– –(– –/– –/– –/ 1.38)		0.052/0.283		
RAM [185]	CbR	0.50	– –(– –/– –/– –/ 1.47)		0.053/0.287		
CGRG [209]	CbR*		– –(– –/– –/– –/ 3.26)		– –/0.116#		

We only show the Single-reference version of Holl-E and the Rare version of Topical-Chat. *means different data processing schema. # means div-1/2 from Ghazvininejad et al. [53]. "R1" stands for Recall@1 and is for KS.

example, generative models might use different data processing schema on the same dataset [77]; retrieval models on ARW, Persona-Chat, and WoW datasets used 10, 20, and 100 candidates, respectively. In this section, we analyze these experimental results and present some valuable conclusions.

6.1 Retrieval Models

In Table 8, we present the results of retrieval-based models. For an earlier ARW task [107], the TF-IDF weighted cosine similarity did not use external knowledge. The **Knowledge Encoder** (**KE**) model used the external knowledge, but its performance was restrained by a regular RNN encoder and was only slightly better than TF-IDF, which indicated that simple encoders were not able to capture enough semantic information in UTED task.

Table 10. The UTED Models' Performance on the Test Seen Version of WoW

Model	F1%	BLEU(BLEU-1/2/3/4)%	ROUGE(1/2/L)%	MT.%	Distinct(1/2)	PPL	R@1%
MKST [236]*		0.72(— —/— —/— —/— —)			0.091/0.341		
Adapter-Bot [111]	18.0	1.35(— —/— —/— —/— —)				19.5	
TED [241]*		— —(20.27/9.47/5.33/3.35)		8.45	0.039/0.162		
DRD [237]*	18.0	— —(21.8 /11.5 / 7.5 / 5.5)				52.0	
ZRKGC [89]*	18.9	— —(22.5 / 8.4 / 3.9 / 2.0)				41.1	
KIC [95]*		— —(17.3 /10.5 / 7.7 /— —)			0.138/0.363		
Decoupling [187]*	20.1					18.3	90.84#
Unlikelihood [90]	35.8					8.5	
PostKS [93]*	1.74	— —(17.2 / 6.9 / 3.4 /— —)			0.056/0.213		
Match-Reduce [2]	17.8	1.2 (— —/— —/— —/— —)				60.6	25.4
TMN [33]	18.9					46.5	
DiffKS(Dis) [239]*		— —(— —/11.3 /— —/ 5.7)	— —/6.8 /— —				24.7
DukeNet [120]*		— —(— —/— —/— —/2.43)	25.17/6.81/18.52	17.09*			26.38
SKT [77]*	19.3					52.0	26.8
PIPM+KDBTS [17]*	19.9					42.7	27.9
AKGCM [105]		— —(— —/— —/— —/6.94)	— —/7.38/17.02				18.24
KnowledGPT [238]	22.0					19.2	
Interleave [197]*	35.7					19.7	

"MT." stands for METEOR. * means different data processing schema or metric definition. "R@1" stands for Recall@1. # means the R@1 is for response selection while others are for KS.

For Persona-Chat [227], Starspace learned the similarity between the context and the response by optimizing the embedding directly with the margin ranking loss and k-negative sampling, KV Profile Memory used an attention-based key-value memory network to select a response. TF-IDF/Starspace/KV Profile Memory models served as the baselines which concatenated the profiles with dialogue context and their performance was outperformed by the following models that had more complicated structures and encoded context, knowledge, and response separately.

For both Persona-Chat and CMUDoG, HAKIM [173] employed a representation-based fusion method and learned an interaction vector of context and knowledge. DGMN [235] adopted the Interaction-based fusion method and performed a shallow matching against all fusion matrices. FIRE [58] also used Interaction-based Fusion but with a multi-levels Deep Matching. FIRE was better than DGMN on both Persona-Chat and CMUDoG, which indicated the advantage of multi-levels Deep Matching. HAKIM was worse than the other two on Persona-chat while better than them on CMUDoG. The reason might lie in the Fusion method and the difference between datasets. HAKIM employed Representation-based Fusion while others used Interaction-based. CUMDoG has a longer average length of utterances and external knowledge sentences than Persona-Chat, this entails a data sparsity problem that the dialogue context only contains little external knowledge.

For WoW [33], IR baseline used a simple word-overlap method and got the worst R@1 and F1. BoW MN was a BOW Memory Network [172], TMN was a pre-trained Transformer based Memory Network and fine-tuned on WoW. The latter benefits from the representation ability of the Transformer structure and had better results. For X-Persona [96], the Transformer-based model was also better than the LSTM-based one.

6.2 Generative Models

In Tables 9, 10, and 11, we summarize the auto evaluation results of generative models. For the Holl-E task in Table 9, the results are based on the single reference and oracle background (256

Table 11. The Generative Models' Performance on Persona-Chat/ConvAI2/DSTC-7/ED/LIGHT Datasets

Model	Dataset	F1%	BLEU(BLEU-1/2/3/4)%	MT.%	Ent-4	Distinct(1/2)	PPL	R1%
Profile Memory [227]	Persona						34.5	12.5
RCDG [167]	Persona					— —/0.127	29.9	
PostKS [93]	Persona*	8.7	— —(19.0 / 9.8 / 5.9 /— —)			0.046/0.134		
GDR [165]	Persona					0.037/0.227	16.7	
PAML [112]	Persona*		0.74(— —/— —/— —/— —)				41.6	
CMAML [168]	Persona*		1.70(— —/— —/— —/— —)			0.021/— —	36.3	
MKST [236]	Persona		0.82(— —/— —/— —/— —)			0.095/0.361		
Adapter-Bot [111]	Persona	15.0	1.55(— —/— —/— —/— —)				11.1	
Multi-Input [54]	Persona*		2.77(— —/— —/— —/— —)	7.88	9.211	— —/0.155		
TF+AMI [134]	Persona				9.695	0.104/0.385		
L-PCFG [39]	Persona*		— —(20.9*/— —/— —/— —)	7.77				
PEE [213]	Persona	18.4	— —(23.19/11.52/6.12/3.50)					
$\mathcal{P}^2$ Bot [102]	Persona	19.8					15.1	81.9
Decoupling [187]	Persona	19.9					18.7	66.67
GE [6]	Persona*	20.5				0.021/0.097		
Deepcopy [221]	ConvAI2		4.09(— —/— —/— —/— —)			— —/0.059	54.6	
LIC [55]	ConvAI2	17.7						17.1
Unlikelihood [90]	ConvAI2	19.3					11.9	
TransferTransfo [206]	ConvAI2	19.5					16.3	80.7
BART [84]	ConvAI2	20.7					11.9	
CG+CVAE [222]	DSTC-7*		— —(3.90/0.89/0.22/0.07)	2.62	6.427	0.012/0.027		
TED [241]	DSTC-7*		— —(14.18/5.01/2.11/0.95)	5.60		0.022/0.094		
CBS [177]	DSTC-7		— —(— —/— —/— —/ 1.83)	8.07	9.030	0.109/0.325		
CDD [178]	DSTC-7		— —(— —/— —/— —/ 2.43)			0.109/0.442		
Adapter-Bot [111]	ED		8.53(— —/— —/— —/— —)				12.2	
Decoupling [187]	LIGHT	18.7					23.2	59.63

"Persona" is short for Persona-Chat. "MT." stands for METEOR. *means a difference in data processing or metric definition. "Ent-4" is short for Entropy-4. "R1" stands for Recall@1 and is for response selection.

words) version of the data. We first compare some models with **implicit knowledge selection**. The GTTP [158] model was first proposed for Abstractive Summarization. It was used as one of the baseline models in the original Holl-E paper [127], the input was a concatenation of document and dialogue context and the output was a response. The poor performance was caused by the inefficient interaction between knowledge and context. The Cake [232] used the interaction mechanism in BiDAF [127] to perform knowledge pre-selection then produced response by copying from knowledge or generating from the vocabulary. The RefNet [119] extent the copy mechanism by directly selecting a semantic unit (e.g., a span containing complete semantic information) from the document. The GLKS [150] argued that previous models adopted a local perspective (select a token based on the current decoding state). They introduced a global perspective that pre-selected some text fragments to better guide the generation.

