공학박사학위논문

Leveraging Conversational Characteristics for Knowledge Selection and Ranking in Knowledge Grounded Conversation

지식 기반 대화에서의 대화 특성을 활용한 지식 선택 및 랭킹 방법

2022 년 8 월

서울대학교 대학원

전기·컴퓨터 공학부

안 연 찬

공학박사학위논문

Leveraging Conversational Characteristics for Knowledge Selection and Ranking in Knowledge Grounded Conversation

지식 기반 대화에서의 대화 특성을 활용한 지식 선택 및 랭킹 방법

2022 년 8 월

서울대학교 대학원

전기·컴퓨터 공학부

안 연 찬

Leveraging Conversational Characteristics
for Knowledge Selection and Ranking
in Knowledge Grounded Conversation

지식 기반 대화에서의
대화 특성을 활용한 지식 선택 및 랭킹 방법

지도교수 이 상 구

이 논문을 공학박사학위논문으로 제출함

2022 년 4 월

서울대학교 대학원

전기·컴퓨터 공학부

안 연 찬

Yeonchan Ahn의 공학박사 학위논문을 인준함

2022 년 6 월

위 원 장	김형주	(인)
부위원장	이상구	(인)
위 원	김건희	(인)
위 원	심준호	(인)
위 원	박영기	(인)

Abstract

Knowledge grounded conversation (KGC) model aims to generate informative responses relevant to both conversation history and external knowledge. One of the most important parts of KGC models is to find the knowledge which provides the basis on which the responses are grounded. If the model selects inappropriate knowledge, it may produce responses that are irrelevant or lack knowledge. In this dissertation, we study the methods of leveraging conversational characteristics to select or rank the knowledge for knowledge grounded conversation.

In particular, this dissertation provides novel two methods, where one of which focuses on the sequential structure of multi-turn conversation, and the other focuses on utilizing local context and topic of a long conversation. We first propose two knowledge selection strategies of which one preserves the sequential matching features and the other encodes the sequential nature of the conversation. Second, we propose a novel knowledge ranking model that composes an appropriate range of relevant documents by exploiting both the topic keywords and local context of a conversation. In addition, we apply the knowledge ranking model in quote recommendation with our new quote recommendation framework that provides hard negative samples to the model. Our experimental results show that the KGC models based on our proposed knowledge selection and ranking methods outperform the competitive models in terms of groundness and relevance.

Keywords: Knowledge grounded conversation, Open-domain dialogue system, Semantic matching, Knowledge selection, Knowledge ranking, Neural network

Student Number: 2012-30214

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans often acquire or convey valuable information in the form of natural language in offline and online spaces. Recently, people usually share news stories or previously unknown knowledge on the internet platform such as Reddit[1] and 지식iN (Korean community-based QA service)[2]. Such internet platforms usually allow the users to interact with others with free-form text, which can be considered conversation or dialogue. Many researchers and practitioners have been interested in building open-domain dialogue systems (DSs), which can automatically respond to users' conversational text in various domains. As a result, modern chatbots such as Microsoft XiaoIce [5] and Tay[3] emerged and attracted the public's attention.

Prior to such advances, many efforts have been made over a long history. From 1965, when the term AI was first coined at the workshop [6], up to now, there have been several landmark progress in Artificial Intelligence (AI) tech-

[1] `https://www.reddit.com/` (Accessed: June 24th, 2022)
[2] `https://kin.naver.com/` (Accessed: June 24th, 2022)
[3] `https://en.wikipedia.org/wiki/Tay_(bot)` (Accessed: June 24th, 2022)

nology. Since the introduction of transformational generative grammars [7], theories regarding the representation of meaning and explanation of syntactic anomalies of language have been established, and various DSs have emerged as a result. Representative systems in the early ages include Eliza [8], Alice [9], SHRDLU [10], and Parry [11]. They demonstrated the potential of the technologies at that moment but worked only in constrained environments. Then many notable studies in the AI fields were conducted from the 1970s to the 2000s; They include the 1970s' symbolic approaches for understanding natural language, 1980s' approaches based on complex hard-coded rules and grammars, and statistical models from the late 1980s to 2000s [12]. In 2000, Bengio et al. [13] first applied a neural network to language models, which have been the base of the modern neural approaches. After that, the important building blocks such as Seq2seq framework [14], attention mechanism [15, 16], and pre-trained models [17, 18] are developed sequentially. The research on open-domain DS, the topic of this dissertation, has also made a big progress thanks to the advances in neural network technology.

Despite such advances in AI technologies, many challenges still exist in building open-domain DS, including understanding of user's intents, lack of informativeness in response generation, and consistency between system's responses, as pointed out in the literature [19, 20]. Among the above challenges, the lack of informativeness in response generation is the main interest of this dissertation. To deal with this issue, researchers have focused on three approaches: promoting diversity, pre-trained language models (PLMs), and injecting external knowledge. Diversity promoting method [21, 1, 20] and PLM [22, 23] have been proven to be effective for yielding informative responses. However, these two approaches can produce factually false statements called hallucination [24, 25] because it does not lean on world knowledge directly. The

approach, injecting external knowledge into the conversational models, such as [26, 27, 28] has focused on structured or unstructured knowledge resources. In this study, we are interested in injecting unstructured knowledge into the conversation model. Compared to structured knowledge, unstructured knowledge has the following advantages. Firstly, the unstructured text embraces a variety of resources such as encyclopedia, personal profile [29], personal opinion [30], and news stories. Secondly, unstructured text can be continuously updated because numerous users on the internet can be authors of the knowledge.

In the dissertation, we aim to build knowledge grounded conversation (KGC) [26] models capable of generating knowledge grounded responses for the given conversation history by developing knowledge selection and ranking modules, which uses external knowledge. We define knowledge-grounded responses as the responses that are relevant to both conversation history and documents in knowledge base (KB). The model should identify knowledge that fits the conversation context to generate knowledge grounded responses. Otherwise, it can cause the response generator to produce irrelevant responses or responses lacking knowledge. Table 1.1 presents an example of a KGC, where response R-2, which reflects the best knowledge sentence (K-1), is preferred to response R-1. Response R-2 is both context-coherent and informative, while R-1 conveys not much useful information though it is context-coherent. By conjugating conversation context and external knowledge in a balanced manner, the model can generate a coherent and informative response. Table 1.1 illustrates the challenge of matching keywords, e.g., "diversity" from the conversation context to "cultural, financial, and media" from the knowledge. The primary research problem is to find the appropriate knowledge sentence(s) while understanding the semantics of the discourse to provide informative and contextual responses.

In this study, we explore the knowledge selection/ranking methods that

Conversation	
Topic:	New York City
A-1:	Hi, have you ever been to New York City?
B-1:	No, I haven't. Unfortunately, I've never been to U.S.
A-2:	I'm sorry to hear that. Have you ever heard about New York City?
B-2:	Yeah, I heard it is full of diversity.
R-1:	New York has always fascinated me.
R-2:	I've been there. New York City is the cultural, financial, and media capital of the world.

Candidate Knowledge Sentence	
K-1:	A global power city, New York City has been described as the cultural, financial, and media capital of the world.
K-2:	Located in the southeast part of the New York State, the city is the center of the New York metropolitan area.
K-3:	The United States has a very diverse population; 37 ancestry groups have more than one million members.

Table 1.1: Example of knowledge grounded conversation. Model should generate informative and appropriate responses by using a set of knowledge sentences. Here, response R-2 is preferred to response R-1 because it is grounded to the relevant knowledge sentence, K-1, and relevant to the context.

leverage conversational characteristics, i.e., *sequential structure*, *topic*, and *local context* of conversation. Here, the *sequential structure* means that a conversation is composed of a sequence of turns; *topic* means that there exists a central theme of a whole conversation, and *local context* is the utterances that precede immediately before a response. Differences from our works are that many recent works such as [31, 32] did not focus on the property *sequential structure* and merely model a multi-turn dialogue as a single document, and previous works such as [25, 33] focus on *local context* without considering the *topic* of the whole conversation.

Specifically, we develop two novel knowledge selection and ranking models and incorporate them into KGC models as shown in Figure 1.1 to generate responses based on the chosen knowledge. Then we adopt the knowledge ranking

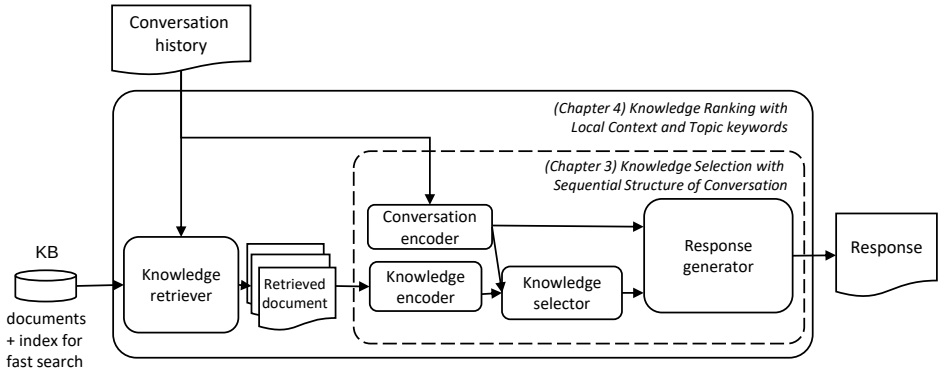

Figure 1.1: Overview of our KGC models

model into the quote recommendation as shown in Figure 1.2, where the quote can be considered a kind of knowledge. In Chapter 3, our setup is selecting a single knowledge document among candidate documents retrieved from an external retriever. In Chapter 4, our setup is retrieving and ranking top-k documents from a large KB for a given conversation history. We summarize the contributions of this dissertation as follows.

(Chapter 3) We focus on the conversation property that the conversation is composed of a sequence of turns to design knowledge selection methods in KGCs. Our knowledge selection methods aim to consider the turn order information to capture information relevant to the ground truth (GT) knowledge snippet. Specifically, we propose novel knowledge selection strategies, Match-Reduce and Reduce-Match, to apply text-matching techniques using token-level or sentence-level features of multi-turn KGC. Models based on Reduce-Match strategy first distill the whole dialogue context into a single vector with salient features preserved and then compare this context vector with the representation of knowledge sentences to predict a relevant knowledge sentence. Models based on Match-Reduce strategy first match every turn of the context with knowl-

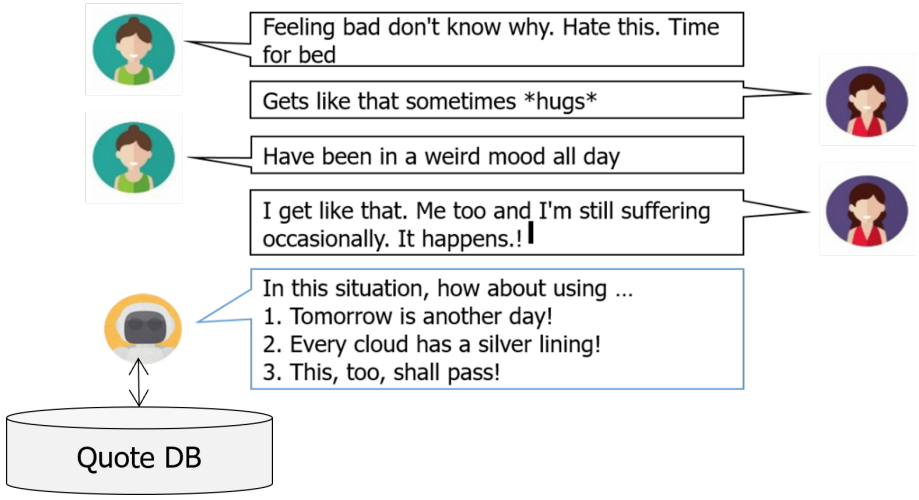

Figure 1.2: Quote recommendation system. Recommending quote, which including proverbs and (famous) statements of other people, can provide support, shed new perspective, and/or add humor to one's arguments in conversation. Knowledge selection/ranking method of KGC models can be utilized in the automatic quote recommendation system

edge sentences to capture fine-grained interactions and aggregate them while minimizing information loss to predict the knowledge sentence. Our extensive experiments show it can improve knowledge selection accuracy and generation performance over competitive baselines. The related publication to this topic is [34].

(Chapter 4) We focus on the conversation property that the interlocutors usually make responses relevant to both the topic of the whole conversation and the conversation turns before the response. From the KGCs on the internet, we first observe that both the conversation topic and the fixed number of tokens that come immediately before the response considerably affect responses in the conversation. Inspired by these properties, we propose a retrieval-augmented response generation model based on our novel knowledge ranking model that

retrieves a range of documents relevant to both the topic and the local context of a conversation. The retriever first accepts topic words extracted from the whole conversation and the tokens before the response to yield multiple representations and then compares the representations regarding the local context and salient tokens with the correspondents of the documents separately. For training, we introduce a new data weighting scheme to encourage the model to produce grounded responses without the GT knowledge snippet. Both automatic and human evaluation results with a large-scale dataset show that our models can generate more knowledgeable, diverse, and relevant responses compared to other state-of-the-art models. The related preprint to this topic is [35].

(Chapter 5) We study our proposed knowledge ranking model's applicability to a different kinds of knowledge. Recently, Lee et al. [36] proposed a quote recommendation system for conversation, which can help the user utilize the quote in the conversation platform. We view the task of recommending quotes as knowledge ranking, where the quote is a type of knowledge. In this topic, we aim to show that our knowledge ranking method can be extended to the challenging task where the words in quotes are metaphorical; thus, the meaning of the words is different from the words' of our language. To this end, we propose a novel quote recommendation framework to adopt the knowledge ranking model in Chapter 4 in the quote recommendation task. The framework consists of candidate generation, encoders for knowledge and context, and the place holder for knowledge ranking modules, trained simultaneously. The candidate generation model generates a list of candidate quotes recognized as suitable for the local context to provide them as hard negatives to the following re-ranker. Then, the re-ranker module produces the score of each quote in the KB via our knowledge selection method. The experiments with two conversation datasets

show that our proposed method can outperform the state-of-the-art baselines. The related publication to this topic is [37].

The remainder of this dissertation is organized as follows. Chapter 2 gives background knowledge, related works, evaluation methods, and problem statements. Chapter 3 introduces our knowledge selection strategies that exploit the sequential structure of a conversation. Chapter 4 proposes our knowledge retrieval-augmented KGC model exploiting topic keywords and the local context of a conversation. Chapter 5 presents our quote recommendation model that adopts knowledge re-ranker in Section 4.2.2 and experimental results. Finally, Chapter 6 concludes our work and presents limitations and future works.

Chapter 2

Background and Related Works

This chapter provides an overview of background knowledge and related works to understand the KGC models. The detailed survey of studies regarding KGC models is presented in [38]. We first elaborate on fundamental terminologies and discuss conversational systems to help understand the KGC task. Then we describe the primary components of a typical KGC model, the taxonomies of those components, and their role. Finally, we present the related works, evaluation methods, and problem statement of the KGC task.

2.1 Terminology

Definitions of essential terminologies used in the KGC task presented in Table 2.1. The definitions of the basic concepts, i.e., token, sentence, document, utterance, response, and conversation, are consistent with the literature regarding natural language processing. The basic units of a KGC dataset are a tuple conversation comprising context c, response r, and GT knowledge snippet k_{GT}, where sometimes the GT knowledge snippets are unavailable. We define a GT

Name	Notation	Meaning
token	w	a sequence of letters
sentence	s	a sequence of tokens that expresses a complete thought
text span	-	a token sequence
interlocutor / participant / user	-	a person who takes part in a conversation
utterance	u	a sequence of tokens in an interlocutor's turn
context / conversation history	c	a sequence of utterances prior to a response
response	r	utterances in a turn next to the given context
conversation	(c, r)	interactive communication between two or more interlocutors. a tuple of context and response
knowledge snippet/unit (sentence, document, or any text units)	k	a token sequence representing a knowledge piece. Depending on the dataset, this can be either a sentence, text span, or document.
knowledge base	$KB \ni k$	a external unstructured database composed of document
knowledge pool	$KP_c \subset KB$	a set of document selected from KB for a given context c
ground truth (GT) knowledge snippet	k_{GT}	documents which the interlocutor of a response refer to

Table 2.1: Basic terminology for KGC

knowledge snippet as document to which an interlocutor of response refers, where the meaning of term *referring* and the unit of knowledge can be different according to datasets. In addition, we define a knowledge snippet/document, the core concept of KGC, as a sequence of tokens to encompass the similar but slightly different forms of knowledge in the literature of KGC. We can formulate setups used in other works with our terminologies as our definitions are more generalized. For example, we can view the *chosen sentence* of Wizard-of-Wikipedia (WoW) [39] as a GT knowledge snippet, where the chosen sentence is a sentence on a Wikipedia page that the participant used when constructing his/her response. Differently from the WoW dataset, the GT knowledge

snippet of each response of the Conversing-by-Reading (CbR) dataset [40] can be a document. However, the GT knowledge snippet of this dataset is rare or missing because the dataset is an archive of conversations from the internet, where users of each conversation are not obliged to share the document.