In the **explicit knowledge selection** models, Kim et al. [77] focused on the role of dialogue history in KS. They introduced SKT to select knowledge of each utterance based on a latent knowledge tracking vector. Based on SKT, Chen et al. [17] enhanced the prior selection module with a **Posterior Information Prediction Module** (**PIPM**) and propose a **Knowledge Distillation Based Training Strategy** (**KDBTS**) to overcome the exposure bias in KS. Meng et al. [120] argued that former work did not pay attention to the repetition of KS. The DukeNet they proposed had a knowledge shifter and tracker module to capture the topic flow and reduce knowledge repetition in dialogue. The DiffKS(Dis) [239] model focused on the knowledge repetition problem as

well. It used the dialogue context and the previously selected knowledge to compute two distributions on the candidate knowledge provided at the current turn, then combine these distributions to choose appropriate knowledge to be used in generation. Benefiting from the ability to select knowledge more accurately, DiffKS(Dis) had a higher BLEU and ROUGE than the DukeNet. The AKGCM [105] owned the best BLEU and Hit@1, which showed reasoning on an augmented KG was more effective on KS.

For the CMUDoG task in Table 9, ITDD/CAT/DIalogT/DRD adopted **implicit KS** methods, while KnowledGPT and Reduce-Match used the **explicit KS** method. DRD worked on the low-resource setting, and ZRKGC worked with the zero-resource setting. Their performances were very close except that models based on GPT-2 (KnowledGPT) or a deliberation decoder (ITDD and CAT) get a lower PPL.

The TF model is the baseline model proposed in Topical-Chat paper [57]. The PD-NRG [68] conducted experiments on an augmented version of Topical-Chat. They proved that dialogue action planning could benefit the response generation. Notably, the TF model computed the lexical diversity while the PD-NRG model used a corpus diversity.

For the CbR task, we compare three **implicit KS** models. The CMR used an MRC model to perform KS. The RAM further improved the CMR model by utilizing the response information with a Teacher–Student framework. However, the F1 on the CbR was lower than other datasets because the external knowledge was a much longer document hence the models had difficulty in locating the salient information. CGRG [209] introduced lexical control phrases and inductive attention in a pre-trained GPT-2. It got a higher BLEU-4 than CMR and RAM that used GloVe embeddings.

In Table 10, we introduce the models' performance on the WoW dataset. In the **implicit KS** models (the first half), different data processing schemes are usually adopted. MKST [236] and Adapter-Bot[7] [111] proposed methods to utilize different types of knowledge. MKST employed Pre-trained Transformer while Adapter-Bot adopted DialoGPT for multi-task learning and used BERT as a task identification. TED [241] performed KS with attention in the decoder. DRD [237] and ZRKGC [89] both utilized PTMs and got competitive results in limited-resource setting. KIC [95] designed a Knowledge-Interaction and Knowledge-Copy mechanism. The better distinction value indicated that copying from both knowledge and context could produce a more diverse response. Tuan et al. [187] trained the Decoupling model to automatically recouple context and related knowledge, their model performance on response selection is impressive. Unlikelihood [90] designed four novel losses to address the flaws in response generation, such as repetition and logic. The best F1 and PPL results showed these loss functions could facilitate model convergence.

In the **explicit KS** models (the second half), PostKS [93] employed a Teacher–Student module to leverage the response information. They employed a seq-to-seq model without a copy mechanism and did not use the ground-truth knowledge label. This caused their model hard to train and obtained the lowest F1 compared with other models. Match-Reduce [2] model matched each context utterance with knowledge sentences to capture fine-grained interactions and aggregated them as a training loss. TMN [33] was a baseline model that used a Transformer Memory network with explicit KS. The DiffKS(Dis) [239], AKGCM [105], DukeNet [120], SKT [77], and PIPM+KDBTS [17] conducted experiments on both Holl-E and WoW. The Holl-E dataset is a relatively simpler dataset compared with WoW. DiffKS(Dis) and AKGCM [105] performed KS better than the other three models on Holl-E but performed worse than them on WoW. The reason might be the different encoders they employed. DukeNet, SKT, and PIPM+KDBTS employed a BERT encoder which might encounter the over-fitting problem on Holl-E. This caused their performance worse than the

[7] Adapter-Bot powered with meta-knowledge has a big improvement (F1 35.5, BLEU 12.26, and PPL 9.04).

relatively simple encoding method (Glove+RNN), DiffKS, and AKGCM employed. Notably, AKGCM got a 42.04% Recall@1 on Holl-E, around 10% higher than the other four models. While on WoW, AKGCM only had an 18.24% Recall@1, around 9% lower than the other four models. Besides the difference of datasets, the convergence instability of RL they used when performing KS might lead to unstable performance. DukeNet employed METEOR Universal [30], a language-specific version of the METEOR metric. KnowledGPT [238] had the best PPL showing the effectiveness of GPT-2 [145]. Interleave [197] used different decoder layers to attend knowledge or context in a Seq-to-Seq Transformer. They had the highest F1 among the explicit KS models.

Among these models, only Match-Reduce, TED, KIC, and, Interleave randomly initialized word embeddings. Generally speaking, their performances are slightly lower than models with pre-trained embeddings except Interleave because of different data processing schemes.

In Table 11, the ConvAI2 competition was primarily based on the Persona-Chat dataset and there was a slight difference in data size between them. For the Persona-Chat task, we find that most models applied the **implicit KS** except GE [6], a Generation-Evaluation framework that adopted **explicit KS** method.

Profile Memory [227] was the baseline model proposed in the original article. It employed a key-value memory network for persona selection. RCDG [167] employed RL and NLI techniques to generate persona consistent dialogues. GDR [165] was a three stages method (generate-delete-rewrite) that deleted inconsistent terms from a generated response and further rewrote it to a context-consistent one. The better Distinct-2 and PPL results than RCDG showed the method's advantage.

PAML [112] utilized meta-learning to learn better initial parameters which could be quickly adapted to new personas by leveraging only a few dialogue samples collected from the same user. CMAML [168] proposed an algorithm based on meta-learning that customized a unique dialogue model for each task in the few-shot setting. MKST [236] and Adapter-Bot [111] aimed to leverage diverse knowledge source. The former had the second-best Distinct and the latter got the best PPL. The same trends are observed on WoW and ED tasks. It indicated that the utilization of multiple types of knowledge benefited the UTED task. Multi-Input [54] employed a PTM (GPT-1) to compare a number of architectures and training schemes. TF+AMI [134] introduced **Adversarial Mutual Information** (**AMI**) to identify joint interactions between source and target. Since the model used randomly initialized embedding on an LSTM encoder, it was the Adversarial training object that helped their model to get the best Distinct and Entropy-4 value among all models.