2.2 Overview of Technologies for Conversational Systems

A KGC model can be considered as an open-domain DS that aims to produce more informative responses based on knowledge with no specific goal or task. This section overviews the related technology of conversational systems, including open-domain DS, task-oriented DS, and QA systems, and compares them with KGC models.

2.2.1 Open-domain Dialogue System

Open-domain DSs focus on conversations that do not have a specific task to complete. The system aims to converse with humans to increase long-term user engagement [41], where long-term user engagement is to retain the emotional bond with humans by responding like humans. Open-domain DSs can be categorized into rule-based, generation-based, ranking-based, and hybrid systems [42, 43], where the ranking-based, generation-based and hybrid systems are called data-driven systems. From the 1960s and 2000s, rule-based systems were actively studied, then from the 2010s, the data-driven systems had been gaining attraction from the research community.

Rule-based systems work based on a set of rules the developer manually built. The systems accept the natural language input, examine whether it matches one of the patterns, and output the responses according to the templates corresponding to the pattern. The rule-based systems could work well

only in limited situations due to the limited size of the rule set. The representative rule-based systems are Eliza [8] and Alice [9].

In 2010, Ritter et al. [44] adopted statistical translation models to open-domain DS. It was the first data-driven open-domain DS trained purely using data. Then data-driven approaches became the mainstream in the research community. Ranking-based systems rank the candidate responses from the human conversational dataset consisting of context-response pairs and yield the best candidate as a response. Currently, many studies proposed neural net-based ranking models, which use various matching methods of different architectures such as recurrent neural network (RNN) [45], convolutional neural network (CNN) [46], and Transformer [47]. Generation-based systems generate new responses appropriate to the input. The emergence of the Seq2Seq framework [14] opened a new era of adopting neural nets for building generation-based open-domain DS, but naive adoption of the Seq2Seq caused the issue of generating dull or not interesting responses [20].

Methods exploiting knowledge to produce informative or knowledgeable responses are proposed to deal with this issue. The methods utilize structured or unstructured knowledge. Zhou et al. [28] utilized large-scale commonsense KBs, and other numerous neural models [48, 49, 50] have been proposed to ground on domain-specific knowledge bases. Models using the structured KB enjoy the advantages of the structure of knowledge, e.g., connections between entities and exactness of the contents; however, they have limitations in extending the knowledge resources owing to the cost of building a large database. On the contrary, KGC models based on unstructured data, our research topic, have the advantage of utilizing rich text resources in the real world. We will discuss KGC models using unstructured knowledge in the rest of this dissertation.

2.2.2 Task-oriented Dialogue System

Task-oriented DS aims to help users complete their goals for specific tasks or domains via interacting with natural language. Representative examples of the task include flight booking, hotel reservation, customer service, and technical support. Task-oriented DSs have been applied in some real-world applications. A typical task-oriented DS is a modular system comprising submodules like language understanding, dialogue management, and language generation. These submodules are usually designed using either hand-crafted knowledge or are trained on task-specific data built manually. Some recent works studies [51, 52] attempt to design the task-oriented DSs with neural networks which can be trained end-to-end.

We present two aspects that are different from open-domain DSs.

- Task-oriented DS can be optimized by using the reward defined by the system designer. However, mathematically defining the goal of open-domain DSs is not easy because the goal can be subjective depending on the user.

- To build conversational systems that can converse about any domain or topic in the real world, the system designer should define the task schema[1] and build its labeled data for every task in the world. In contrast, open-domain DSs do not need such a pre-defined schema. Researchers usually attempt to train the open-domain DSs based on the human-to-human conversation corpus.

2.2.3 Question Answering System

QA system aims to output the correct answer for a given question in a natural language. Recent studies almost focus on open-domain QA where the domain

[1] A task schema is usually a data structure composed of user intents and slot-value pairs the system should gather from the user.

of questions is not restricted [53]. The open-domain QA systems can be categorized into retrieval-based (i.e., text-based in [53]), knowledge-based, and hybrid. Retrieval-based, knowledge-based, and hybrid systems utilize textual resources, a database of facts, and both, respectively. The retrieval-based system finds the document or passages, then yields the answer based on them using the techniques used in reading comprehension models. Many of the recent open-domain systems such as [54, 55] are neural net-based models. In contrast to the retrieval-based system, the knowledge-based approach transforms the input into a semantic representation and queries a database of facts. The hybrid approach retrieves pieces of evidence for the answer from both textual resources and databases and uses them to produce the answer. The representative hybrid system is IBM Watson [56], which won the first-place prize of $1 million against human champions on the TV show Jeopardy in 2011. We present two aspects of the QA systems that differ from the open-domain DSs or KGC model.

- The typical QA system assumes that the number of correct answers is one, whereas the number of possible responses in the open-domain DS is large. As Zhao et al. [1] pointed out, a single context can correspond to multiple responses in conversation, as shown in Figure 2.1.

- QA systems assume that its input is natural language text in the form of a question, whereas the KGC model assumes that the natural language input is utterances of various dialogue act [57]. This means that an ideal KGC should be able to deal with conversational utterances of either question or non-question types.

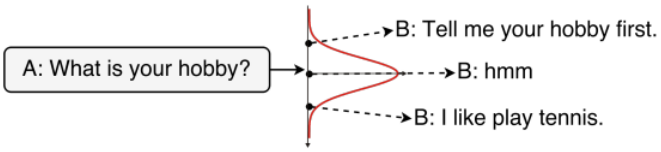

Figure 2.1: A figure from [1] representing the relationship between context and responses in conversations

2.3 Components of Knowledge Grounded Conversation Model

Figure 2.2 shows architecture components of a typical generative KGC model and the data flow of how the model processes the given conversation history. The architecture components comprise knowledge retrieval, conversation and knowledge encoding, knowledge selection, and response generation modules. Note that the modules are usually built by utilizing, modifying the structure, or sharing parameters of the underlying neural network we call backbone architecture, which is usually the Seq2Seq (S2S) model. In the following, we elaborate on the role of components and categorize each component utilizing the taxonomy shown in Table 2.2.

Components	Category
Backbone Architecture	RNN-based S2S
	Transformer-based S2S
Knowledge Retriever	Word matching-based
	Neural network-based
Conversation and Knowledge encoder	with/without pre-trained word embedding
	with explicit knowledge expansion
Knowledge selector	Soft selection
	Hard selection
Response generator w/ knowledge integration	with attention mechanism
	with copy mechanism

Table 2.2: Our taxonomy for each component of KGC model

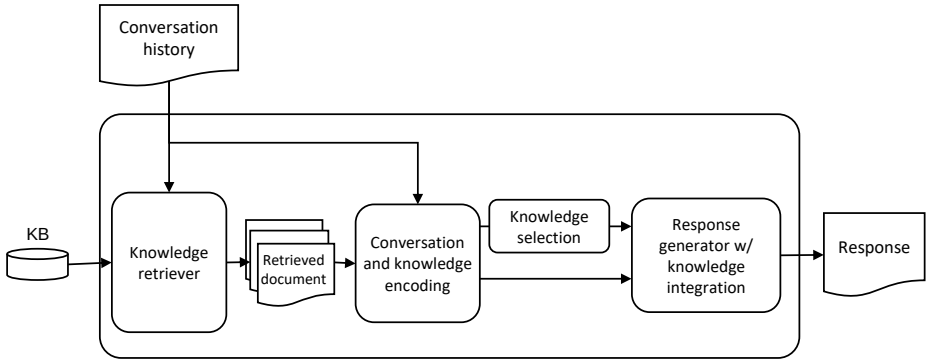

Figure 2.2: The architecture components of a typical KGC model and data flow

(Backbone architecture) KGC models generally are based on a S2S architecture [14, 58], where the encoder transforms tokens of variable-length to a sequence of real-valued vectors where each corresponds to the input token, and the decoder predicts each token of the response in an auto-regressive manner using the encoder's outputs. When the KGC task is proposed for the first time, RNN-based S2S architectures were widely adopted; after the effectiveness of the Transformer [16] architecture was proved, many researchers have adopted the Transformer architecture as a backbone.

(Knowledge retrieval) A KB has documents containing diverse knowledge. The knowledge retrieval module aims at fast retrieving relevant documents from the large KB. Specifically, it retrieves documents relevant to the conversation history and then gives them to the conversation and knowledge encoding module. It accepts the conversation history as input and outputs the search results by measuring similarities with the documents indexed in the KB. Because efficiency is the purpose of the retrieval module, selecting multiple documents coarsely relevant to the inputs is necessary. Word matching-based retrieval models measure the similarity based on Bag-of-Words (BoW) representation, so they

cannot retrieve the document semantically similar to the query that does not contain matching words. To deal with this limitation, neural retrieval models represent words with continuous vector representations. Such superiority of the neural retrieval model has been reported in [59].

(Conversation and knowledge encoding) The modules for conversation and knowledge encoding (conversation/knowledge encoder) capture the information carried in a word sequences of context and knowledge snippets and represents their semantics as real-valued vectors; that is transforming natural language into machine-understandable data. The result of this process is usually a sequence of real-valued vectors of input length (the number of input tokens). Then the following modules utilize the result and choose appropriate knowledge snippets for the context. Because KGCs are conversations, KGCs inherit the characteristics of natural language and conversations. We describe the important techniques that have been enjoyed in the advance of the techniques of NLP and conversation models.

The entire process of encoding starts with tokenization. The tokenization process divides the input sentences into tokens. If we use a word as a unit of the token divided by spaces then the number of parameters representing the tokens of vocabulary increases proportionally to the vocabulary size due to the emergence of new words out of the vocabulary. To this end, advanced tokenization methods such as BPE [60] or Wordpiece [61] divide a word into subtokens to represent a large-sized vocabulary with a small number of parameters. Then the tokens are represented with dense vectors. One of the most important concepts for this step is contextualization. Contextualization is to represent the tokens with vectors of their neighbors. The neural model such as RNN or Transformer contextualizes each token by modeling bidirectional interaction between word sequences. ELMO [62] and BERT [17] are representative methods for contex-

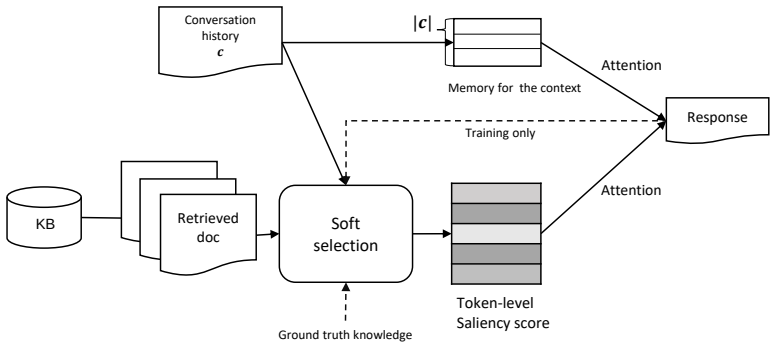

Figure 2.3: Overview of soft knowledge selection model. The dotted lines in valid only in testing.

tualization. Transformer-based representation methods have been popular due to their high performance in various NLP tasks.

(Knowledge selection) The knowledge selection module is a crucial component of the KGC model because the documents fetched from external textual resources provides the contents for the response. Most KGC models utilize the attention mechanism [15] and the memory network framework [63] to dynamically read the document memory built by the encoder and knowledge selection module. As classified by Ma et al. [38], we categorize knowledge selection methods into two groups: soft selection and hard selection, depending on the existence of a sampling mechanism that explicitly selects the most relevant knowledge snippet among candidates.

The soft selection-based methods aim to learn the continuous saliency score of knowledge tokens, and the scores are applied to the memory constructed by the knowledge encoder. Models using the soft knowledge selection method expect the result of soft selection to help attention from the decoder focus more on the relevant parts of the document. The hard selection-based methods aim

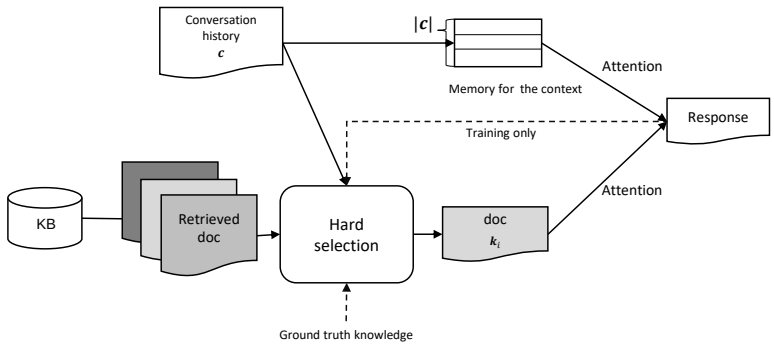

Figure 2.4: Overview of hard knowledge selection model. The dotted lines in valid only in testing.

to explicitly select some knowledge snippets (usually sentence) from candidate knowledge snippets where knowledge candidates are usually supplied from an external KB. If a GT knowledge snippet is provided, the model is trained with cross-entropy with the scores over the candidate knowledge snippets. The section 2.4 provides related works focused on the knowledge selection module.

(Response generation) The response generation module in KGC usually is based on the attention mechanism, which utilizes the decoder's hidden states as query and the encoder's hidden states as key and value. The primary method for injecting the knowledge into the response is to construct memory using the knowledge snippets chosen by the knowledge selection module and the conversation history and then give attentions to this memory.

The copy mechanism [64] learns to copy words from the source text by adding the probability of the words being copied to the generator's output token distribution. It has been widely used in generation-based summarization and dialogue models to deal with the out-of-vocabulary problem. Copy mechanisms also played an important role in KGC, as it makes models copy information

from the conversation context, external knowledge, or both.

2.4 Related Works

In this section, we introduce several known KGC datasets frequently used in the literature and summarize recent works related to our research topic. Table 2.3 summarizes KGC models introduced in this section with our models using the taxonomy presented in Table 2.2.

2.4.1 KGC datasets

Datasets using crowdsourcing use documents of various sources, including Wikipedia, movies, and news stories. Dinan et al. [39] collected KGCs between crowd-sourced workers, the sentences retrieved from an external retrieval system, and sentence-level GT knowledge snippets. Datasets such as CMU_DoG [65], or Holl-E [66] include conversations between crowd-sourced workers with specific documents, but they provide the GT knowledge snippets in the form of a text span. Datasets collected from the internet use the conversations of websites such as Reddit and Twitter[2]. Ghazvininejad et al. [26] compiled Twitter conversations that refer to Foursquare tips. Qin et al. [40] crawled conversations discussing a topic specified by web pages such as Wikipedia or news stories. They assume the documents relevant to a given conversation thread as GT knowledge snippets if the users do not explicitly give the GT knowledge snippet.

2.4.2 Soft Selection-based KGC Model

Qin et al. [40] proposed a model **CMR** based on a state-of-the-art machine reading comprehension model. The model builds a document memory to integrate information of the conversation context to the given document by using

[2] `https://twitter.com/` (Accessed: July 23rd, 2022)

cross-attention and self-attention. Then decoder generates a response while referring to the document memory via the attention mechanism. Ren et al. [67] proposed a model **GLKS** which utilizes a matrix representing matching between the context and the document. The model builds the matching matrix based on sequence length representations of the document and the context, then compresses it into one vector to use them with the decoder's state together for response generation. Tian et al. [68] proposed a teacher-student framework **RAM** to build a document memory that reflects the similarity with the response. The teacher is given the knowledge document, the context, and the ground-truth response, then builds a similarity weight vector (or matrix) between the response and document to apply it to document memory. The student learns to construct a document memory whose token saliency weights resemble the weights built by the teacher using the document and the context.