Du and Black [39] generated the words of a sentence according to the order of their first appearance in its lexicalized PCFG (L-PCFG) parse tree instead of the traditional left-to-right manner. PEE [213] was a neural topical expansion framework, which was able to extend the predefined knowledge with semantically correlated content before utilizing them to generate dialogue responses. The BLEU-1/2/3 results were the best, showing some advantages of the knowledge expansion method. Decoupling [187] and $\mathcal{P}^2$ Bot [102] were both based on PTM (GPT-1). The former focused on latent KS and the latter paid attention to model the understanding scheme between interlocutors. They both performed well on the R@1 metric, indicating that these methods were sensitive to differences between knowledge sentences.

For the ConvAI2 task, Deepcopy [221] had a better BLEU than all models. However, its Distinct-2 was not good. This was contrary to the model's motivation of using copy mechanism to enhance the ability of knowledge utilization. It showed that the auto evaluation metircs based on word overlap were insufficient to reflect the dialogue quality. LIC [55] and TransferTransfo[8] [206] used

[8]Kim et al. [78] proposed a self-consciousness method and a distractor memory to improve the persona consistency in dialogue. They tested their methods with three models including TransferTransfo.

PTM (GPT-1) and fine-tuned on ConvAI2. Unlikelihood [90] method had an competitive performance compared with BART [84] which was a PTM with denoising autoencoder and conducted experiments on ConvAI2 task. Among the results of Persona-Chat and ConvAI2, we observed that models with PTMs normally had a better PPL than others.

The DSTC-7 task was based on the CbR task except that the external knowledge was organized as sentences, not documents. CG+CVAE [222] tried different methods to calculate the latent variable in CVAE, but the GRU-based encoder may restrict the performance. TED [241] improved the performance with a Transformer-based model with randomly initialized embeddings. CBS [177] introduced Cluster-based Beam Search and exceeded TED in all metrics with a simple LSTM+copy framework. CDD [178] had a switcher to control whether to perform convergent or divergent decoding, it outperformed the CBS in BLEU-4 and Distinct-2. Since these models employed the random initialized embeddings or GloVe embeddings on simple encoders, we reckon that models based on large parameter PTMs will achieve better performance on this task.

6.3 Summary of Experimental Results

From the current UTED experimental results, we find some interesting conclusions: (1) Adding a copy mechanism is always helpful in Generative models [77, 119, 120, 150]; (2) PTMs are better at leveraging semantic information and can improve the models' performance in most cases; (3) Meta-learning is now only applied to the Persona-chat dataset and does not show obvious advantages. How to leverage the meta-learning in UTED requires more investigation; (4) Supplementary information is useful, such as knowledge expansion [113], syntactic tree structures [39], meta-knowledge [111]; (5) Experimental results on most datasets in retrieval models can still improve since the R@1 is below 83% and the candidate sizes are small; and (6) By observing the range of various metrics in generative models: BLEU in (0.7–8.6)%, METEOR in (2.6–8.5)%, Distinct-2 in (0.025–0.445), and F1 in (1.7–36)%, we can say that despite many different structures and algorithms that researchers have applied, the dialogue quality and knowledge utilization of the UTEDS still have a lot to improve.

7 FUTURE TRENDS

So far we introduced UTEDS from the perspective of retrieval and generative. The retrieval models can choose fluent responses from given-candidates but have difficulty adapting to specific dialogue context and external knowledge. In contrast, generative models can make better use of context and external knowledge and attract more attention from researchers. Benefiting from the current powerful PTMs [145, 233], generative models have made significant improvements in producing fluent responses but still have room for improvement. Figure 3 shows six future research trends for UTEDS. 7.1/7.2/7.3/7.5 refer to both retrieval and generative models, 7.4/7.6 are only suitable for generative models.

7.1 Limited Resource Problem

The UTED aims at imitating the real human dialogues where interlocutors can freely refer to their own knowledge. However, the current dataset constructed by researchers has the following problems: (1) **Small data size.** Some datasets (ARW [193], CMU_DoG [246], Holl-E [127], and so on) are not big enough for training large parameter models [246]; (2) **Knowledge sparsity.** Some datasets (GCD [53], CMU_DoG [246], CbR [141], and so on) suffer from knowledge sparsity [115, 242] problem which means the dialogue data collected only contains little external knowledge. This type of data distribution makes it difficult for the model to learn to inject knowledge into conversation.

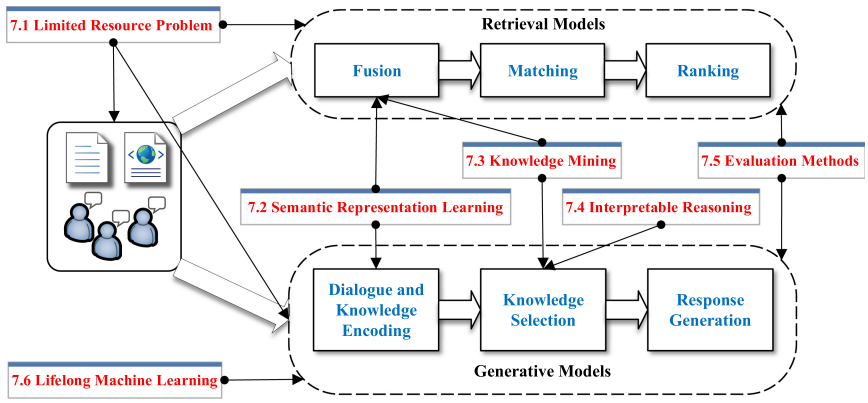

Fig. 3. The future research trends of the UTEDS. Each direction points to the related modules of Retrieval and Generative models.

The limited resource problem can be addressed in two directions. The first is *data construction*, a large-scale dataset with clear knowledge utilization will be helpful to train the UTED model. Due to the expense of artificial data construction, we can focus on large-scale dialogue data that the knowledge utilization is naturally labeled (e.g., when consulting legal texts, the legal knowledge involved in the dialogue has a clear source and is easy to locate). The second is *model structure*, a specific architecture focusing on knowledge sparsity can be designed. In a broad sense, we consider the human dialogues to consist of linguistic knowledge, dialogue intentions, and world knowledge. Linguistic knowledge is the grammatical and syntactical rules, which can be learned by PTMs [142]. The dialogue intentions is "what to act" in a dialogue (e.g., strategies in persuasion [196], negotiation [247] or recommendation [67] tasks, and yes-and dialogue strategy [20]). The world knowledge is "what to refer" in a dialogue, including commonsense knowledge, entities, and events, and so on. The problem is, some of the dialogue intentions require world knowledge injection, others do not. Even among the dialogue intentions which need knowledge injection, there is a difference between the knowledge utilization pattern. These differences cause the UTED model hard to learn the knowledge injection in different dialogue intentions. Therefore, we believe that a good UTED model needs to set up different modules to identify the dialogue intentions and control the world knowledge injection.

Besides the above two problems, there are **other limited resource problems**: (1) the amount of knowledge available in different languages is extremely imbalanced. For instance, the number of articles in English is hundreds of times larger than that of Bengali on Wikipedia. This limited resource problem can be addressed with cross-lingual [19, 27] methods. Retrieving relevant text from external knowledge in a rich resource language can improve the informativeness of the response generation in the low resource language; (2) many specialized domains contain their own specific terms that are not part of the pre-trained LM vocabulary. Some cross-domain [226] methods can be used to address this problem. Researchers can tap more potential from the cross-lingual and cross-domain methods.

7.2 Semantic Representation Learning

Semantic representation learning aims at learning universal language representations that contain salient language knowledge, which can be understood and used by a computer. It is a research Hotspot in NLP [32, 139, 144, 145].

In UTED, data distribution is different between knowledge and dialogue, and between different types of knowledge. For example, the language used in movie reviews is different from movie plots. We need the UTEDS that can better integrate dialogue and knowledge, understand the relationship between dialogue and knowledge, identify, select, and use knowledge according to the dialogue. The current knowledge representation methods are not up to the requirements. Some possible directions are: (1) borrowing ideas from the latest PTMs to fully utilize the structured representation ability of hidden variable space [28, 85]. The latent spaces can be used to represent knowledge with different data distribution; (2) the current models usually adopted PTMs and fine-tuning to distinguish tasks from the parameter perspective but ignored the model-structure perspective, resulting in similar dialogue models for different tasks. We can design specific model architecture to jointly learn dialogue and knowledge representations [64]; and (3) using task-related domain data to re-train PTM [63].

7.3 Knowledge Mining

We present two knowledge mining categories: (1) **Latent dialogue knowledge.** The UTEDS need to mine information not only from external knowledge but also from dialogue context. Latent attributes of interlocutors can help to build consistent and engaging DS. Tigunova et al. [186] addressed this type of knowledge acquisition problem by extracting personal attributes from the conversation. Specifically, user-generated social media text (Reddit) is used to infer the users experience from voice conversations (movies and Persona-Chat). The current research on extracting dialogue knowledge focused on exploiting persona [12], emotion [97], or action [7]. Future research can explore more user behaviors, habits, and other latent characteristics; (2) **Long text knowledge selection.** Unstructured knowledge usually comes from web-page text (Wikipedia, news, and so on), and these web page texts usually contain thousands of words. The existing methods for processing long texts used MRC technology to implicitly filter the entire text or explicitly truncate the text into independent sentences. The performance of implicit filtering decreases significantly as the length of the text increases,[9] while explicit filtering ignores the logical relationship between sentences and wastes text information. A possible direction is to integrate the advantages of implicit and explicit methods for long text KS.

7.4 Interpretable Reasoning

Interpretable reasoning means that the model can give an explicit, logical, and complete reasoning path to explain the reason of choosing the corresponding knowledge. Interpretable reasoning can not only help understand machine's reasoning logic but also meet the ethical requirements of artificial intelligence. For a long time, researchers have shown a great interest in constructing models with specific reasoning abilities, such as algebraic reasoning [25], logical reasoning [202], commonsense reasoning [133], multi-fact reasoning [40, 162, 176, 199], and multiple steps reasoning [76, 122, 200, 220].

In UTED, most models depended on the attention mechanism to conduct interpretable reasoning, but the effectiveness of attention for interpretability is still controversial [73, 161]. To address this, some work first used unstructured knowledge to construct a concrete memory (augmented KG [105] or event graph [212]), and then adopted reinforcement learning to reason on the memory. These explicit KS methods had difficulty in leveraging paragraph-level or document-level text information compared with attention-based methods. To conduct interpretable reasoning and leverage the global text information, we believe a more fine-grained structure that using both

[9]For example, on the Holl-E dataset, the RefNet model got 29.38 and 17.19 BLEU when using 256 and 1200 words external knowledge, respectively.

implicit KS in text and explicit KS in structured data is the future trend. Meanwhile, some traditional methods which possess good interpretability, such as symbolic logical reasoning, can be leveraged.

7.5 Evaluation Methods

Generally speaking, DS with good performance needs to produce a response with high semantic relevance, rich information, and diverse expressions. In UTEDS, the evaluation needs to reflect not only the dialogue quality but also the external knowledge utilization. For example, Zhao et al. [236] defined the Fusion F1 (F1 [33]+Knowledge F1 [93]) as an automatic metric to evaluate the quality of responses. While F1 evaluates the char-based F-score of prediction against gold response, Knowledge F1 evaluates the exact recall performance of the output response at the char level relative to the text knowledge. However, due to the expense of human evaluation and the deficiency of auto evaluation metrics in dialogue scenario [106, 180], we still need to find more reliable auto evaluation metrics highly correlated with human evaluation.

We believe that model-based evaluations are the future direction. Although earlier methods such as ADEM [106] which mimic human evaluation have been showed insensitive in adversarial scenarios [154], recent research [10] outlined that the current PTM can effectively approximate human annotations when annotating paragraph-level text. Another recently proposed USR [118] model leveraged the ability of PTM and used a regression model to approximate the specific scores rated by annotators. These studies shed light on the future direction of auto evaluations. However, there are still difficulties in training evaluation models because of the human-annotated training dataset: (1) the definition of human evaluation metrics lacks a unified standard. For example, there are no universal standards for what is an interesting response; (2) the workers with different knowledge backgrounds tend to give very different scores on the same metric; and (3) the annotation data size is small since the human annotations are expensive. We need to solve these problems which cause bias accumulation in training.

7.6 Lifelong Machine Learning

Lifelong machine learning (**LML**) [18, 126, 169, 184] requires that the deployed machine learning system have the capability to continuously improve themselves through interaction with the environment. As we introduced in Table 3, the UTED datasets involve different domains. We hope that a UTEDS has a certain memory ability and can learn a new domain knowledge without forgetting the old one, which is called **knowledge retention** [18]. We also hope that a UTEDS can comprehend by analogy, and can perform well in the new domain through existing knowledge, which is called **knowledge transfer** [169].

Knowledge retention and knowledge transfer can be seen as high-level abstract memory ability. The UTEDS with lifelong learning ability needs to design a memory module with appropriate spatial structure, distinguish good experience from bad ones, preserve experience knowledge, slow down catastrophic forgetting problem, update the obsolete knowledge, establish new knowledge in the face of unseen tasks, and balance the mutual influence of memory ability and performance. There is a lack of architecture, dataset, and benchmarks to test this high-level memory ability of the UTEDS.

8 CONCLUSION

The UTEDS need to select correct external knowledge according to dialogue and incorporate the knowledge into response generation. We believe that extracting unstructured text information during dialogue is the future trend in DS research because a large amount of human knowledge are contained in these texts. The research of the UTEDS not only possesses a broad application prospect

but also facilitates the DS to better understand human knowledge and natural language. This article introduces the UTEDS, defines the related concepts, analyzes the current datasets/model structures/evaluation methods/model performance, and provides views on future research trends, hoping to help researchers in this field.

REFERENCES

[1] Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018. A knowledge-grounded multimodal search-based conversational agent. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail Burtsev (Eds.). Association for Computational Linguistics, 59–66. DOI: https://doi.org/10.18653/v1/w18-5709

[2] Yeonchan Ahn, Sang-Goo Lee, and Jaehui Park. 2020. Exploiting text matching techniques for knowledge-grounded conversation. *IEEE Access* 8 (2020), 126201–126214. DOI: https://doi.org/10.1109/ACCESS.2020.3007893

[3] Martín Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *Proceedings of the 5th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=Hk4_qw5xe.