2.4.3 Hard Selection-based KGC Model

Dinan et al. [39] proposed the dataset of WoW and a model called **TMN**, which selects a knowledge sentence from a knowledge pool and generates a response based on the chosen knowledge sentence. Lian et al. [32] proposed a model **PostKS** that uses both prior and posterior distributions over knowledge sentences to select a knowledge sentence. Kim et al. [27] introduced a sequential latent variable model in which the latent variable indicates GT knowledge sentences to consider the conversation flow. The model **SKT** is trained by minimizing the KL divergence between prior and posterior probabilities for knowledge selection, where the prior and posterior probabilities are calculated by using the knowledge selection history encoded by gated recurrent units (GRUs) [69]. Conceptually, this process can be thought of as exploiting the evidence information in the last response encoded by posterior probability and transferring it

to infer the prior probability. **PIPM+KDBTS** [70] improved the SKT model by providing additional posterior information to the prior selection module for better approximating the posterior distribution. The posterior information is composed of a summary of the context history and a summary of knowledge candidates. Recently, Meng et al. [71] proposed a model **MIKe** that considers the initiatives in a conversation. The model discriminates each turn's initiative type (system or user initiative) and then calculates the knowledge selection probability by integrating its two knowledge selectors corresponding to each initiative. Zhao et al. [72] proposed **KnowledGPT** to apply large-scaled PLMs to the KGC tasks. They devised a knowledge selection module based on BERT [17] and LSTM [73], and formulated knowledge selection as a sequential prediction process. The model is trained on a dataset with GT knowledge snippets automatically built using a similarity score (unigram F1) between knowledge snippets and responses. Then, the knowledge selection and response generation modules are trained alternately through reinforcement learning and curriculum learning.

Our model proposed in Chapter 3 is a hard selection-based model. We explore similarity functions between context and knowledge by exploiting encoding context representation with turn sequence (ReduceMatch) or matching feature between turn and knowledge (MatchReduce). The difference between our model and the previous works [32, 39] is that their models consider the conversation as a single document. The difference between our model and the previous studies [27, 72] is that their models attempt to learn the mechanism of exploiting information in responses for selecting knowledge, while our model learns the relevance between context and knowledge using the GT knowledge snippet.

2.4.4 Retrieval-based KGC Models

Some recent studies have explored neural retrieval-based models [33, 74, 25]. Fan et al. [33] proposed to augment generative Transformer with two KNN-based information fetching **KIF** modules. The model uses the Wikipedia corpus and training utterances corpus to pull the content source of responses and the structure of other responses. The KIF modules learn to read to access two external knowledge sources, respectively. They showed that the proposed model could be extended to use a database of images. Shuster et al. [25] adopted retrieval-based models such as **RagToken** and **RagThenPoly**, which access a Wikipedia corpus directly and reduce the hallucination problem. Their work is based on the advanced techniques in the open-domain QA task. Another recent work [74] proposed a model called **RetGen** trained with a similar objective with RAG with different retriever models.

The difference between the above works and ours in Chapter 4 is that we devise a KGC model with our new re-ranker and data weighting scheme to cope with the properties (topic and local context) of KGCs in the wild, however they focused on the phenomena observed in crowd-sourced KGC datasets.

2.4.5 Response Generation with Knowledge Integration

Meng et al. [75] proposed **RefNet** which includes a decoder that integrates the probabilities of generating words in the vocabulary and copying the GT knowledge text span. Yavuz et al. [31] designed a KGC model **Deepcopy** whose decoder copies words from multiple knowledge sentences and the context, giving source word weights by using similarity between the decoder state and source's token-level and sentence-level hidden states. It calculates attention scores over the context and each knowledge sentence with the decoder's hidden state and combines them with token-level attention scores over each hidden state at each

token of the sources. Wang et al. [76] proposed an adaptive posterior knowledge selection method **AdaPKS** that helps the decoder select an appropriate token of knowledge consistent at every decoding step. The model sequentially computes information about which knowledge token should be used in decoding steps and selects tokens regarding some knowledge consistent with previously selected knowledge tokens.

Models	Backbone	Knowledge Retriever	Conversation and Knowledge encoder	Knowledge Selector	Response generator w/ Knowledge integration
CMR [40]	RNN-based (LSTM)	-	GLOVE	Soft	w/ attention
GLKS [67]	RNN-based (GRU)	-	GloVe	Soft	w/ copy mechanism
RAM [68]	RNN-based (LSTM)	-	GloVe	Soft	w/ attention
TMN [39]	Transformer-based	word matching	Pretrained w/ custom corpus	Hard	w/ attention
PostKS [32]	RNN-based (GRU)	word matching	GloVe	Hard	w/ attention
SKT [27]	Transformer-based	word matching	BERT	Hard	w/ copy mechanism
PIPM+KDBTS [70]	Transformer-based	word matching	BERT	Hard	w/ copy mechanism
MIKe [71]	Transformer-based	word matching	BERT	Hard	w/ copy mechanism
KnowledGPT [72]	Transformer-based	word matching	PTM (GPT-2/BERT)	Hard	w/ attention
KIF [33]	Transformer-based	neural network	PTM (Transformer)	Soft	w/ attention
RetGen [74]	Transformer-based	neural network	PTM (initialized w/ GPT-2)	Soft	w/ attention
RagToken [25]	Transformer-based	neural network	BART	Soft	w/ attention
RagThenPoly [25]	Transformer-based	neural network	BART	Soft	w/ attention
RefNet [75]	RNN-based (LSTM)	-	w/o pre-training	Soft	w/ copy mechanism
Deepcopy [31]	RNN-based (LSTM)	-	w/o pre-training	Soft	w/ copy mechanism
AdaPKS [76]	RNN-based (GRU)	-	w/o pre-training	Soft	w/ copy mechanism
Our model in Chapter 3 [34]	Transformer-based	word matching	w/o pre-training	Hard	w/ attention
Our model in Chapter 4 [35]	Transformer-based	neural network	BART	Soft	w/ attention

Table 2.3: Model comparison with our taxonomy. PTM stands for pre-trained model. BART [3] is a Transformer based Seq2seq model pre-trained by corrupting text with noising functions and learning to reconstruct the original text.

2.4.6 Quote Recommendation

Citation Recommendation One of the most related tasks to the quote recommendation is citation recommendation for academic articles [77, 78] which recommends relevant reference articles for academic writing. Specifically, the citation recommendation task aims to recommend the top-k relevant articles for texts that appear before and after a citation within a certain fixed length. One can exploit rich information of articles, such as title, abstract, full text, and venue. A notable difference with quote recommendation is that the model for the quote recommendation cannot use such various information. In addition, quote recommendation systems can use only the context before the quote as input because utterances next to the quote are usually unavailable during the conversation.

For citation recommendation, some works attempt to bridge the language gap between cited papers and the body of text where the citation is needed. Shaparenko and Joachims [79] proposed to use language models to consider the relevance of the context of the citation and text of the paper. Huang et al. [78] used paper's unique IDs to represent the candidate papers and utilized the translation model to estimate the conditional probability of the paper's ID given the context. Tan et al. [80] proposed a neural network approach that learns the distributed representations for each context and quote, respectively, and measures the relevance of the context and quotes using those representations.

Quote Recommendation in Conversation We categorize models for recommending quotes in conversation into ranking-based and generation-based models. The ranking-based models output a ranked list of quotes for a given conversation. Lee et al. [36] first proposed to recommend quotes for conversation by using a neural network. They focused on extracting features solely from

the conversation by combining a convolutional neural network and recurrent neural network to rank the candidate of quotes. Tan et al. [81] attempt to enrich the words in quotes with meta information such as tags and authors of the quotes. Recently, Wang et al. [82] introduced a transformation matrix that maps a query representation directly to a quote representation and a mapping loss that minimizes the distance between semantic spaces of quote and conversation. The generation-based models generate a quote rather than selecting the quote from a database. After generation, the models need post-processing, which queries the quote database where term-matching similarity is needed. Wang et al. [83] proposed an encoder-decoder framework for generating quotations in conversation. Its encoder encodes interaction information of turns (for both turn and earlier history) and latent topics in contexts; then, its decoder generates the quote using the attention mechanism auto-regressively.

Our model proposed in Chapter 5 is a ranking-based model. The most similar model to ours is the model proposed in [82]. Our model differs from the model [82] in the following two aspects. First, our model calculates the score for the quote based on the matching using multiple vector representations from inputs and conducts a fine-grained level matching, while the model [82] uses a single vector representation of the input. Second, our model considers the relationship between words in different turns because we encode the conversation using only a single Transformer encoder, allowing self-attention between tokens of different turns. However, the model [82] did not consider the relationship between words in different turns, i.e., it represents the conversation history with a single vector using a combination of Transformer and RNN.

2.5 Evaluation Methods

It is challenging to evaluate the text generated by models. Especially evaluation for open-ended text generation tasks such as conversational response generation is so because multiple plausible answers may exist. Thus, it is not easy to standardize the evaluation method. For example, Venkatesh et al. [84] also mentioned that "In short, there is no standard evaluation model for NLG, nor agreement in terminology, and explanatory details for the criteria are often lacking." Evaluating responses generated by KGC models is conducted typically through automatic and human evaluations.

For automatic evaluation, many existing works used word matching-based metrics in machine translation or the embedding-based method [38]. The word matching-based method more helps us determine whether or not a specific word is used than embedding based; as a result, it can evaluate the model's capability to use technical terms or named entities from external documents. In KGC tasks such as WoW [39] task and the grounded response generation task at DSTC [4], similarity with human-generated responses was measured using word matching-based similarities such as Unigram F1 or BLEU score. When a single reference response is used for evaluation, the absolute value of the metric can be very low due to the word mismatch problem, so it is not easy to get reliable results. To alleviate this issue, several human responses collected are used as reference responses. However, even if we use multiple reference responses for evaluation, the absolute value of the measurements can still be low because the number of possible responses is very large.

The human evaluation aims to evaluate subjective qualities that the automatic evaluation metric cannot cover. Table 2.4 shows the evaluation metrics used in the human evaluation of the presented responses that are judged as

good or bad. We confirm that some of the existing works evaluate the KGC models by using multiple questions that a single question integrates several metrics, e.g., "Is the response informative and interesting?" as shown in [40]. However, if several metrics are mixed and evaluated, it may be difficult to grasp the characteristics of the systems' responses and find their detailed deficiency.

Metrics	Definition	Example (**A:** Context, **G:** Good Response, **B:** Bad Response, K: External doc)	Why is G's response better than B's?
Relevance	How much is the response closely connected to context?	**A:** Do you have a dog? **G:** No, I have a cat. **B:** I am a workaholic. I like to work days and nights.	Semantic distance between context and G's response is closer than context and B's.
Appropriateness	How much is the response suitable to the context?	**A:** Do you know where the toilet is? **G1.**: Over there. **G2.**: I don't know. **B:** The toilet near here is clean.	Responses of G1 and G2 are appropriate to the context because they deliver relevant information to A. (G1's: the location, G2's: the information that G2 does not know) But Response of B is inappropriate since it provides information that does not match A's intention.
Informativeness	Does the response add new and useful information to the conversation?	**A:** Sonny scored against Chelsea yesterday. **G:** So Tottenham won the game. **B:** Yeah, Sonny did.	G's response is more informative than B's, since it gives new information about Tottenham's winning, whereas B's response rephrases A's utterance.
Interestingness	How much does the response give your attention because it seems unusual or exciting or provides the information you did not know about?	**A:** Sonny scored five goals against Chelsea yesterday. **G:** Surprisingly, Tottenham lost the game. **B:** So Tottenham won the game.	It is usual that the team which scored five goals lose. So G's response is more interesting than B's.
Groundness	How much is the response relevant to the given document?	**A:** Do you like Sonny? **G:** Yes, Sonny is the best player in EPL, He won the MVP three times in '20. **B:** Yes, son is a young soccer player. **K.** Son Heung-min won the Best player award in '20 EPL. And he also won MVP two times in '20.	Response of G is more relevant to K than Response of B, though some information is factually incorrect.
Factual correctness	Can it be inferred by knowledge of the world?	**B:** Cho Kyunghyun was the co-chair of NIPS '14 and ICML '15 **K.** (No evidence in the knowledge of the world)	We cannot infer that 'Cho Kyunghyun was the co-chair of NIPS '14 and ICML '15 since there is no evidence in the knowledge of the world.
Coherence within a response	Does it make sense in and of itself?	**B:** I like dogs but don't like dog.	It is a self-contradiction.

Table 2.4: Core metrics used for evaluating KGC systems

2.6 Problem Statements

We assume that a KB is composed of many knowledge snippets, and more than or equal to two persons have a conversation on the topic related to some of the knowledge snippets from that KB. As the conversation progresses, the topic may shift; consequently, the knowledge snippets relevant to the context may also change. Our goal is to generate an informative response suited to a given conversation history by developing the module for selecting or ranking knowledge. We formally define two subtasks:

- **Knowledge selection/ranking task** - Given a conversation history $c = (u_i, \cdots, u_{(i-1)+M})$, predict an (or multiple) appropriate knowledge snippet(s) k(s) from candidate knowledge snippets K_c.
- **Response generation task** - Given a conversation history $c = (u_i, \cdots, u_{(i-1)+M})$ and the chosen knowledge snippet(s), k(s), generate response r.

where u and r represent single or multiple utterance(s) in a conversation turn, respectively, M is the length of the part of a long conversation session, and K_c is a set of candidate knowledge snippets of the context c, which the responding participant can refer to. Depending on the setup of the given task, the candidates K_c can be either a small set of knowledge snippets relevant to the context c, KP_c, or the entire KB, KB.

Problem definitions in the following chapters are slightly different. In Chapter 3, the problem definition is to output a response r for a given context c based on a sentence chosen from a K_c retrieved by an external retriever. In Chapter 4, the problem definition is to output a response r for a given c by retrieving documents from KB directly. In Chapter 5, the problem is to yield the top-k knowledge snippets ks (quotes) fit to the context c.

Chapter 3

Knowledge Selection with Sequential Structure of Conversation

3.1 Motivation

A conversation session in the KGC task is composed of a sequence of turns, where the participants in the conversation may request information either explicitly or implicitly. Specific words or phrases bearing the (implicit) information needs in the conversation history can correspond to the words in specific knowledge sentences. In this chapter, we introduce our two knowledge selection strategies: 1) Reduce-Match and 2) Match-Reduce to match such information in the context with knowledge candidates. Furthermore, we explore several knowledge selection methods based on these strategies by using the same neural KGC model. To capture semantic information triggering the external knowledge in the turn sequence, the model should collect effectively important matching features and transfer them to the next layer to compute the knowledge selection

loss.

In recent studies, neural network-based models have been widely adopted for KGC [26, 32, 85, 40, 86, 87, 48, 75, 39]. As we mentioned in Chapter 2, KGC models generate response using the results of conversation encoder, knowledge encoder, and knowledge selector. The conversation encoder models the given conversation history to keep track of the current conversation flow. The knowledge encoder provides a base source to the response generator. Many of the relevant studies [26, 31, 66, 85, 40, 75, 87, 86] focus on modeling knowledge sentences collectively to generate responses. However, the knowledge selector is crucial in rendering the response informative, which may change the flow of the conversation. Depending on which sentence is chosen by the knowledge selector, the topic of the conversation may be expanded or changed. As shown in Table 1.1, selecting knowledge sentence K-3 will result in a response less coherent than the one based on sentence K-1; sentence K-3 mentions about the diversity of population in the U.S., rather than a fact about "New York." Despite the importance of this selection, only few studies [39, 32] have investigated this issue. Dinan et al. [39] and Lian et al. [32] proposed models that focus on choosing the correct knowledge sentence. However, they do not focus on how to model the context in a multi-turn conversation setting, where they merely model a multi-turn conversation as a single document.

3.2 Reduce-Match Strategy & Match-Reduce Strategy

Depending on when matching with a knowledge sentence is performed, we define two strategies, Reduce-Match and Match-Reduce. The abstracted steps of each strategy are as follows:

- **Reduce-Match**:

Embedding → Aggregation → Matching → Scoring

- **Match-Reduce**:

  Embedding → Matching → Aggregation → Scoring

Knowledge selection models based on Reduce-Match strategy first distill the whole dialogue context into a single vector, with the salient features preserved, and then compare the vector of context with the representation of a knowledge sentence to predict a relevant knowledge sentence. On the contrary, models based on Match-Reduce strategy first match every turn of the context with knowledge sentences to capture local or fine-grained interactions and transfer them to the aggregation step.

In both these strategies, word tokens in the context and knowledge sentences are transformed into real-valued vectors during the embedding step, and a score list for the knowledge sentences is computed at the scoring step. Roles of the aggregation and matching steps differ according to the knowledge selection strategy. From an abstract-level point of view, the aggregation step merges multiple outputs from the previous step into a single output such as a fixed-length vector, and the matching step compares between two different types of inputs.