[4] Siddhartha Arora, Mitesh M. Khapra, and Harish G. Ramaswamy. 2019. On knowledge distillation from complex networks for response prediction. In *Proceedings of the NAACL-HLT (1)*. Association for Computational Linguistics, 3813–3822.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1409.0473

[6] Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. 2019. Know more about each other: evolving dialogue strategy via compound assessment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5382–5391. DOI: https://doi.org/10.18653/v1/p19-1535

[7] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 85–96. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.9/.

[8] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *NIPS*. MIT Press, 932–938.

[9] Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning*, (*5th. Ed.*). Springer. DOI: https://www.worldcat.org/oclc/71008143.

[10] Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. In *CIKM'20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 85–94. DOI: https://doi.org/10.1145/3340531.3412043

[11] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 4762–4779. DOI: https://doi.org/10.18653/v1/p19-1470

[12] Alex Boyd, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Large scale multi-actor generative dialog modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 66–84. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.8/.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020,*

December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[14] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 1870–1879.

[15] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explorations* 19, 2 (2017), 25–35.

[16] Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8635–8648. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.697

[17] Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3426–3437. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.275

[18] Zhiyuan Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR Workshop and Conference Proceedings, Vol. 32)*. JMLR.org, 703–711. Retrieved from http://proceedings.mlr.press/v32/chenf14.html

[19] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 7570–7577. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/6256.

[20] Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 2398–2413. DOI: https://www.aclweb.org/anthology/2020.acl-main.218/

[21] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1724–1734. DOI: https://doi.org/10.3115/v1/d14-1179

[22] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 2174–2184.

[23] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555 Retrieved from http://arxiv.org/abs/1412.3555.

[24] Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics, 26-29 June 1989, University of British Columbia, Vancouver, BC, Canada, Proceedings*, Julia Hirschberg (Ed.). ACL, 76–83. DOI: https://doi.org/10.3115/981623.981633

[25] Peter Clark. 2015. Elementary school science and math tests as a driver for AI: Take the aristo challenge!. In *Proceedings of the AAAI*. AAAI Press, 4019–4021.

[26] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence* 2, 1 (1971), 1–25. DOI: https://doi.org/10.1016/0004-3702(71)90002-6

[27] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 8440–8451. DOI: https://doi.org/10.18653/v1/2020.acl-main.747

[28] 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 *Proceedings of the 8th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=H1edEyBKDS.

[29] Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, and Jack W. Rae. 2019. Training language GANs from scratch. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 4302–4313. Retrieved from https://proceedings.neurips.cc/paper/2019/hash/a6ea8471c120fe8cc35a2954c9b9c595-Abstract.html.

[30] Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation* . The Association for Computer Linguistics, 376–380. DOI: https://doi.org/10.3115/v1/w14-3348

[31] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.* 54, 1 (2021), 755–810. https://doi.org/10.1007/s10462-020-09866-x

[32] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. Retrieved from https://www.aclweb.org/anthology/N19-1423/.

[33] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. arXiv:1811.01241 Retrieved from http://arxiv.org/abs/1811.01241.

[34] Xiao Ding, Zhongyang Li, Ting Liu, and Kuo Liao. 2019. ELG: An event logic graph. arXiv:1907.08015 Retrieved from http://arxiv.org/abs/1907.08015.

[35] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 138–145.

[36] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 13042–13054. Retrieved from http://papers.nips.cc/paper/9464-unified-language-model-pre-training-for-natural-language-understanding-and-generation.

[37] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani (Eds.). ACM, 601–610. DOI: https://doi.org/10.1145/2623330.2623623

[38] Wenchao Du and Alan W. Black. 2019. Boosting dialog response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 38–43. DOI: https://doi.org/10.18653/v1/p19-1005

[39] Wenchao Du and Alan W. Black. 2019. Top-down structurally-constrained neural response generation with lexicalized probabilistic context-free grammar. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 3762–3771. DOI: https://doi.org/10.18653/v1/n19-1377

[40] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2368–2378.

[41] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2020. Augmenting transformers with knn-based composite memory for dialogue. *Transactions of the Association for Computational Linguistics* 9, 0 (2021), 82–99. https://transacl.org/ojs/index.php/tacl/article/view/2419.

[42] Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 889–898. DOI: https://doi.org/10.18653/v1/P18-1082

[43] Song Feng, Kshitij P. Fadnis, Q. Vera Liao, and Luis A. Lastras. 2020. Doc2Dial: A framework for dialogue composition grounded in documents. In *The 34th AAAI Conference on Artificial Intelligence, AAAI 2020, The 32nd Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The 10th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020.* AAAI Press, 13604–13605. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/7089.

[44] Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A Goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8118–8128. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.652

[45] Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 236–245. Retrieved from https://www.aclweb.org/anthology/2020.sigdial-1.29/.

[46] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1126–1135. Retrieved from http://proceedings.mlr.press/v70/finn17a.html.

[47] Alice E. Fischer and Frances S. Grodzinsky. 1993. *The Anatomy of Programming Languages*. Prentice Hall.

[48] Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Proceedings of the Nips, Modern Machine Learning and Natural Language Processing Workshop*, Vol. 2.

[49] Fabian Galetzka, Chukwuemeka Uchenna Eneh, and David Schlangen. 2020. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 565–573. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.71/.

[50] Jianfeng Gao, Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Heung-Yeung Shum. 2020. Robust conversational AI with grounded text generation. arXiv:2009.03457 Retrieved from https://arxiv.org/abs/2009.03457.

[51] Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 1229–1238. DOI: https://doi.org/10.18653/v1/n19-1125

[52] Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. 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*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 1814–1823. DOI: https://doi.org/10.18653/v1/D19-1190

[53] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI*. AAAI Press, 5110–5117.

[54] Sergey Golovanov, Rauf Kurbanov, Sergey I. Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6053–6058. DOI: https://doi.org/10.18653/v1/p19-1608

[55] Sergey Golovanov, Alexander Tselousov, Rauf Kurbanov, and Sergey I Nikolenko. 2020. Lost in conversation: A conversational agent based on the transformer and transfer learning. In *Proceedings of the The NeurIPS'18 Competition*. Springer, 295–315.

[56] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680. Retrieved from http://papers.nips.cc/paper/5423-generative-adversarial-nets.

[57] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations, Gernot Kubin and Zdravko Kacic (Eds.). ISCA, 1891–1895. https://doi.org/10.21437/Interspeech.2019-3079

[58] Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, Si Wei, and Xiaodan Zhu. 2020. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1412–1422. https://doi.org/10.18653/v1/2020.findings-emnlp.127

[59] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL (1)*. The Association for Computer Linguistics.

[60] Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. DOI: https://doi.org/10.18653/v1/p16-1014

[61] Bin Guo, Hao Wang, Yasan Ding, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. 2019. c-TextGen: Conditional text generation for harmonious human-machine interaction. *ACM Trans. Intell. Syst. Technol.* 12, 2, Article 14 (Feb. 2021), 50 pages. https://doi.org/10.1145/3439816

[62] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for Ad-hoc retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 55–64. DOI: https://doi.org/10.1145/2983323.2983769

[63] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 8342–8360. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.740/.