Our two strategies have been extensively adopted in the tasks of text-matching between two sentences [88, 89, 90, 91, 92] and answer (or response) selection in conversation [93]. However, our work has some notable differences with the studies on two tasks. First, the primary concern of text-matching is a comparison between two sentences, whereas our work focuses on matching between the sequence of sentences and the other one. Second, properties of our task are different from the answer selection task's, where 1) the knowledge selection task is a subtask of another task, i.e., response generation, and

2) the participant who has access to the external document is encouraged to select a novel or interesting knowledge sentence. These two reasons motivate us to pay much attention to extensively explore the matching function and aggregation function, including simple, efficient, or sophisticated ones under the same framework to cope with the unique challenges of the KGC. To the best of our knowledge, our study is the first attempt to explore various text-matching methods extensively for KGC.

3.2.1 Backbone architecture

We choose an encoder-decoder model attached with a knowledge selection module [39], shown in Figure 3.1, as our framework. We will explore various knowledge selection models on top of the framework in the following sections. Our framework is built on Transformer [94], which itself and its variants have provided high performance in understanding or generating natural language [94, 17, 95]. We concatenate all the words in the context c and encode them into a sequence of vectors by using Transformer encoder. We also encode each knowledge sentence in K_c into another sequence of vectors by using the same encoder. Then, the knowledge selection module selects a knowledge sentence by comparing the context and the candidate knowledge sentences. Finally, the decoder on receiving the encoded context and a selected knowledge sentence, generates a response. The decoder uses the knowledge sentence chosen by the participant (GT knowledge sentence), k, during training, and, k_{pred}, the knowledge sentence predicted by the trained knowledge selection module during inference. We train the model in an end-to-end manner using the loss [39] defined as:

$$\mathcal{L} = -(1-\lambda)\log P(r|c,k) - \lambda \log P(k|c) \tag{3.1}$$

where λ is a hyperparameter between 0 to 1.

Note that our framework uses a GT knowledge sentence instead of a predicted one in training. If we feed a predicted knowledge sentence, which may be inaccurate, to the decoder in training, the decoder can be trained to less transfer information in knowledge sentence to the decoder or ignore the selected knowledge sentence. Thus, this will not be suitable for our goal that investigates the effect of knowledge selection on response generation. Also, we empirically confirmed that our training scheme outperforms others, including using a predicted knowledge sentence with or without Gumbel-softmax [96], in terms of knowledge selection accuracy and response generation metrics. For the above two reasons, we adopt the learning method shown in Figure 3.1.

3.2.2 Reduce-Match Strategy-based Models

Reduce-match strategy transforms the vectors of utterances into a vector preserved with salient features in a multi-turn dialogue and matches it with a knowledge vector. This strategy can be computationally efficient compared to the Match-Reduce strategy that matches every turn of the context before aggregation. This strategy at the aggregation step, aims to condense all the multi-granularity features i.e., from word- to turn-level, which makes fine-grained matching difficult. Here, it is essential for the aggregation step to filter the important information from the irrelevant one.

During the aggregation step, we use the Universal Sentence Encoder based on Transformer (USE_T) [97] to aggregate the context or knowledge sentence into a single vector. The USE_T computes the element-wise sum of the vectors from the Transformer's encoder and divides it by the square root of the length of the input text. Here, we do not use any pre-trained model of Transformer for simplicity. Although this method is simple, it has the capability of distinguishing a relevant sentence from irrelevant ones thanks to the effectiveness of

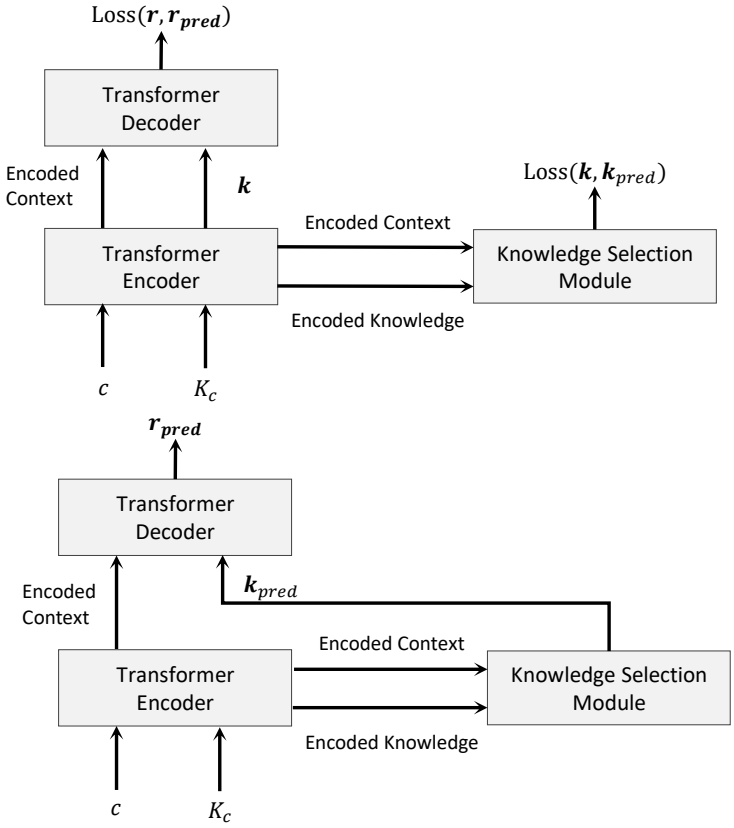

Figure 3.1: Our Transformer-based encoder-decoder framework for KGC (Top) Training phase. (Bottom) Testing phase.

the Transformer encoder [97]. The Transformer associates each output vector with related words within the same context by representing it with a weighted sum of other word embeddings, which is called self-attention. Therefore, the semantics in the text sequence such as phrases or proposition in a multi-turn dialogue can be preserved in the final representation. We compare the following two discourse-level aggregation methods to obtain a fixed-length context vector $\mathbf{c} \in \mathbb{R}^d$ as follows:

- Average aggregation: As shown in Figure 3.2 (top), this method aggre-

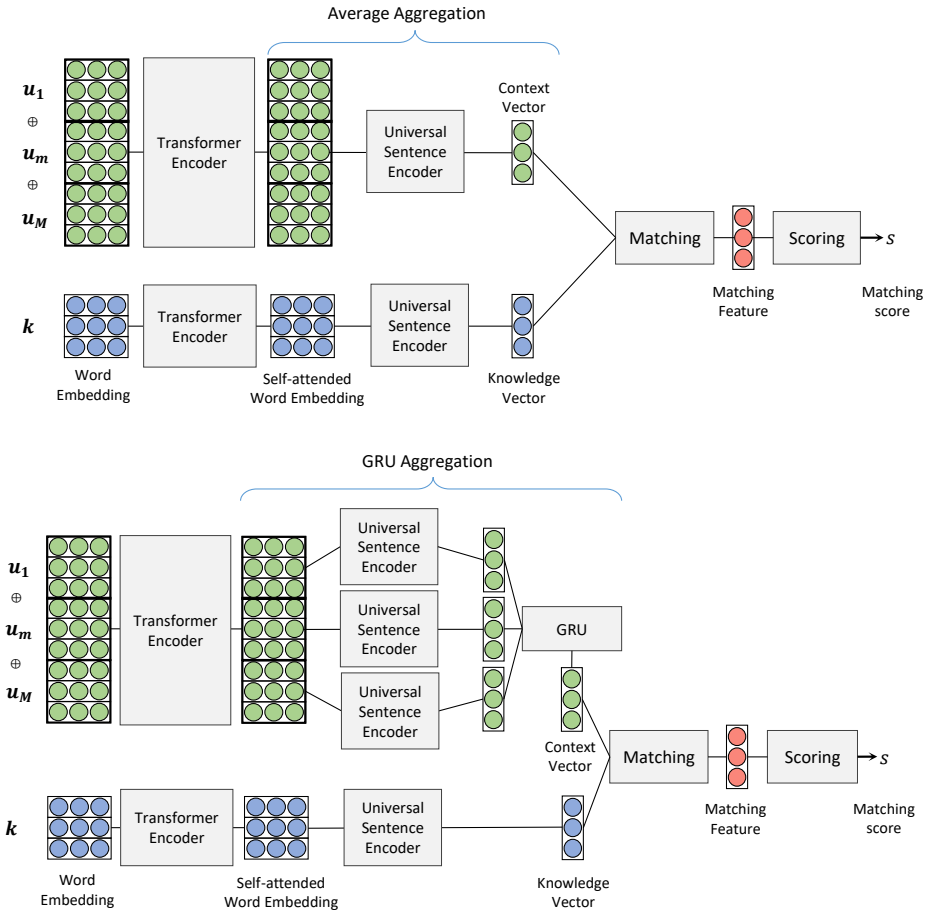

Figure 3.2: Reduce-Match strategy. (Top) Reduce-Match with average aggregation. (Bottom) Reduce-Match with GRU aggregation

gates the output vectors of all the words encoded by the Transformer by using USE_T.

- GRU aggregation: As shown in Figure 3.2 (bottom), this method aggregates the turn embeddings by using GRU [98], where the turn embeddings are computed by USE_T. The final hidden state of the GRU after reading the sequence of turn representation is used to express the context,

which we denote as $GRU(x_0, \cdots x_n)$. We expect the GRU to be trained to extract important features from a sequence of turns.

While aggregating the context, we encode the knowledge sentence to a vector $\mathbf{k} \in \mathbb{R}^d$ by using the Transformer encoder used incontext encoding followed by USE_T.

During the matching step, we explore various matching functions between the aggregated context vector and knowledge vector. We examine the following four matching functions of a context vector $\mathbf{c}$ and the knowledge vector $\mathbf{k}$.

- Bilinear function

$$p_{\text{bil}}(\mathbf{c}, \mathbf{k}) = \sigma(\mathbf{c}^T W_b \mathbf{k}) \tag{3.2}$$

- Cosine similarity

$$p_{\cos}(\mathbf{c}, \mathbf{k}) = \cos(\mathbf{c}, \mathbf{k}) \tag{3.3}$$

- Multi-head dot product [94]

$$p_{\text{mh}}(\mathbf{c}, \mathbf{k}) = W_m([head_1; \cdots ; head_h]) + b \tag{3.4}$$

$$head_i = \frac{W_i^C \mathbf{c} (W_i^K \mathbf{k})^T}{\sqrt{d_h}} \tag{3.5}$$

- Dimension-wise features-based matching [99]

$$p_{\text{dw}}(\mathbf{c}, \mathbf{k}) = \mathit{FeedFwd}([\mathbf{c}; \mathbf{k}; \mathbf{c} \odot \mathbf{k}; |\mathbf{c} - \mathbf{k}|]) \tag{3.6}$$

where $W_b \in \mathbb{R}^{d \times d}$, $W_i^C \in \mathbb{R}^{d_h \times d}$, $W_i^K \in \mathbb{R}^{d_h \times d}$, $W_m \in \mathbb{R}^{h \times 1}$ and $b \in \mathbb{R}$ are trainable parameters. σ, *FeedFwd*, and h represent a sigmoid function, a feedforward layer with $tanh$ activation and output size of 1, and the number of heads, respectively. Multi-head dot product matching compares two representations of the

aggregated context and the knowledge sentence in different subspaces. This idea was implemented in a different manner in several NLP tasks [94, 100, 101, 102]. In this study, we apply the one used in the attention module proposed in [94]. Dimension-wise features-based matching [99] is commonly used in many models such as [103, 104] for the task of natural language inference. The output of the matching step is scalar; therefore, we use a scoring layer which yields the input scalar value as it is.

It should be noted that Dinan et al. [39]'s E2E Transformer MemNet model employs the Reduce-Match strategy with average aggregation and dot product matching. We compare the experimental result of this configuration with our other configurations.

3.2.3 Match-Reduce Strategy-based Models

As aforementioned, a disadvantage of the Reduce-Match strategy is that the aggregation step has a bottleneck of extracting multi-level granularity features in the dialogue context for matching. To overcome this problem, the Match-Reduce strategy first extracts the interactions between a knowledge unit and context at the turn-level, and then tries to preserve this information as much as possible in the aggregation step. Furthermore, executing the matching step earlier than aggregation not only enables capturing the matching information between sentences but also additional interactions such as discourse-level patterns, e.g., flow of conversation. Now, we introduce shallow matching-based networks and three of our implementations of remarkable models in response selection for retrieval-based chatbot. We choose the models according to their representativeness and performance in the given topic. We will detail our models and the justifications of our choices. We use the Transformer encoder shown in Figure 3.3 for all the models to ensure fair comparisons. Furthermore, we use

a feed-forward network for knowledge scoring, whose input is the result of the aggregation step and output is a real-valued score.

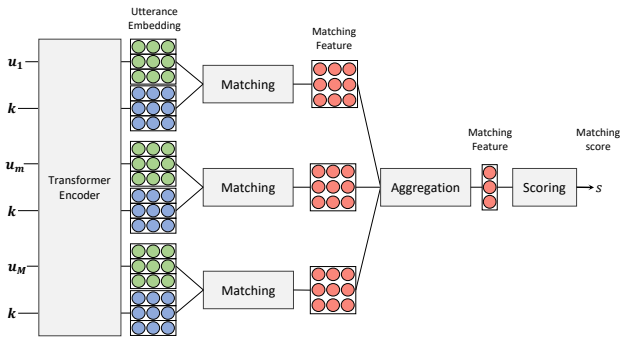

Figure 3.3: Match-Reduce strategy

Shallow Matching Networks

We first explore various simple but efficient matching functions to find effective methods for the Match-Reduce strategy. The shallow matching-based networks first encode every turn and the knowledge unit by using USE_T. Then the matching step extracts matching features between representations of the i-th utterance u_i in context and a knowledge unit k. We compare five matching functions that are defined as follows.

$$q^i_{\text{dot}} = \mathbf{u}_i^T \cdot \mathbf{k} \tag{3.7}$$

$$q^i_{\text{bil}} = \sigma(\mathbf{u}_i^T W_b \mathbf{k}) \tag{3.8}$$

$$q^i_{\text{cos}} = \cos(\mathbf{u}_i, \mathbf{k}) \tag{3.9}$$

$$q^i_{\text{mh}} = W_m([head_1; \cdots; head_h]) + b \tag{3.10}$$

$$q^i_{\text{dw}} = FeedFwd([\mathbf{u}_i; \mathbf{k}; \mathbf{u}_i \odot \mathbf{k}; |\mathbf{u}_i - \mathbf{k}|]) \tag{3.11}$$

where $\mathbf{u}_i \in \mathbb{R}^d$ is a vector of the ith turn encoded by the universal encoder. We use a GRU for aggregating the matching features. The matching result, q^i, of each turns is processed by the function, GRU.

Sequential Matching Network

Recently, Wu et al. [105] proposed a Sequential Matching Network (SMN) that can effectively extract word-/segment- level matching between context and a response, thanks to the CNN. We follow Wu et al. [105]'s suggestion in our implementation, except that we replace its utterance encoder with the Transformer encoder. The matching step extracts interaction features between utterances and a response at the word and segment levels. Subsequently, aggregation step accumulates the matching information from the matching step by using a CNN and GRU.

In this strategy, the matching function between context c and knowledge unit k, $f(c,k)$ is defined as follows [105].

$$M_1 = U_w^T K_w \tag{3.12}$$

$$M_2 = U^T K \tag{3.13}$$

$$f_u(u,k) = W[CNN(M_1); CNN(M_2)] + b \tag{3.14}$$

$$f(c,k) = FeedFwd(GRU([f_u(u_i,k), \cdots, f_u(u_{i+M},k)])) \tag{3.15}$$

where $U_w \in \mathbb{R}^{d \times T}$ and $K_w \in \mathbb{R}^{d \times T_k}$ are utterance word embeddings and word embeddings of knowledge sentence, respectively. $U \in \mathbb{R}^{d \times T}$ and $K \in \mathbb{R}^{d \times T_k}$ are word-level vectors of an utterance and a knowledge sentence yielded by the Transformer encoder, respectively. $CNN(\cdot)$ is a function composed of 2D convolution layers and the 2D pooling layesr as proposed in [105].

We choose SMN as one of the models of the Match-Reduce strategy rather than Sequential Attention Network (SAN) [105] that uses matching features as attention scores between words in each turn by words in response using GRU. SMN is more efficient and easier to parallelize than SAN, thanks to the CNN that is used for matching. Moreover, the matching functions of both do not

exhibit clear difference in terms of efficacy.

Deep Attentive Matching Network

In a turn-to-knowledge matching, matching can not only occur at fine-grained level (e.g., word or phrase) but also at coarse-grained level (e.g., sentence or topic). To handle this behavior, we can exploit the internal representations from the layers of the text encoder, where the encoder gradually learns from the fine-grained features to the coarse-grained level feature. To realize this concept, we chose the Deep Attentive Matching (DAM) [106] that provides high performance in the response selection task in multi-turn dialogue. The only difference between our model and the original one is that we use the multi-head attention module [94] instead of the attentive module to maximize the matching capability.

Our DAM model first constructs representations of text segments at different granularity with stacked self-attention and then calculates the matching matrices, $M_{self}^{u_i,k,l}$ and $M_{cross}^{u_i,k,l}$, between the context and the knowledge unit as follows.

$$M_{self}^{u_i,k,l} = U_i^{lT} K^l \tag{3.16}$$

$$M_{cross}^{u_i,r,l} = \tilde{U}_i^{lT} \tilde{K}^l \tag{3.17}$$

where

$$U_i^{l+1} = MHAttentionModule(U_i^l, U_i^l, U_i^l) \tag{3.18}$$

$$K^{l+1} = MHAttentionModule(K^l, K^l, K^l) \tag{3.19}$$

$$\tilde{U}_i^l = MHAttentionModule(U_i^l, K_i^l, K_i^l) \tag{3.20}$$

$$\tilde{K}_i^l = MHAttentionModule(K_i^l, U_i^l, U_i^l). \tag{3.21}$$

MHAttentionModule represents a module that is same as that of a layer of the Transformer encoder [94] and X_i^l denotes the output of lth encoder layer after encoding X_i. Then DAM concatenates the two matching matrices of each utterance and a knowledge unit into a 3D matching tensor, Q, that is defined as:

$$Q = [M_{self}^{u_i,k,l}; M_{cross}^{u_i,k,l}] \tag{3.22}$$

Subsequently, it leverages a two-layered 3D convolution with max-pooling operations to retain important matching features from the tensor.

Knowledge Enhanced Hybrid Neural Network

The aforementioned three models attempt to maximally extract matching features from all of the semantic units. However, this may cause a problem; the majority of information in context might be irrelevant to the correct knowledge sentence.

To mitigate this problem, we notice the two recent studies: Knowledge Enhanced Hybrid Neural Network (KEHNN) [107] and Multi-hop Selector Network (MSN) [108]. KEHNN leverages prior knowledge, such as key phrases tagged in advance, to identify useful information in the dialogue context and performs matching with three interaction matrices. It fuses the prior knowledge into word representations by the so-called knowledge gates and establishes a new interaction matrix. On the contrary, MSN explicitly selects relevant utterances in the dialogue history by comparing them with the immediate previous utterance (message) and matches the selected utterances with the candidate response. Interestingly, though MSN presently provides state-of-the-art performance in three public response-selection datasets, it shows inferior performance in our task[1].

[1]We conducted experiments with the author's codes available at `https://github.com/`

Consequently, we choose KEHNN to tackle the problem of filtering irrelevant information during matching. KEHNN builds knowledge enhanced representation $\tilde{\mathbf{e}}$ for each word by using knowledge gate $\mathbf{k}_w$ that is defined as:

$$\tilde{\mathbf{e}}_w = \mathbf{k}_w \odot \mathbf{e}_w + (1 - \mathbf{k}_w) \odot \mathbf{k}_x \tag{3.23}$$

$$\mathbf{k}_w = \sigma(W_k \mathbf{e}_w + W_k \mathbf{k}_{u_i}) \tag{3.24}$$

where $\mathbf{e}_w \in \mathbb{R}^d$ is the embedding of a word w in text S_x, $\mathbf{k}_x \in \mathbb{R}^n$ is the representation of the prior knowledge of text S_x, σ is a sigmoid function, and $W_w \in \mathbb{R}^{d \times d}$ and $W_k \in \mathbb{R}^{d \times n}$ are learnable parameters.

In our implementation, text S_x can be a knowledge sentence or utterances in a dialogue context. To acquire prior knowledge vector $\mathbf{k}_x$, we train the Biterm Topic Model [109] on the dialogue and knowledge corpus that is used for our response generation task with 200 topics and assign a topic to each utterance and a response by using the inference algorithm. Finally, we transform topic keywords to a vector by averaging the embeddings of the top 20 words under the topic.

After constructing knowledge enhanced representation of each words in the context and knowledge sentences, word-by-word matrix between knowledge enhance representations of utterances and a knowledge unit are computed (3.27). Then, *CNN* extract important features from this matrix and is concatenated to the other features (3.28). Therefore the matching function $f(c, k)$ of KEHNN

`chunyuanY/Dialogue` (Accessed: Jul. 24, 2019) in our main dataset WoW. We suspect the reason is that the utterance in the final turn may be frequently semantically more different to the previous dialogue history than general multi-turn dialogue, which may not be helpful to find the information required to find an irrelevant utterance.

is defined as follows.

$$M_1 = U_w^T K_w \tag{3.25}$$

$$M_2 = U^T K \tag{3.26}$$

$$M_3 = \tilde{E_u}^T \tilde{E}_k \tag{3.27}$$

$$f_u(u,k) = W[CNN(M_1); CNN(M_2); CNN(M_3)] + b \tag{3.28}$$

$$f(c,k) = FeedFwd(GRU([f_u(u_i,k), \cdots, f_u(u_{i+M},k)])) \tag{3.29}$$

where $\tilde{E_u} \in \mathbb{R}^{d \times T}$ and $\tilde{E_k} \in \mathbb{R}^{d \times T_k}$ are knowledge enhanced representations of an utterance and a knowledge unit respectively. We implement the aggregation step of KEHNN similar to that of SMN's.

3.3 Experiments

We conduct several experiments to show the performance of proposed knowledge selection methods and compare ours with other existing models. We provide a case study and analysis results in three dimensions, which include matching difficulty, length of context, and dialogue acts.

3.3.1 Experimental Setup

Our goal includes the knowledge selection and response generation task. We perform an automatic evaluation for knowledge selection with a metric used in information retrieval and conduct both automatic and human evaluation for response generation due to the ambiguity of natural language.

Datasets We evaluate our approach on two benchmarks, the WoW[2] collected in [39] and CMU_DoG [65], which are commonly used in the related works such as [27, 32, 110]. To the best of our knowledge, the WoW is the only large dataset

[2] `http://parl.ai/projects/wizard_of_wikipedia/` (Accessed: June. 4, 2019)

that provides the GT knowledge sentence for KGC. The CMU_DoG has different properties, including natural language phenomena in the real world, such as informal language patterns and grammatical errors. For computational efficiency, we set the maximum length of context to M and the maximum number tokens of a turn to 4 and 64, respectively. The example that we define for training comprises a dialogue snippet that ends before the turn of the participant with knowledge. All our models employ BPE encoding [60] that is known to be effective in a large vocabulary corpora.

WoW dataset is collected from crowd-sourced workers' conversations grounded on related Wikipedia pages. We use its random split, which comprises 18,430 dialogues for training 1,948 for validation, and 965 for test. The average number of turns in a dialogue session is 9.0. This dataset provides ground-truth knowledge-selection labels that can be utilized by the knowledge selection model. The average number of knowledge sentences for each context is 61.1.

We use this CMU_DoG dataset as our sub dataset. It is collected from crowd-sourced workers' conversations grounded on documents regarding movies. This online[3] dataset comprises 3,373 dialogues for training, 229 for validation, and 619 for test. The average number of turns in a dialogue session is 22.58. This dataset does not provide GT knowledge snippets. Therefore, we automatically create GT knowledge snippets on the dataset based on our observation, where lexically relevant words frequently occur in one of the knowledge sentences. For all examples in the datasets, pseudo GT knowledge sentence k_{GT} of response r is defined as:

$$k_{GT} = \begin{cases} k & \text{if } sim(r,k) > sim_{th} \text{ for } k \in KP_c \\ \texttt{\_NO\_KNOWLEDGE\_} & \text{else} \end{cases} \tag{3.30}$$

[3] `https://github.com/festvox/datasets-CMU_DoG` (Accessed: Sep. 15, 2019)

where $sim(\cdot, \cdot)$ is the similarity between two text inputs, sim_{th} is a hyperparameter determining whether the response r contains information in knowledge sentence k. We used similarity measure cosine similarity with TF-IDF representation to build pseudo GT knowledge sentences. We empirically set the hyperparameter sim_{th} to 0.1 by using knowledge selection performance in validation set of our baseline model E2E Transfo MemNet [39]. The average number of knowledge sentences for each context is 33.8.

Automatic Evaluation Setup We evaluate Recall@1 (R@1) for the knowledge selection task, which is the number of cases that the GT knowledge sentence is selected in the top-1 result divided by number of total test cases. We adopt unigram F1, BLEU, NIST scores between generated responses and references, and perplexity for the response generation task. F1 score is used as the main automatic metric in [39], and it is defined as:

$$F1 = \frac{2R \cdot P}{R + P} \tag{3.31}$$

Precision P is defined as $|W_H \cap W_R|/|W_H|$ and recall R is defined as $|W_H \cap W_R|/|W_R|$, where the set of nonstop words in reference response R and system response H are denoted by W_R and W_H, respectively. BLEU and NIST scores are used in the response generation task in the DSTC7 challenge [4]. The BLEU score measures with an n-gram matching degree. NIST scores operate in a similar manner as that of BLEU; however, it assigns more weight to rare words. It should be noted that a lower score for perplexity and higher for the other metrics indicate improved performance.

Human Evaluation Setup We recruited nine human annotators for qualitative evaluation of the systems' response. For each dataset, we randomly sampled 100 test examples, where three human annotators evaluated each sample. The

average number of evaluated samples per annotator was 66.7. During the test phase, we showed the dialogue context and randomly ordered responses of our method and baselines to the participants without any model information. We asked them to rate the quality of a response from 1-5 in two aspects by using the following questions:

- **Appropriateness (App.)**: How appropriate do you find the response to the dialogue context?
- **Informative gain (Info.)**: How much new and probable information does the response provide?[4]

Baselines We include baselines that can be trained in an end-to-end fashion. End-to-end models [20, 26, 1, 39, 111, 112] have been the mainstream of the research topic, open-domain response generation, due to their potential to leverage massive conversation corpus without hand-coding. We compare our models with five baselines as follows.

- **S2S**: This is a Seq2Seq model based on a 1-layer GRU encoder and a GRU decoder, which does not have knowledge access.
- **MemS2S** [26]: Several knowledge sentences are stored in memory units, and the fused knowledge vectors with dot product attention are added with the initial hidden state of the decoder.
- **TF-IDF Transfo Net**: Knowledge units are selected by cosine similarity with the TF-IDF model. The selected one and context are encoded by the Transformer encoder and injected to the Transformer decoder.

[4]We provide the following additional directions to the workers "The term "new" means how much the information in the response is novel compared with in the context, and "probable" says that it is likely to happen in the real world."

- **E2E Transfo MemNet** [39]: The model matches knowledge sentence and context for knowledge selection and uses a Transformer decoder for response generation.

- **DeepCopy** [31]: The model's decoder copies tokens from multiple knowledge sentences and the dialogue context. It uses feedforward network to calculate attention score over the context and each knowledge sentence with the decoder's hidden state.

We implement the baselines on our own because the authors' codes are not available. Moreover, the hyperparameters are chosen among those suggested in the corresponding paper and its variants on each dataset[5]. We set the length of dialogue context to 2 turns for MemS2S, DeepCopy, and E2E Transfo MemNet, following the original papers' suggestion. We report R@1 of the knowledge unit having the maximum score for MemS2S and DeepCopy because the models do not explicitly select a single knowledge unit.

[5]We experiment E2E Transfo MemNet with 100 random samples using the same random search used for our models, but we could not get superior results than the one reported in [39]. Thus, we assume that there is no significant difference between the grid search and the random search.

Models		R@1	PPL	F1	BLEU	NIST
Baseline Model	E2E Transfo MemNet [39]	0.187	65.1	0.164	0.009	0.409
Reduce-Match Models	Average Agg. → Bilinear Match	0.158	67.5	0.163	0.007	0.400
	Average Agg. → Cosine Match	0.148	64.6	0.161	0.007	0.396
	Average Agg. → MH Dotprod Match	0.216	63.3	0.175	0.011	**0.467**
	Average Agg. → DW Match	0.196	63.9	0.164	0.009	0.390
	GRU Agg. → Dotprod Match	0.230	62.8	0.174	0.012	0.444
	GRU Agg. → Bilinear Match	0.177	66.3	0.169	0.009	0.433
	GRU Agg. → Cosine Match	0.172	63.7	0.164	0.009	0.391
	GRU Agg. → MH dotprod Match	0.225	62.7	0.172	0.010	0.427
	GRU Agg. → DW Match	0.199	63.8	0.173	0.010	0.453
Match-Reduce Models	Dotprod Match → GRU Agg.	0.192	64.3	0.169	0.010	0.420
	Bilinear Match → GRU Agg.	0.219	64.6	0.173	0.010	0.431
	Cosine Match → GRU Agg.	0.209	63.1	0.173	0.009	0.425
	MH Dotprod Match → GRU Agg.	0.233	61.9	0.175	0.012	0.436
	DW Match → GRU Agg.	0.242	62.0	0.175	0.011	0.439
	Word/Seg. Match → CNN+GRU Agg. (SMN)	0.197	61.9	0.170	0.009	0.425
	Trasfo Layer Match → 3D CNN Agg. (DAM)	**0.254**	**60.6**	**0.178**	0.012	**0.467**
	Word/Seg. w/ Topic Match → CNN+GRU Agg. (KEHNN)	0.244	62.0	0.177	**0.013**	0.446

Table 3.1: Automatic evaluation results of models of each knowledge selection strategy in WoW dataset. Agg., Dotprod, MH, DW, Seg., and Transfo represent Aggregation, Dot product, Multi-head, Dimension-wise, Segment, and Transformer, respectively.

Models		R@1	PPL	F1	BLEU	NIST
Baseline Model	E2E Transfo MemNet [39]	0.222	48.2	0.114	0.002	0.137
Reduce-Match Models	Average Agg. → Bilinear Match	0.199	**46.4**	0.122	0.002	0.187
	Average Agg. → Cosine Match	0.226	49.5	0.111	0.002	0.103
	Average Agg. → MH Dotprod Match	0.247	48.2	0.124	0.003	0.157
	Average Agg. → DW Match	0.229	54.3	0.105	0.002	0.081
	GRU Agg. → Dotprod Match	**0.277**	52.4	**0.136**	**0.007**	**0.250**
	GRU Agg. → Bilinear Match	0.158	46.5	0.124	0.003	0.201
	GRU Agg. → Cosine Match	0.245	51.1	0.107	0.003	0.105
	GRU Agg. → MH dotprod Match	0.262	48.8	0.120	0.004	0.159
	GRU Agg. → DW Match	0.247	49.5	0.112	0.003	0.132
Match-Reduce Models	Dotprod Match → GRU Agg.	0.238	49.7	0.108	0.003	0.126
	Bilinear Match → GRU Agg.	0.247	48.0	0.114	0.003	0.162
	Cosine Match → GRU Agg.	0.234	51.0	0.110	0.003	0.121
	MH Dotprod Match → GRU Agg.	0.250	48.4	0.117	0.003	0.146
	DW Match → GRU Agg.	0.261	51.7	0.120	0.004	0.151
	Word/Seg. Match → CNN+GRU Agg. (SMN)	0.233	50.8	0.118	0.003	0.157
	Trasfo. Layer Match → 3D CNN Agg. (DAM)	0.255	47.1	0.128	0.005	0.193
	Word/Seg. w/ Topic Match → CNN+GRU Agg. (KEHNN)	0.232	46.7	0.130	0.006	0.211

Table 3.2: Automatic evaluation results of models based on each knowledge selection strategy in CMU_DoG dataset. Agg., Dotprod, MH, DW, Seg., and Transfo represent Aggregation, Dot product, Multi-head, Dimension-wise, Segment, and Transformer respectively.

Models	Automatic evaluation					Human evaluation	
	R@1	PPL	F1	BLEU	NIST	App.	Info.
S2S	-	114.4	0.144	0.004	0.385	2.61 (1.54)	2.66 (1.57)
MemS2S [26]	0.023	105.7	0.117	0.002	0.278	1.15 (0.63)	1.16 (0.64)
TF-IDF Transfo Net	0.111	65.8	0.160	0.007	0.411	3.09 (1.42)	3.24 (1.54)
E2E Transfo MemNet* [39]	-	63.5	0.169	-	-	3.28 (1.40)	3.31 (1.49)
DeepCopy [31]	0.128	**59.0**	0.172	0.009	0.466	3.22 (1.43)	3.21 (1.46)
Our best model (DAM)	**0.254**	60.6	**0.178**	**0.012**	**0.467**	**3.43 (1.38)**	**3.52 (1.47)**

Table 3.3: Comparison of our models with baselines in WoW dataset. *The automatic evaluation sores of E2E Transformer MemNet are from the original paper. We report the mean ratings and their standard deviation (in parenthesis) of different methods for Appropriateness (App.) and Informative gain (Info.) scores for human evaluation.