[64] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3929–3938. http://proceedings.mlr.press/v119/guu20a.html.

[65] Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 4832–4839. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16691.

[66] Katsuhiko Hayashi and Masashi Shimbo. 2017. On the equivalence of holographic and complex embeddings for link prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 554–559. DOI: https://doi.org/10.18653/v1/P17-2088

[67] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8142–8152. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.654/.

[68] Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tür. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, Brian Davis, Yvette Graham, John D. Kelleher, and Yaji Sripada (Eds.). Association for Computational Linguistics, 412–421. https://www.aclweb.org/anthology/2020.inlg-1.46/.

[69] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2020. Knowledge graphs. *ACM Comput. Surv.* 54, 4 (2021), 71:1–71:37. https://doi.org/10.1145/3447772

[70] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rygGQyrFvH.

[71] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2019. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.* 38, 3 (2020), 21:1–21:32. https://doi.org/10.1145/3383123

[72] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of the 8th International Conference on Learning Representations*. . Retrieved from https://openreview.net/forum?id=SkxgnnNFvH.

[73] Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 3543–3556. DOI: https://doi.org/10.18653/v1/n19-1357

[74] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. Retrieved from https://openreview.net/forum?id=rkE3y85ee.

[75] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2020. A survey on knowledge graphs: Representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems* (2021), 1–21. https://doi.org/10.1109/TNNLS.2021.3070843

[76] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the NAACL-HLT*. Association for Computational Linguistics, 252–262.

[77] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *Proceedings of the 8th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=Hke0K1HKwr.

[78] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? Improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 904–916. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.65

[79] Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tür. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 278–289. Retrieved from https://www.aclweb.org/anthology/2020.sigdial-1.35/.

[80] Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and Dilek Hakkani-Tür. 2021. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access track in DSTC9. arXiv:2101.09276. Retrieved from https://arxiv.org/abs/2101.09276.

[81] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 3294–3302. Retrieved from http://papers.nips.cc/paper/5950-skip-thought-vectors.

[82] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the WMT@ACL*. Association for Computational Linguistics, 228–231.

[83] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 1203–1213. DOI: https://doi.org/10.18653/v1/d16-1128

[84] 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, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7871–7880. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.703/.

[85] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4678–4699. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.378/.

[86] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the HLT-NAACL*. The Association for Computational Linguistics, 110–119.

[87] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *EMNLP*. The Association for Computational Linguistics, 1192–1202.

[88] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 2157–2169. DOI: https://doi.org/10.18653/v1/d17-1230

[89] Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020*, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/609c5e5089a9aa967232aba2a4d03114-Abstract.html.

[90] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4715–4728. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.428/.

[91] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. arXiv:1909.03087 Retrieved from http://arxiv.org/abs/1909.03087.

[92] Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 12–21.

[93] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the IJCAI*. ijcai.org, 5081–5087.

[94] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*. 74–81.

[95] Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 41–52. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.6/.

[96] Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. XPersona: Evaluating multilingual personalized chatbot. arXiv:2003.07568 Retrieved from https://arxiv.org/abs/2003.07568.

[97] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. 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*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 121–132. https://doi.org/10.18653/v1/D19-1012

[98] Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. 2020. Variational transformers for diverse response generation. arXiv:2003.12738 Retrieved from https://arxiv.org/abs/2003.12738.

[99] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. CAiRE: An end-to-end empathetic chatbot. In *The 34th AAAI Conference on Artificial Intelligence, The 32nd Innovative Applications of Artificial Intelligence Conference. The 10th AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 13622–13623. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/7098.

[100] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the EMNLP*. The Association for Computational Linguistics, 2122–2132.

[101] Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical inference for multi-relational embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2168–2178. Retrieved from http://proceedings.mlr.press/v70/liu17d.html.

[102] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1417–1427. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.131/.

[103] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 1694–1704.

[104] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Retrieved from http://arxiv.org/abs/1907.11692.

[105] Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with reasoning on augmented graph. arXiv:1903.10245. Retrieved from http://arxiv.org/abs/1903.10245.

[106] 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 ACL (1)*. Association for Computational Linguistics, 1116–1126.

[107] Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, and Joelle Pineau. 2015. Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Proceedings of the Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*.

[108] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL Conference*. The Association for Computer Linguistics, 285–294.

[109] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1412–1421. DOI: https://doi.org/10.18653/v1/d15-1166

[110] Longxuan Ma, Wei-Nan Zhang, Runxin Sun, and Ting Liu. 2020. A compare aggregate transformer for understanding document-grounded dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1358–1367. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.122

[111] Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. The adapter-bot: All-in-one controllable conversational model. arXiv:2008.12579. Retrieved from https://arxiv.org/abs/2008.12579.

[112] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5454–5459. DOI: https://doi.org/10.18653/v1/p19-1542

[113] Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian J. McAuley. 2020. Like hiking? You probably enjoy nature: Persona-grounded dialog with commonsense expansions. arXiv:2010.03205 Retrieved from https://arxiv.org/abs/2010.03205.

[114] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian J. McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8129–8141. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.653

[115] Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4252–4273. https://doi.org/10.18653/v1/2020.emnlp-main.348

[116] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2775–2779. DOI: https://doi.org/10.18653/v1/d18-1298

[117] Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 225–235. Retrieved from https://www.aclweb.org/anthology/2020.sigdial-1.28/.

[118] Shikib Mehri and Maxine Eskénazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 681–707. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.64/.

[119] Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. RefNet: A reference-aware network for background based conversation. The *Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7-12, 2020. AAAI Press, 8496–8503. https://aaai.org/ojs/index.php/AAAI/article/view/6370.

[120] Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. DukeNet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1151–1160. DOI: https://doi.org/10.1145/3397271.3401097

[121] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupala, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (Eds.). Association for Computational Linguistics, 33–44. https://doi.org/10.18653/v1/2020.blackboxnlp-1.4

[122] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 2381–2391.

[123] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1301.3781.

[124] Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations*, Lucia Specia, Matt Post, and Michael Paul (Eds.). Association for Computational Linguistics, 79–84. Retrieved from https://www.aclweb.org/anthology/D17-2014/.

[125] Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the EMNLP*. The Association for Computational Linguistics, 1400–1409.

[126] Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2015. Never-ending learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, Blai Bonet and Sven Koenig (Eds.). AAAI Press, 2302–2310. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10049.

[127] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 2322–2332.

[128] Nikita Moghe, Priyesh Vijayan, Balaraman Ravindran, and Mitesh M. Khapra. 2020. On incorporating structural information to improve dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. 11–24.

[129] Roy Mor, Ben Kantor, and Emma Rapoport. 2018. Adversarial approach to persona based dialogue agents. In *Proceedings of Machine Learning Research*, Vol. 80. Retrieved from https://www.cs.tau.ac.il/~joberant/teaching/nlp_spring_2019/past_projects/adversarial_dialogue.pdf.

[130] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2016), 11–33. DOI: https://doi.org/10.1109/JPROC.2015.2483592

[131] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 1955–1961. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484.

[132] Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 1723–1731. Retrieved from http://papers.nips.cc/paper/6547-reward-augmented-maximum-likelihood-for-neural-structured-prediction.

[133] Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the SemEval@NAACL-HLT*. Association for Computational Linguistics, 747–757.