Implementation Details In this work, we aim at exploring a wide rage of variants in the knowledge selection methods; thus, we adopt an random hyperparameter search algorithm to find efficiently the best hyperparameters for each knowledge selection model. Specifically, the hyperparameters of our models are tuned with the tree structured Parzen estimator algorithm with asynchronous Hyper Band scheduling implemented in Ray framework[6] for fair comparison. We set the number of random samples for each of the knowledge selection models to the value in [18, 100] proportional to the hyperparameter space of each. Consequently, in the WoW and CMU_DoG datasets, both of encoder and decoder in our Transformer use 256 hidden units, 2 attention heads, and 1,024 hidden units of position-wise feed-forward network for all knowledge selection models. The only difference of the datasets is the number of layers in encoder and decoder, which are 6 and 2 for the WoW and CMU datasets, respectively. Our Transformer encoders use the shared parameter between the knowledge encoder and the conversation encoder. We determine the hyperparameters of

[6] `https://github.com/ray-project/ray` (Accessed: Sep. 4, 2019)

Models	Automatic evaluation					Human evaluation	
	R@1	PPL	F1	BLEU	NIST	App.	Info.
S2S	-	108.3	0.130	0.004	0.189	2.50 (1.44)	1.94 (1.26)
MemS2S [26]	0.231	64.2	0.131	0.003	0.175	2.39 (1.47)	1.59 (1.09)
TF-IDF Transfo Net	0.180	**41.6**	0.133	0.005	0.228	2.65 (1.40)	2.37 (1.40)
E2E Transfo MemNet [39]	0.222	48.2	0.114	0.002	0.137	2.54 (1.46)	1.77 (1.19)
DeepCopy [31]	0.072	78.3	0.131	0.005	0.219	2.78 (1.44)	2.18 (1.38)
Our best model (GRU Agg. → Dotprod)	**0.277**	52.4	**0.136**	**0.007**	**0.250**	**2.82 (1.52)**	**2.42 (1.42)**

Table 3.4: Comparison of our models with baselines in CMU_DoG dataset. We report the mean ratings and their standard deviation (in parenthesis) of different methods for Appropriateness (App.) and Informative gain (Info.) scores for human evaluation.

each knowledge selection model by using the above-mentioned tuning method. We set the number of knowledge sentences as 10 and λ as 0.95 for training all of our models and early stop using perplexity on validation set and patience of 12. The parameters were updated by Adam algorithm [113], whose parameters, β_1 and β_2, are 0.9 and 0.98, respectively. The learning rate was set to 5e-4 increasing linearly for the first 4,000 warmup steps and decreasing proportionally to the inverse square root of the step number. We used a greedy search for decoding the response sentence. Any pre-trained word embeddings are not used for all the models including baselines and ours.

3.3.2 Experimental Results

Comparison of variants of our models Table 3.1 and Table 3.2 provide the automatic evaluation results of our models for the WoW and CMU_DoG datasets, respectively. The performance of the knowledge selection models varies in a wide range: $[0.148, 0.254]$ for WoW and $[0.158, 0.277]$ for CMU_DoG, which shows the importance of choosing a proper knowledge selection model. There is a strong correlation between knowledge selection R@1 and automatic metric of response generation. Pearson coefficients R@1 between automatic evaluation

metrics are -0.75, 0.87, 0.92, and 0.75 for perplexity, F1, BLEU, and NIST, respectively (lower perplexity is desirable).

In the WoW dataset, all Reduce-Match models except the ones using bilinear and cosine match function enhance the knowledge selection and the response-generation performance of the baseline model, E2E Transfo MemNet, that uses a Reduce-Match model (Average Agg. → Dot product Match). On the contrary, all Match-Reduce models improve the baseline's performance, where these results show the effectiveness of the turn-level matching. Among our models based on the two matching strategies, the knowledge selection model based on DAM and KEHNN provides the best results in all the evaluation metrics. These results show the importance of considering the granularity of matching in knowledge selection. Results in the CMU_DoG dataset show a similar but slightly different trend from those in WoW. In contrast to WoW's result, the Reduce-Match models prove superior to the Match-Reduce models. (GRU Agg. → Dotprod Match) outperforms the others in automatic evaluation metrics, i.e., R@1, F1, BLEU, and NIST score and human evaluation metrics, which shows the efficacy of GRU aggregation in extracting context features. In our opinion, this can be attributed to the fact that the CMU_DoG dataset size is not sufficiently large to train the deep matching method.

Comparison with baselines Table 3.3 and Table 3.4 provide comparisons of our best models and the baselines in automatic and human evaluations. The results show that our best models (DAM and GRU Agg. → Dotprod Match) outperform the other competitive baselines in automatic evaluation metrics, i.e., R@1, F1, BLEU, NIST score and all the human evaluation metrics. This supports our argument that our models can achieve improvements of response generation by enhancing the knowledge selection module. For both the con-

sidered datasets, the perplexity results of our models are inferior to the other baseline models. This is owing to the reason that the model of lower perplexity generates only general or probable utterances and does not risk generating informative or rare words.

3.4 Analysis

We first review the outputs of the models by conducting case study. Then we provide in-depth comparisons of performance on the WoW dataset according to three aspects: matching difficulty, length of dialogue, and dialogue act of message, which reflects the different aspects of challenges in KGC. For the analysis, we compared our four best performing models adopting the same aggregation method for each knowledge selection strategy to the baseline E2E Transfo MemNet. We use mean reciprocal rank (MRR) for the performance comparisons.

3.4.1 Case Study

Table 3.5 shows examples of response and knowledge sentence selected by the models for the two datasets: WoW and CMU_DoG. The example with the WoW dataset is regarding a conversation where participant A presents his/her personal preference and a nickname for the panda. Subsequently, participant B responds to participant A stating their interest in a panda's color. Although responses of the S2S and MemS2s models are not relevant to the context, those of the other models: TF-IDF Transfo Net, E2E Transfo MemNet, and DeepCopy are relevant to some extent. Each response shows a slight lack of concentration to the given context, e.g., TF-IDF Transfo Net's fails to mention another nickname, E2E Transfo MemNet fails to provide additional information regarding previous context, and DeepCopy provides knowledge about the panda that is very generic. On the contrary, our best model responds with more direct at-

tention over the last utterance, which is a key utterance in the context. The example with the CMU_DoG dataset presents a conversation about the movie Jaws, where participant A asks a question to confirm the name of the movie. In this example, the S2S and MemS2S models generate irrelevant and non informative responses; the other models except our model do not convey meaningful information. The response of our best model (Average Agg. → Dot product match) is much more informative.

Figure 3.4 and Figure 3.5 show representative failure cases of our models in the Wow dataset. Figure 3.4 shows the outputs of the DAM when the model fails to select the GT knowledge sentence. Figure 3.4a shows that the DAM finds a knowledge sentence that matches the knowledge with the terms 'blue' and 'sky' and uses it in response, which differs from what the human chose. In the case of humans, the word 'blue' and 'sky' in the knowledge matches the words in context, but he/she changes the conversation topic to a movie named 'blue sky.' We presume the reason that our model cannot simulate such behavior is that there is no mechanism to capture such topic transition explicitly. Another failure case is shown in Figure 3.4b. In this case, the model selects the knowledge sentence that has already been transferred to the other party in the context, showing it tends to select knowledge semantically similar to the context. Figure 3.5 shows cases where there exist problems in generating the response. Figure 3.5a demonstrates the case in which the model ignores the knowledge selected by itself and generates the words relevant to the context, e.g., 'early.' Figure 3.5b presents a case where it fails to use the correct words from knowledge (the word after 'means' should be 'motion' or 'movement' instead of 'kinema.') and repeat the word 'kinema.'

Item	Wizard-of-Wikipedia	CMU_DoG
Context	**Topic:** Giant panda **A:** Hi! I love giant pandas, they are my favourite animal. , they're also called "big bear cat" sometimes? how about you? **B:** i love pandas, mostly their colours	**Topic:** Jaws **A:** hello , i understand we 're supposed to chat about a movie ? **B**: yes , jaws **A**: jaws . the one from the 1980s ?
S2S	**R:** their wild carrot is a big problem , they are so good . they are so good	**R:** yes , it is a movie .
MemS2S	**R:** i ' m not sure , but i know that it is a very popular sport . **K.** The National Zoological Park, commonly known as the National Zoo, is one of the oldest zoos in the United States.	**R:** heloo **K.** _NO_KNOWLEDGE_
TF-IDF Transfo Net	**R:** i ' m not sure but i know they are called " giant panda " which means " giant panda " . **K.** In each category, China is the most biodiverse country outside of the tropics.	**R:** jaws is a classic movie . **K.** Brody finally convinces a devastated Vaughn to hire Quint.
E2E Transfo MemNet	**R:** i ' m not sure , but i do know that they are the only species that eat bamboo . **K.** Though it belongs to the order Carnivora, the giant panda's diet is over 99% bamboo.	**R:** yes **K.** _NO_KNOWLEDGE_
DCP	**R:** they are a mammal , which is a mammal which is a mammal . **K.** The giant panda lives in a few mountain ranges in central China, mainly in Sichuan, but also in neighbouring Shaanxi and Gansu.	**R:** yes , i saw it in theaters , it was a great movie __eos__ i ' m not sure about that **K.** Jaws is too gruesome for children, and likely to turn the stomach of the impressionable at any age.
Our best model	**R:** i love them ! they are so cute and have a large black patches in their ears . **K.** It is easily recognized by the large, distinctive black patches around its eyes, over the ears, and across its round body.	**R:** yes , it is a 1975 thriller film . rotten tomatoes gave it a 97 % rating . **K.** thriller Rotten Tomatoes: 97% with an average: 9.2/10 Metacritics: 87/100 IMDB: 8.0/10 1975

Table 3.5: Examples of response and knowledge selected (or the best scoring knowledge for MemS2s and DeepCopy) by the different models. The responses and Knowledge selected are denoted by boldfaced R and K, respectively. Our best models for the WoW and CMU_DoG are DAM and Average Agg. → Dot product Match, respectively.

Item	Text
Context	**Topic**: Blue **M**: Blue is my favorite primary color. **U**: Blue is always nice. I like royal blue. **M**: I once road on The Royal Blue train from New York to D.C **U**: Oh that sounds really nice. I bet there was a lot of scenery and blue skies.
DAM' response	i love the clear daytime sky and the sea appear blue because of an optical effect known as rayleigh scattering .
DAM's selected knowledge	Blue __knowledge__ The clear daytime sky and the deep sea appear blue because of an optical effect known as Rayleigh scattering.
Ref.	Yes, speaking of Blue Skies, have you seen the 1946 movie staring Bing Crosby?
GT Know	Blue Skies (1946 film) __knowledge__ Blue Skies is a 1946 American musical comedy film directed by Stuart Heisler and starring Bing Crosby, Fred Astaire, and Joan Caulfield.

(a) Failing in topic transition

Item	Text
Context	**Topic**: Blue **M**: Blue is my favorite primary color. **U**: Blue is always nice. I like royal blue.
DAM' response	blue is one of the three primary colors .
DAM's selected knowledge	Blue __knowledge__ Blue is one of the three primary colours of pigments in painting and traditional colour theory, as well as in the RGB colour model.
Ref.	I once road on The Royal Blue train from New York to D.C
GT Know	Royal Blue (train) __knowledge__ The Royal Blue was the Baltimore and Ohio Railroad (B&O)'s flagship passenger train between New York City and Washington, D.C., in the United States, beginning in 1890.

(b) Repeating knowledge

Figure 3.4: Failures in knowledge selection. Ref. and GT Know stands for human response and GT knowledge sentence, respectively.

3.4.2 Impact of Matching Difficulty

Each conversation example has a different level of matching difficulty in finding the GT knowledge sentence owing to different levels of word redundancy in the context. To measure this concept systemically, we define matching difficulty as the ratio of similarity of the GT knowledge sentence to maximum similarity of other candidate knowledge sentences. We use the cosine similarity between averages of the embeddings of context and knowledge sentence since it has

Item	Text
Context	**Topic**: Genius **M**: a genius is a person who displays exceptional intellectual ability. **U**: What is an IQ of a genius? **M**: universality in genres or originality typically to a degree that is associated with achievements **U**: Is that a trait that you can tell someone has early on in their life?
DAM' response	the term genius is often used in the early 19th century
DAM's selected knowledge	Genius __knowledge__ Because the achievements of exceptional individuals seemed to indicate the presence of a particularly powerful "genius", by the time of Augustus, the word began to acquire its secondary meaning of "inspiration, talent".
Ref.	there is no scientifically precise definition of genius
GT Know	Genius __knowledge__ There is no scientifically precise definition of genius, and the question of whether the notion itself has any real meaning has long been a subject of debate, although psychologists are converging on a definition that emphasizes creativity and eminent achievement.

(a) Not using knowledge

Item	Text
Context	**Topic**: Cinematography **U**: Hi buddy, What you think about cinematography **M**: Cinematography, is a type of motion picture , captured electronically by means of an image **U**: Yes buddy, Images captured with an electronic image-sensor, produces an electrical charge.The word "cinematography" is based on the Greek words meaning movement, motion.
DAM' response	the word cinematography comes from the greek word " kinema " which means " kinema " .
DAM's selected knowledge	Cinematography __knowledge__ The word "cinematography" was created from the Greek words ("kinema"), meaning "movement, motion" and ("graphein") meaning "to record", together meaning "recording motion."
Ref.	It works by lens used to repeatedly focus the light reflected from objects into real images on the light-sensitive surface .
GT Know	Cinematography __knowledge__ Typically, a lens is used to repeatedly focus the light reflected from objects into real images on the light-sensitive surface inside a camera during a questioned exposure, creating multiple images.

(b) Repeating words

Figure 3.5: Failures in response generation. Ref. and GT Know stands for human response and GT knowledge sentence, respectively.

been commonly used in various applications such as evaluation for response generation thanks to its robustness to word mismatch problem. We use 300d fasttext vectors trained on Wikipedia[7] for the embeddings. We divide 2,409 queries of 4 length context into 3 groups equally according to the difficulty

[7] `https://dl.fbaipublicfiles.com/fasttext/vectors\-wiki/wiki.en.vec` (Accessed: Dec. 23, 2019)

level.

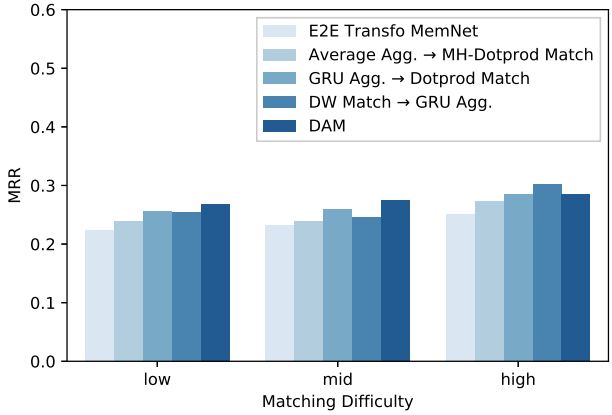

Figure 3.6: Performance of knowledge selection depending on matching difficulty. We report the average MRR of the model in three groups of difficulty levels (relative similarity of GT knowledge sentence and maximum similarity of the others).

The result is shown in Figure 3.6. Our models consistently outperform the baseline E2E Transfo MemNet for all the difficulty levels. The DAM outperforms the others by a large margin in the low and middle levels but by a small margin in the high level. Dimension-wise matching is effective when the semantic distance between context and knowledge unit is high. Interestingly, there exists a trend that the performance of all models in low difficulty level is not better than the ones in high difficulty level, which requires in-depth investigation. We suspect the reason for this phenomenon is that we trained our models only on a limited-size of the dataset.

3.4.3 Impact of Length of Context

As the conversation proceeds, the number of keywords in the context increases, which means that the degree of redundancy increases. To evaluate the models in

such cases, we compare the knowledge-selection accuracy for different context lengths, i.e., the number of turns used in the model. Figure 3.7 shows that DAM consistently outperforms other models for all context lengths. Notably, a substantial margin is seen when context length is short. The DAM seems particularly effective for matching short sentences when the shortness of the context limits redundant information. Moreover, a trend wherein the accuracy of each model drops as the turn proceeds is seen. This trend indicates that the problem of filtering irrelevant information is graver.