[134] Boyuan Pan, Yazheng Yang, Kaizhao Liang, Bhavya Kailkhura, Zhongming Jin, Xian-Sheng Hua, Deng Cai, and Bo Li. 2020. Adversarial mutual information for text generation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 7476–7486. http://proceedings.mlr.press/v119/pan20a.html.

[135] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the ACL*. ACL, 311–318.

[136] Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 690–695.

[137] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: A novel multi-domain information seeking dialogues dataset. arXiv:1912.04639 Retrieved from http://arxiv.org/abs/1912.04639.

[138] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. DOI: https://doi.org/10.3115/v1/d14-1162

[139] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the NAACL-HLT*. Association for Computational Linguistics, 2227–2237.

[140] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 2463–2473. DOI: https://doi.org/10.18653/v1/D19-1250

[141] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 5427–5436.

[142] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63, 10 (2020), 1872–1897.

[143] Meng Qu and Jian Tang. 2019. Probabilistic logic neural networks for reasoning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 7710–7720. Retrieved from http://papers.nips.cc/paper/8987-probabilistic-logic-neural-networks-for-reasoning.

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

[145] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019).

[146] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *EMNLP*. The Association for Computational Linguistics, 2383–2392.

[147] 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. arXiv:1801.03604. Retrieved from http://arxiv.org/abs/1801.03604.

[148] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5370–5381. DOI: https://doi.org/10.18653/v1/p19-1534

[149] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *TACL* 7 (2019), 249–266.

[150] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7-12, 2020. AAAI Press, 8697–8704. https://aaai.org/ojs/index.php/AAAI/article/view/6395

[151] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with search engines. arXiv:2004.14162. Retrieved from https://arxiv.org/abs/2004.14162.

[152] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 300–325. https://www.aclweb.org/anthology/2021.eacl-main.24/.

[153] Vasile Rus and Mihai C. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational*

Applications Using NLP, BEA@NAACL-HLT 2012, June 7, 2012, Montréal, Canada, Joel R. Tetreault, Jill Burstein, and Claudia Leacock (Eds.). The Association for Computer Linguistics, 157–162. Retrieved from https://www.aclweb.org/anthology/W12-2018/.

[154] Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating ADEM: A deeper look at scoring dialogue responses. In *Proceedings of AAAI*. AAAI Press, 6220–6227.

[155] Sashank Santhanam and Samira Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems - past, present and future directions. arXiv:1906.00500. Retrieved from http://arxiv.org/abs/1906.00500.

[156] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, the 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 3027–3035. DOI: https://doi.org/10.1609/aaai.v33i01.33013027

[157] Thomas Scialom, Serra Sinem Tekiroglu, Jacopo Staiano, and Marco Guerini. 2020. Toward stance-based personas for opinionated dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 2625–2635. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.238

[158] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 1073–1083.

[159] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HJ0UKP9ge

[160] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *D&D* 9, 1 (2018), 1–49.

[161] Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable?. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2931–2951. DOI: https://doi.org/10.18653/v1/p19-1282

[162] Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. ReasoNet: Learning to stop reading in machine comprehension. In *Proceedings of the KDD*. ACM, 1047–1055.

[163] Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: Challenges and opportunities with social chatbots. *Frontiers of IT & EE* 19, 1 (2018), 10–26.

[164] 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 the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 2021–2030. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.183/.

[165] Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5821–5831. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.516/.

[166] Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 5190–5196. DOI: https://doi.org/10.24963/ijcai.2019/721

[167] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *The 34th AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8878–8885. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/6417.

[168] Yiping Song, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang. 2020. Learning to customize model structures for few-shot dialogue generation tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5832–5841. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.517/.

[169] Luc Steels. 1995. The biology and technology of intelligent autonomous agents. *Robotics Autonoumous System* 15, 1-2 (1995), 1–2. DOI: https://doi.org/10.1016/0921-8890(95)00010-D

[170] Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-yi Lee. 2019. Personalized dialogue response generation learned from monologues. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, Gernot Kubin and Zdravko Kacic (Eds.). ISCA, 4160–4164. DOI: https://doi.org/10.21437/Interspeech.2019-1696

[171] Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7087–7097. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.634/.

[172] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 2440–2448. Retrieved from http://papers.nips.cc/paper/5846-end-to-end-memory-networks.

[173] Yajing Sun, Yue Hu, Luxi Xing, Jing Yu, and Yuqiang Xie. 2020. History-adaption knowledge incorporation mechanism for multi-turn dialogue system. In *The 34th AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8944–8951. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/6425.

[174] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the NIPS*. 3104–3112.

[175] Idan Szpektor, Deborah Cohen, Gal Elidan, Michael Fink, Avinatan Hassidim, Orgad Keller, Sayali Kulkarni, Eran Ofek, Sagie Pudinsky, Asaf Revach, Shimi Salant, and Yossi Matias. 2020. Dynamic composition for conversational domain exploration. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 872–883. DOI: https://doi.org/10.1145/3366423.3380167

[176] Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the NAACL-HLT*. Association for Computational Linguistics, 641–651.

[177] Yik-Cheung Tam. 2020. Cluster-based beam search for pointer-generator chatbot grounded by knowledge. *Computer Speech Language* 64 (2020), 101094. DOI: https://doi.org/10.1016/j.csl.2020.101094

[178] Ryota Tanaka and Akinobu Lee. 2020. Fact-based dialogue generation with convergent and divergent decoding. arXiv:2005.03174 Retrieved from https://arxiv.org/abs/2005.03174.

[179] Xiangru Tang and Po Hu. 2019. Knowledge-aware self-attention networks for document grounded dialogue generation. In *KSEM (2) (Lecture Notes in Computer Science, Vol. 11776)*. Springer, 400–411.

[180] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI*. AAAI Press, 722–729.

[181] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2020. Synthesizer: Rethinking self-attention in transformer models. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10183–10192. http://proceedings.mlr.press/v139/tay21a.html.

[182] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 4593–4601. DOI: https://doi.org/10.18653/v1/p19-1452

[183] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=SJzSgnRcKX.

[184] Sebastian Thrun. 1995. Is learning the n-th thing any easier than learning the first?. In *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27-30, 1995*, David S. Touretzky, Michael Mozer, and Michael E. Hasselmo (Eds.). MIT Press, 640–646. Retrieved from http://papers.nips.cc/paper/1034-is-learning-the-n-th-thing-any-easier-than-learning-the-first.

[185] Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, Yiping Song, Xiaojiang Liu, and Nevin L. Zhang. 2020. Response-anticipated memory for on-demand knowledge integration in response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 650–659. https://doi.org/10.18653/v1/2020.acl-main.61

[186] Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 1818–1828. DOI: https://doi.org/10.1145/3308558.3313498

[187] Yi-Lin Tuan, Wei Wei, and William Yang Wang. 2020. Unsupervised injection of knowledge into dialogue generation via language models. arXiv:2004.14614 Retrieved from https://arxiv.org/abs/2004.14614.

[188] Alan M. Turing. 2009. Computing machinery and intelligence. In *Proceedings of the Parsing the Turing Test*. Springer, 23–65.

[189] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 673–683. DOI: https://doi.org/10.18653/v1/D19-1062

[190] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NIPS*. 5998–6008.

[191] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 4566–4575. DOI: https://doi.org/10.1109/CVPR.2015.7299087

[192] Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial Intelligence Review* 18, 2 (2002), 77–95. DOI: https://doi.org/10.1023/A:1019956318069

[193] Pavlos Vougiouklis, Jonathon S. Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of the COLING*. ACL, 3370–3380.