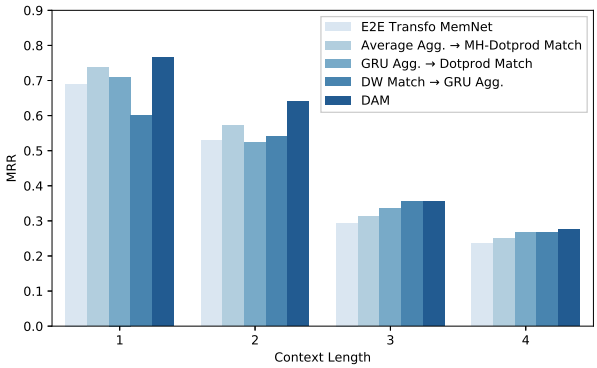

Figure 3.7: Performance of knowledge selection depending on context length

3.4.4 Impact of Dialogue Act of Message

Dialogue acts (DAs) or speech acts, represent the intention behind an utterance in conversation to achieve a conversational goal [57]. Modeling conversations as structured DA sequences can be regarded as a fundamental step toward the automated understanding of dialogue. We use DAs to compare performance of the models in detail. Dialogue acts in our datasets are not available; therefore, we manually annotated the DA for each utterance in the last turn (often called message) of the WoW's 100 queries used for human evaluation. We used the

DAs defined on [114] plus *Topic* DA, i.e., the start of dialogue session.

Table 3.6 presents knowledge selection performance for different proportions of DAs in the samples. Statement-non-opinion/statement-opinion and question DAs have a large proportion in the random samples, which show the dialogue patterns of communicative exchange represented as opinions and facts between participants. Additionally, the table shows that our four models outperform the baselines in the queries of statement-non-opinion and statement-opinion DA, where informative or salient words may frequently occur. However, the performance on the query of Question-type DAs, i.e., Wh-Question and Yes-No-Question, is not much superior to the baselines'. We suspect the reason is that our similarity-based methods have difficulty in handling such question queries where its GT knowledge sentence often contains requested information that its semantic-relatedness to the context is not close enough.

Dialogue Act	Proportion (%)	E2E Transfo MemNet	Average Agg. → MH Dotprod	GRU Agg. → Dotprod Match	DW Match → GRU Agg.	DAM
Statement-non-opinion	25.33	0.27	0.35	**0.44**	0.38	0.37
Statement-opinion	22.67	0.24	0.29	**0.31**	**0.31**	0.29
Wh-Question	18.67	0.34	0.36	**0.41**	0.34	0.35
Yes-No-Question	14.67	0.31	**0.34**	0.23	0.33	0.30
Topic	9.33	0.82	**0.85**	0.77	0.62	0.77
Agree/Accept	2.67	0.08	0.05	**0.14**	0.09	0.06
Acknowledge (Backchannel)	2.00	**0.47**	0.42	0.39	0.24	0.21
Yes answers	1.33	0.18	0.09	0.10	0.14	**0.31**
DeclarativeYes-No-Question	0.67	0.03	0.03	0.04	0.02	**0.05**
Backchannel in question form	0.67	0.07	0.14	0.14	0.03	**0.20**
Thanking	0.67	0.06	0.04	0.05	0.05	**0.07**
Non-verbal	0.67	**0.04**	0.03	**0.04**	0.03	**0.04**
Uninterpretable	0.67	0.10	0.33	0.33	0.25	**0.50**

Table 3.6: Knowledge selection accuracy (MRR) for each Dialogue Act. Agg., Dotprod, MH, DW, Seg., and Trasfo stands for Aggregation, Dot product, Multi-head, Dimension-wise, Segment, and Transformer respectively.

Chapter 4

Knowledge Ranking with Local Context and Topic Keywords

4.1 Motivation

Many datasets based on the knowledge of various sources, such as Wikipedia [39] and articles regarding movies [65, 115], have been proposed to train the KGC models by using conversations between crowd-sourced workers. Most conversation data sets contain a significant portion of knowledge grounded utterances that are grounded on one or more facts. A computing model must have a KB, in one form or another, in order to carry out KGC, and document collections such as Wikipedia are natural choices for this purpose. Henceforth, we will call an utterance to be knowledge grounded when there is an association (as defined by the model) between the utterance and one or more documents in the KB in consideration. Figure 4.1 shows an example of the KGC on the internet where multiple users naturally exchange information by referring to various documents without providing GT knowledge documents. Generating knowledge grounded

responses in such a situation is challenging because some users can refer to multiple documents on various topics at the same time as shown in Turn # 3-1 or # 3-2 without sharing the GT knowledge documents. The user of Turn # 3-1 responds by referring to both the main topic document on 'Audrey Hepburn' and another document on 'Julie Andrews.'

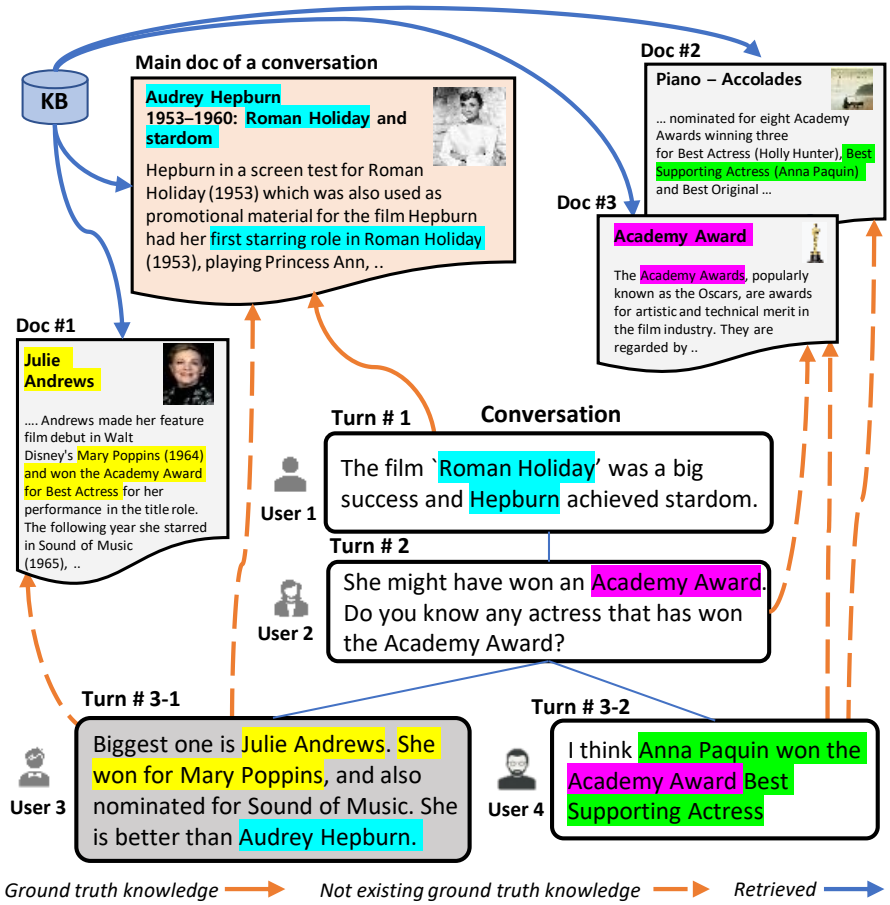

Figure 4.1: **Knowledge Grounded Conversation on the Internet** where User 1 begins the conversation by sharing a document that is the main topic of the conversation, and others share their opinions while omitting the GT knowledge documents. Our model aims to generate responses based on the main document of a conversation and the relevant but diverse topics, as shown in Turn # 3-1.

However, most existing methods such as [68] consider only one document when generating a response (as shown in Turn # 2). Such a method may limit the content of the responses, making it difficult to deliver diverse and interesting information. On the other hand, the retrieval-augmented KGC models by Shuster et al. [25] can utilize multiple documents at once (as shown in Turn # 3-2). However, the models have limitations of ignoring the main documents that people are interested in because they consider only a limited number of tokens that come immediately before the response, which we call *local context*. Furthermore, the models can suffer retrieval collapse phenomenon [2], in which they retrieve the same and irrelevant documents regardless of the inputs. Figure 4.2 presents two outputs for different conversation histories of the E2E retrieval-augmented model (RagToken) trained on the CbR's trainset. We presume that the reason for this is that a high ratio of non-grounded responses[1] can flood the training process with meaningless signals.

To this end, we introduce a novel method aimed at generating interesting responses grounded on the main documents or other relevant documents (as shown in Turn # 3-1) by retrieving documents based on topic keywords throughout the whole context and local context. Specifically, we propose an end-to-end (E2E) retrieval-augmented KGC model based on our novel Topic-aware Dual Matching (TADM) re-ranker. Our model considers both the topic of the whole conversation and the local context to compose top-k documents regarding an appropriate range of relevant topics. Our conversation encoder accepts a fixed number of tokens before the response (local context) and also topic keywords automatically extracted from the whole conversation history (holistic context). Then the following two matching layers separately compare

[1]Our examining of 74 random samples from the test split of CbR task [40] showed that 54 % of responses were not grounded to any documents on the web.

Context	>> til that the sex scene between julie christie and donald sutherland in ' don't look now ' looked so real that warren beatty , who was dating julie christie at the time , tried to get it removed because he genuinely thought the two had sex
RagToken	it's because of his " genuinely " that they had sex .
Top-5 results	1. Joshua Brand | Pyromaniac's Love Story", the television movies "Wall to Wall Records" and "Homeward Bound", the television pilot "Gemini Man", and episodes of "thirtysomething" and "Joan of Arcadia". Brand was also nominated for an Emmy and a DGA Award for ...
	2. Elizabeth Wurtzel | of her double mastectomy and reconstruction, "It is quite amazing. They do both at the same time. You go in with breast cancer and come out with stripper boobs." Elizabeth Wurtzel Elizabeth Lee Wurtzel (born July 31, 1967) is an ...
	3. RYOT | D.C. A year later, Mooser and Darg, debuted their third film at the 2013 Tribeca Film Festival, a documentary short titled "The Rider and The Storm", which chronicles a New York surfer who lost everything in the Breezy Point fires during Hurricane ...
	4. David Newsom | 2007 AFI Directing Workshop for Women grant, wrote "OYE" as a tribute to the women she had known who battled and lived with cancer. In May, 2008, "Open Your Eyes" was awarded the AFI "Jean Picker Firstenburg Award of Excellence". In recent ...
	5. Jonathan Stern | "Beef", "The Hotwives of Las Vegas" and "The Hotwives of Orlando" on Hulu. He has produced numerous feature films including "A Futile and Stupid Gesture", "Oxygen", "Mexico City", "Scotland, PA", "The Vagina Monologues", "Confess", ...

Context	chiefs qb elvis grbac was selected people's " sexiest athlete alive " in 1998 by accident . rather than offend grbac , the magazine just ran with it
RagToken	so what's the point in posting this ? what's the point of posting this ? it's not even relevant to the chiefs
Top-5 results	1. Joshua Brand | Pyromaniac's Love Story", the television movies "Wall to Wall Records" and "Homeward Bound", the television pilot "Gemini Man", and episodes of "thirtysomething" and "Joan of Arcadia". Brand was also nominated for an Emmy and a DGA Award for ...
	2. Elizabeth Wurtzel | of her double mastectomy and reconstruction, "It is quite amazing. They do both at the same time. You go in with breast cancer and come out with stripper boobs." Elizabeth Wurtzel Elizabeth Lee Wurtzel (born July 31, 1967) is an ...
	3. RYOT | D.C. A year later, Mooser and Darg, debuted their third film at the 2013 Tribeca Film Festival, a documentary short titled "The Rider and The Storm", which chronicles a New York surfer who lost everything in the Breezy Point fires during Hurricane ...
	4. David Newsom | 2007 AFI Directing Workshop for Women grant, wrote "OYE" as a tribute to the women she had known who battled and lived with cancer. In May, 2008, "Open Your Eyes" was awarded the AFI "Jean Picker Firstenburg Award of Excellence". In recent ...
	5. Jonathan Stern | "Beef", "The Hotwives of Las Vegas" and "The Hotwives of Orlando" on Hulu. He has produced numerous feature films including "A Futile and Stupid Gesture", "Oxygen", "Mexico City", "Scotland, PA", "The Vagina Monologues", "Confess", ...

Figure 4.2: Two examples showing the retrieval collapse phenomenon in KGC. The results of retrieval using different context are the same and irrelevant to the context.

the output representations of the conversation encoder with the ones of the candidate documents to simultaneously consider holistic and local features. To train our model on the KGC data sets without GT knowledge documents, we propose a new data weighting scheme that encourages our model to generate grounded and informative responses. We use term matching-based similarity between a response and the top-1 retrieved knowledge as one type of weight. We also utilize the inverse document frequency (IDF) of the response's terms. Figure 4.3 shows our overall model architecture, where the proposed modules in this chapter are highlighted in violet.

4.2 Retrieval-Augmented Knowledge Grounded Conversation Model

We focus on a knowledge grounded response generation task in which the system is given KB as knowledge sources. A system should generate responses grounded on the relevant factoid documents and relevant to the conversation history. More specifically, we are given a KB containing documents without any links between a response and its grounded document, i.e., GT knowledge document. Formally, the system is given a conversation history of turns $X = (x_0, ..., x_M)$ and KB of documents KB. With the conversation history X, the system needs to generate a natural language response y that is relevant to the web documents in the KB and the conversation history. The KB consists of all Wikipedia pages and main documents of a conversation topic. We define the *main document* as passages provided at the beginning of the conversation.

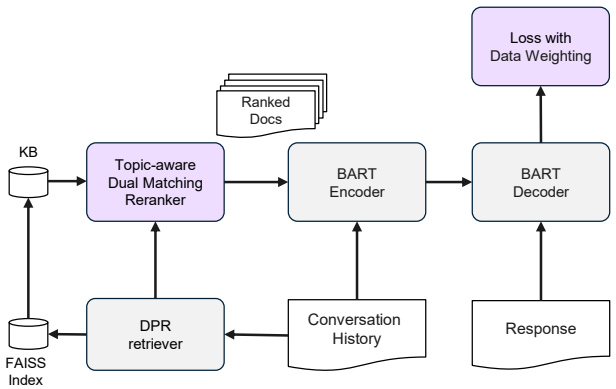

Figure 4.3: **Model architecture.** Our model is based on RAG token [2] with two significant changes, i.e., Topic-aware Dual Matching Re-ranker to enhance the retrieval accuracy and data weighting scheme to encourage generating grounded responses.

4.2.1 Base Model

We choose RagToken [2] as our base model because the intuition of RAG-token is the same as our assumption, which is that each token can be generated based on multiple documents. Lewis et al. [2] proposed answer generation models equipped with a neural retriever trained end-to-end, called Retrieval-Augmented Generation (RAG), for open-domain QA tasks. While training, it conducts a K-NN document search from a large document KB by utilizing the FAISS index [116]. The training objective of the RAG token is shown as follows:

$$P(y|x) = \prod_{i}^{N} \sum_{z \in top\text{-}k(p(\cdot|x))} p(z|x)p(y_i|x, z_i, y_{i:i-1}) \tag{4.1}$$

where x is a question; z is a document, and y is a answer of the question. The dense passage retrieval (DPR) model scores the $|top\text{-}k(\cdot)|$ documents retrieved by the FAISS index, and the second term conditional probability of generating each token $p(y_i|\cdot)$ is calculated based on different documents z_i, the input x, and previously generated tokens $y_{i:i-1}$.

We replace the input and output of the RAG model with the ones of the KGC setup and utilize it as the basis of our model. x will be the conversation history; z will be the external document, and y will be the next response for x. The $p(z|x)$ of our model is computed by using a weighted sum of the score from DPR and score from the re-ranker described in the followings.

4.2.2 Topic-aware Dual Matching for Knowledge Re-ranking

Figure 4.4 shows our retriever and re-ranker, which comprises a conversation encoder, a document encoder, dual matching layers, a topic keyword extractor, a salient token checker, and a scorer layer. The conversation encoder encodes local context and topic keywords from the conversation history to a sequence of

vectors. Then retriever queries FAISS index to find c candidate documents using the representation corresponding to the [CLS] token. The document encoder encodes retrieved documents. The dual matching layers, shallow Transformer encoders, conduct more fine-grained matching between conversation and documents and yield two scores using linear layers. Then, the scorer layer outputs the weighted sum of the two scores.

Conversation encoder: The conversation encoder encodes both tokens before the response and the topic keywords of the whole conversation to retain top-k documents that consider the appropriate range of the relevant topics of the conversation. To implement this strategy, we use an external topic word extraction module implemented as TextRank [117] to recognize important words from the initial turns of the conversation history and use them as additional input with the tokens before the response. We exploit the initial turns because they tend to focus on documents of the conversation topic, which may also impact all the turns of the conversation history. Then, we assign different segment embeddings to each of these two input types, turns before the response, and topic keywords, and add them with token embeddings and position embeddings.