[194] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions Knowledge Data Engineering* 29, 12 (2017), 2724–2743. DOI: https://doi.org/10.1109/TKDE.2017.2754499

[195] Shuohang Wang and Jing Jiang. 2017. Machine comprehension using Match-LSTM and answer pointer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. Retrieved from https://openreview.net/forum?id=B1-q5Pqxl.

[196] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5635–5649. DOI: https://doi.org/10.18653/v1/p19-1566

[197] Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. 2019. Improving conditioning in context-aware sequence to sequence models. arXiv:1911.09728 Retrieved from http://arxiv.org/abs/1911.09728.

[198] Joseph Weizenbaum. 1966. ELIZA - A computer program for the study of natural language communication between man and machine. *Communication ACM* 9, 1 (1966), 36–45.

[199] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the NUT@EMNLP*. Association for Computational Linguistics, 94–106.

[200] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *TACL* 6, 0 (2018), 287–302.

[201] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *Proceedings of the 8th International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=SJeYe0NtvH.

[202] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1502.05698.

[203] Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1410.3916.

[204] Jason Weston, Emily Dinan, and Alexander H. Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail Burtsev (Eds.). Association for Computational Linguistics, 87–92. Retrieved from https://www.aclweb.org/anthology/W18-5713/.

[205] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1511.08198.

[206] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. arXiv:1901.08149. Retrieved from http://arxiv.org/abs/1901.08149.

[207] Sixing Wu, Ying Li, Dawei Zhang, and Zhonghai Wu. 2020. Improving knowledge-aware dialogue response generation by using human-written prototype dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1402–1411. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.126

[208] Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 7281–7288. DOI: https://doi.org/10.1609/aaai.v33i01.33017281

[209] Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A controllable model of grounded response generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14085–14093. https://ojs.aaai.org/index.php/AAAI/article/view/17658.

[210] Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard H. Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 950–962. DOI: https://doi.org/10.18653/v1/P17-1088

[211] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Review conversational reading comprehension. arXiv:1902.00821. Retrieved from http://arxiv.org/abs/1902.00821.

[212] Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2020. Enhancing dialog coherence with event graph grounded content planning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 3941–3947. DOI: https://doi.org/10.24963/ijcai.2020/545

[213] Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 325)*, Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang (Eds.). IOS Press, 2244–2251. DOI: https://doi.org/10.3233/FAIA200351

[214] Rui Yan. 2018. "Chitty-chitty-chat bot": Deep learning for conversational AI. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 5520–5526. DOI: https://doi.org/10.24963/ijcai.2018/778

[215] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2Image: Conditional image generation from visual attributes. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 9908)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 776–791. DOI: https://doi.org/10.1007/978-3-319-46493-0_47

[216] Fan Yang, Zhilin Yang, and William W. Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 2319–2328. Retrieved from http://papers.nips.cc/paper/6826-differentiable-learning-of-logical-rules-for-knowledge-base-reasoning.

[217] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3–7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1341–1350. DOI: https://doi.org/10.1145/3357384.3357881

[218] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of the SIGIR*. ACM, 245–254.

[219] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS'19), December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 5754–5764.

[220] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 2369–2380.

[221] Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tür. 2019. DeepCopy: Grounded response generation with hierarchical pointer networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino (Eds.). Association for Computational Linguistics, 122–132. DOI: https://doi.org/10.18653/v1/W19-5917

[222] Hao-Tong Ye, Kai-Ling Lo, Shang-Yu Su, and Yun-Nung Chen. 2019. Knowledge-grounded response generation with deep attentional latent-variable model. *Comput. Speech Lang.* 63 (2020), 101069. https://doi.org/10.1016/j.csl.2020.101069

[223] Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, R. Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Xiang Gao, Huda AlAmri, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2019. Dialog system technology challenge 7. arXiv:1901.03461 Retrieved from http://arxiv.org/abs/1901.03461.

[224] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A survey of knowledge-enhanced text generation. arXiv:2010.04389. Retrieved from https://arxiv.org/abs/2010.04389.

[225] Yury Zemlyanskiy and Fei Sha. 2018. Aiming to know you better perhaps makes me a more engaging dialogue partner. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, Anna Korhonen and Ivan Titov (Eds.). Association for Computational Linguistics, 551–561. DOI: https://doi.org/10.18653/v1/k18-1053

[226] Rong Zhang, Revanth Gangi Reddy, Md. Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 5461–5468. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.440

[227] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 2204–2213.

[228] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations* . Retrieved from https://openreview.net/forum?id=SkeHuCVFDr.

[229] Wen Zhang, Bibek Paudel, Wei Zhang, Abraham Bernstein, and Huajun Chen. 2019. Interaction embeddings for prediction and explanation in knowledge graphs. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 96–104. DOI: https://doi.org/10.1145/3289600.3291014

[230] Yuyu Zhang, Xinshi Chen, Yuan Yang, Arun Ramamurthy, Bo Li, Yuan Qi, and Le Song. 2020. Efficient probabilistic logic reasoning with graph neural networks. In *Proceedings of the 8th International Conference on Learning Representations.*. Retrieved from https://openreview.net/forum?id=rJg76kStwH.

[231] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of the NeurIPS*. 1815–1825.

[232] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2019. Improving background based conversation with context-aware knowledge pre-selection. In *Proceedings of the 2019 IJCAI Workshop SCAI: The 4th International Workshop on Search-Oriented Conversational AI.*

[233] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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, ACL 2020, Online, July 5-10, 2020*, Asli Çelikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, 270–278. Retrieved from https://www.aclweb.org/anthology/2020.acl-demos.30/.

[234] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the ACL (1)*. Association for Computational Linguistics, 654–664.

[235] Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A document-grounded matching network for response selection in retrieval-based chatbots. In *Proceedings of the IJCAI*. ijcai.org, 5443–5449.

[236] Xiangyu Zhao, Longbiao Wang, Ruifang He, Ting Yang, Jinxin Chang, and Ruifang Wang. 2020. Multiple knowledge syncretic transformer for natural dialogue generation. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 752–762. DOI: https://doi.org/10.1145/3366423.3380156

[237] Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. In *Proceedings of the 8th International Conference on Learning Representations*. DOI: https://openreview.net/forum?id=rJeIcTNtvS

[238] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3377–3390. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.272

[239] Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16–20 November 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 115–125. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.11

[240] Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2020. Approximation of response knowledge retrieval in knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16–20 November 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3581–3591. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.321

[241] Wen Zheng and Ke Zhou. 2019. Enhancing conversational dialogue models with grounded knowledge. In *Proceedings of the CIKM*. ACM, 709–718.

[242] Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, the 32nd Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence*. DOI: https://aaai.org/ojs/index.php/AAAI/article/view/6518

[243] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6556–6566. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.531

[244] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the IJCAI*. ijcai.org, 4623–4629.

[245] Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7098–7108. DOI: https://www.aclweb.org/anthology/2020.acl-main.635/

[246] Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 708–713.

[247] Yiheng Zhou, He He, Alan W. Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino (Eds.). Association for Computational Linguistics, 367–378. DOI: https://doi.org/10.18653/v1/W19-5943

Received December 2020; revised March 2021; accepted April 2021