Dual matching layers: Prior works [25, 118] pointed out that the method using only a vector at [CLS] lacks the interaction between the two inputs, which they considered as one of the main causes of the poor performance. Following this line of work, we propose to match the conversation and document with multiple representations of the encoders. Our dual matching layers separate the representations from the conversation/document encoder and matches embeddings in the same group. Each matching layer is composed of two layers of the Transformer encoder, which consists of the Matching Layer for Local

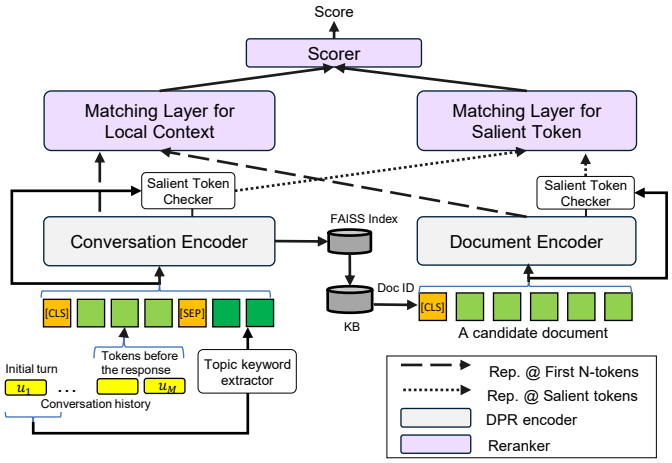

Figure 4.4: **Retrieving documents with Topic-aware Dual Matching Re-ranker**: The conversation encoder encodes tokens before the response with topic keywords, and then, it retrieves Top-k candidate docs from KB while the document encoder encodes the candidate docs. The Matching Layer for Local Context accepts representations on top of the first N-tokens of the inputs. The Matching Layer for Salient Token takes representations selected by the salient token checker. The scorer layer aggregates the scores of each doc from the dual matching layers.

Context (MLLC) and Matching Layer for Salient Token (MLST). The MLLC matches representations on top of the first N-token, and the MLST matches representations corresponding to the representations on top of the tokens selected by our salient token checker. Our salient token checker is an external natural language processing tool that assigns each token to a category label. For the salient token checker, we use a union of the named entity, non-stopwords, and keywords extracted by TextRank [117].

Our intuitions behind the dual matching layers are as follows. First, we expect the MLLC to match representations having global features related to the local context and document. Second, we expect the MLST to match rep-

resentations having information relevant to informative or crucial keywords. The analysis on self-attention trained by BERT [119] supports our intuitions. In that work, they visualized the attention scores on some meaningful words, where scores on noun phrases are higher than stopwords'.

4.2.3 Data Weighting Scheme for Retrieval Augmented Generation Models

In the KGCs on the internet, users often do not provide GT knowledge documents that they refer to (shown as not existing knowledge in Figure 4.1) and make casual responses that are not grounded. To handle this issue, we consider two metrics, the groundness [25] and informativeness [20] to evaluate the quality of the responses. The groundness metric matches the main purpose of ours; however, the diversity metric does not. However, we assume that the informative responses, specifically those that include words that rarely occur in the conversation, can be grounded on knowledge or can be interesting. For example, the term 8,849 in an utterance grounded in some knowledge "The height of Mt. Everest is 8,849 m" rarely occurs in our conversation.

As a result of the above discussion, we propose a novel instance weighting scheme for retriever-augmented KGC models. Algorithm 1 shows the training method using our proposed weighting scheme where an instance's weight w_i is the geometric mean of BLEU(z_e, y) and IDF-W(y). Apart from machine reading compression style KGC models such as [40], our method does not assume that responses are grounded on the given document. Therefore, we retrieve the documents from KB by using the DPR retriever with a query $(x, y)_j$ and use the similarity value as a surrogate of the true groundness degree. Specifically, we use the sentence-level BLEU score, which we confirmed that it shows better results than the similarity measures based on distributional representations. For informativeness, we propose to use the length-penalized summation of the

term's IDF shown in Eq. 4.2. We calculate the weighting score by dividing the sum of IDFs by the length as in [120] to avoid generating only long responses.

$$\text{IDF-W}(y) = \sum_{w_i \in y} IDF(w_i) \frac{(4 + length)^{\alpha}}{(4+1)^{\alpha}} \tag{4.2}$$

Here, α controls the strength of the length normalization.

Algorithm 1: Training algorithm

Data: Training Data D
Result: Trained Model M
`// Before training`
Calculate IDF of words in turns;
for $(x, y)_j \in D$ **do**
  $Z \leftarrow$ Top-K docs relevant to $(x, y)_j$ from KB;
  $e \leftarrow argmax_{z \in Z}\text{BLEU}(z, y)$;
  $w_j \leftarrow gmean(\text{BLEU}(z_e, y), \text{IDF-W}(y))$;
end
`// Training`
while M *Converges* **do**
  $B \subset$ **shuffle**(D);
  $w_b^{'} \leftarrow w_b / \sum_{b \in B} w_b$;
  Multiply loss of b-th example with $w_b^{'}$;
  Update Model M with the loss;
end

4.3 Experiments

4.3.1 Experimental Setup

We use a Reddit KGC dataset, CbR [40], as our benchmark. The dataset consists of conversation threads extracted from Reddit.com. Each conversation is linked to a main document shared at beginning. We provide the main document with KGC models except for the retrieval-based models. The dataset has 2.8M, 5.9k, and 13k instances for the training, validation, and testing, respectively. For the evaluation, we use a test set of 2208 instances built by Galley et al.

[4], for which 6 responses are available. Additionally, we built another test split that we call internet-grounded split by manually choosing responses that contain knowledge of webpages on the internet such as Wikipedia and news stories. The resulting split includes 204 instances having only one response for each conversation history. We use this test split to evaluate the model's capability of utilizing knowledge of the documents in KB.

Automatic Evaluation Metrics We evaluate our models in terms of conversational aspect and grounding aspect.

1. **Relevance**: We use three metrics measuring similarity between response and human responses, i.e., BLEU [121], METEOR [122], and NIST [123], as the surrogates of the relevance between the response and context. NIST is a variant of BLEU that weighs more informative n-gram.

2. **Ground-to-MainDoc**: This metric is proposed by Qin et al. [40]. It measures the systems' ability to exploit knowledge from the document of a conversation topic by calculating the number of word overlaps between the reference document and generated responses except for words from the conversation history.

3. **Ground-to-Internet**: We evaluate the model's ability of utilizing knowledge in the KB using the internet-grounded split. We measure similarities between three responses generated by a model and the reference responses and the maximum value to alleviate the word-mismatch problem. The equation for this is presented as follows:

$$k\mathrm{R}_s = \frac{1}{|D_{test}|} \sum_{l=1}^{|D_{test}|} max_{i \in [1,3]} s(\hat{y}_l, y_{l,i}) \tag{4.3}$$

where s is chosen from $\{BLEU, NIST, F1\}$; D_{test} is the test split; $\hat{y}_l$ is the human response, and $y_{l,i}$ is one of the generated responses. BLEU and NIST

are sentence-level MT metrics, respectively. F1 is the Unigram F1 score.

4. **Diversity**: We use the system-level diversity metrics, Ent-n [124] and Div-n [20]. Ent-n is the entropy of the n-gram count distribution. Div-n is the number of distinct n-grams in the generated responses divided by the total number of generated token.

Human Evaluation Setup We conducted a qualitative evaluation with human annotators from English-speaking countries using Amazon Mechanical Turk[2]. We randomly sampled 200 examples from the test set of 2,208 instances and asked the distinct human annotators to choose a preferred response in terms of **Relevance** and **Interestingness** among the two randomly ordered responses of a subset of the baseline models and ours. We chose the 5 most competitive models in automatic evaluation metrics, i.e., RAM-T, BART, RagToken, DPRThenPoly, and Human. Additionally, we asked two questions (one for each model) regarding the **Knowledgeableness** to survey whether the responses contain knowledge that does not exist in the conversation history.

Competing Models

- **MemNet** [26]: A Memory Network designed for KGC. The model uses a memory network to store knowledge facts.
- **CMR** [40]: A KGC model based on the state-of-the-art machine reading comprehension model [125]. It is trained with a data weighting scheme to encourage the model to yield responses grounded on the main document.
- **CMR-F** [40]: A model that omits the document reading component of the original CMR model.

[2] `https://www.mturk.com/` (Accessed: July 27th, 2022)

- **RAM-T** [68]: A state-of-the-art model in the CbR dataset. The model is trained to generate the memory to resemble the memory induced by the teacher network, which accepts response, conversation history, and the main document.
- **BART** [3]: A Transformer-based Seq2seq model pre-trained by corrupting text with noising functions and learning to reconstruct the original text.
- **RagToken** [2]: A state-of-the-art model on open-domain QA tasks, which is our base model. A detailed description of this model is presented in Section 4.2.1.
- **DPRThenPoly** [25]: A model that shares the same backbone of RagToken, and re-ranks documents retrieved by DPR using Polyencoder [118].
- Our models: **TADM** denotes our KGC model using the TADM re-ranker. **TADM+IDF-Bw** denotes our model trained with the data weighting scheme based on BLEU and IDF. **TADM+Bw** and **TADM+IDFw** denote our models with the data weighting scheme based on only BLEU and IDF, respectively.

Implementation Details For MemeNet, CMR and CMR-F, we used the implementations provided by [40]. We implemented RAM-T by ourselves because the author did not provide their codes[3]. The above models were trained with hyperparameters that the authors recommended. For Transformer-based models including BART, RagToken, DPRThenPoly and our models, we used a common codebase[4] provided by Shuster et al. [25]. We used the common architectures

[3]Though Tian et al. [68] said that the authors would make the code available to the public, they have not yet uploaded their codes in the repository `https://github.com/tianzhiliang/RAM4CbR`. (Accessed: June 24th, 2022)

[4]`https://github.com/facebookresearch/ParlAI/tree/main/parlai/agents/rag` (Accessed: June 24th, 2022)

and parameters of `bart-large` and `multiset_dpr`. For the Transformer-based models, we used the Adam optimizer for model training, with an initial learning rate of 5e-4 and early-stopped when the perplexity with a validation set did not improve with patience 5 with a 13 batch size. During training, all responses were truncated to have a maximum length of 30 tokens[5], and the maximum length of the context and document were set to 60 tokens and 100 words[6], respectively. For inference, we used top-k (k=10) random sample decoding [126]. We set the number of retrieved documents c to 3 and 5 in the training and testing, respectively. For our model, we use spacy NLP tools for the topic word extraction and salient token checker and set α as 0.5. The aforementioned implementation details can be found in our codes[7].

4.3.2 Experimental Results

Automatic Evaluation Results Table 4.1 demonstrates the results of the automatic evaluation of the Relevance, Ground-to-MainDoc, and Diversity. Our TADM+Bw outperforms the best baseline RAM-T by 0.13% in the F1 score of the Ground-to-MainDoc metric, which matches the human's F1 score (0.88%). For the Relevance metrics, our TADM+Bw improves the baselines' best metrics by 0.01 for NIST, 0.07% for BLEU, and 0.19% for METOR, for which the scores in the NIST and METEOR even showed better results compared to the human's. Regarding diversity, our TADM+Bw improved Ent-4 by 0.07 and Diversity-1 by 0.01. Another noteworthy point is that RagToken and DPRThen-Poly do not show much difference in the Ground-to-MainDoc metric compared to BART, which relies on only the conversation rather showing slight improvements in Relevance, i.e., BLEU.

[5]We processed them with the model's pre-trained tokenizer.
[6]We used space as a delimiter.
[7]`https://github.com/acha21/RAG4KGC-Wild` (Accessed: June 24th, 2022)

Table 4.2 shows that our models produce quality responses in terms of all the Ground-to-Internet metrics with BLUE, NIST, and F1. As we argued in 4.2.3, TADM-IDFw demonstrates better results than the other retrieval-based models', which shows the effectiveness of the data weighting scheme based on IDF.

Model	Ground-to-MainDoc			Relevance			Diversity			Len
	Precision	Recall	F1	NIST	BLEU	METEOR	Ent-4	Div-1	Div-2	
Human	3.47%	0.51%	0.88%	2.65	3.13%	8.31%	10.44	0.17	0.67	18.8
MemNet	0.85%	0.11%	0.20%	2.17	0.98%	7.19%	9.79	0.03	0.21	15.4
CMR-F	0.77%	0.10%	0.18%	2.25	0.98%	7.39%	9.79	0.03	0.20	16.0
CMR	1.15%	0.15%	0.27%	2.26	1.30%	7.44%	9.85	0.04	0.25	15.4
RAM-T	2.17%	0.47%	0.77%	2.08	0.96%	8.58%	10.27	0.03	0.17	28.6
BART	2.29%	0.34%	0.60%	2.82	2.04%	8.68%	10.24	0.10	0.43	18.0
RagToken	2.34%	0.34%	0.59%	2.85	2.27%	8.60%	10.21	0.10	0.45	17.3
DPRThenPoly	2.41%	0.35%	0.62%	2.82	2.40%	8.68%	10.24	0.10	0.46	17.5
TADM+Bw	**3.34%**	**0.52%**	**0.90%**	**2.86**	2.47%	8.87%	10.34	**0.11**	0.46	19.1
TADM+IDFw	2.65%	0.44%	0.75%	2.75	2.12%	**9.08%**	**10.39**	0.10	0.46	20.0
TADM+IDF-Bw	3.17%	0.49%	0.85%	2.84	**2.57%**	8.86%	10.32	**0.11**	**0.47**	19.1

Table 4.1: **Automatic evaluation results in terms of Ground-to-MainDoc, Relevance, and Diversity.** Our best model (TADM+Bw) outperforms the baseline models in terms of grounding considerably, and also improves the Diversity and Relevance slightly. Len denotes the length of the generated responses. Best figures among the baselines are underlined.

Models	kR_{BLEU}	kR_{NIST}	kR_{F1}
BART	11.67%	53.40%	15.30%
RagToken	11.64%	51.80%	14.61%
DPRThanPoly	11.32%	50.27%	14.95%
TADM+Bw	**12.62%**	**62.89%**	**16.09%**
TADM+IDFw	12.29%	58.31%	15.67%
TADM+IDF-Bw	12.37%	57.84%	15.54%

Table 4.2: **Automatic evaluation results in terms of Ground-to-Internet.** Our models outperform the competitive baselines.

Human Evaluation Results The results in terms of Relevance, Interestingness, and Knowledgeableness are summarized in Table 4.3, Table 4.4, and Table 4.5, respectively. Table 4.3 shows that the human annotators judged that our model produces responses relevant to the conversation history more than the other baselines, including the human responses. We hypothesize that this result is because people on the internet do not always reply with highly context-relevant responses; instead, they interact with responses that can fit various conversation contexts, such as "I love it." Table 4.4 shows that our model also outperforms the other models by 10.3% to at least 1.3% in terms of Interestingness. Table 4.5 shows our model produces knowledgeable responses more frequently than the other models by 6.2% to at least 1.9%.

Relevance				
Our best system		Neutral	Comparator	
TADM+Bw	**52.0%**	15.3%	32.7%	RAM-T
TADM+Bw	**42.3%**	26.2%	31.5%	BART
TADM+Bw	**41.2%**	21.5%	37.3%	RagToken
TADM+Bw	**36.3%**	30.7%	33.0%	DPRThenPoly
TADM+Bw	**48.2%**	11.2%	40.7%	Human

Table 4.3: **Human evaluation results in terms of the Relevance**, showing preferences (%) for our model (TADM+BW) vs. the baselines.

Interestingness				
Our best system		Neutral	Comparator	
TADM+Bw	**41.3%**	27.7%	31.0%	RAM-T
TADM+Bw	**39.5%**	30.5%	30.0%	BART
TADM+Bw	**40.5%**	20.3%	39.2%	RagToken
TADM+Bw	**35.2%**	32.3%	32.5%	DPRThenPoly
TADM+Bw	41.2%	16.7%	**42.2%**	Human

Table 4.4: **Human evaluation results in terms of Interestingness**, showing preferences (%) for our model (TADM+BW) vs. the baselines.

Knowledgeableness			
Our best system		Comparator	
TADM+Bw	**48.5%**	42.3%	RAM-T
TADM+Bw	**53.0%**	46.8%	BART
TADM+Bw	**54.7%**	52.0%	RagToken
TADM+Bw	**49.2%**	47.3%	DPRThenPoly
TADM+Bw	48.5%	**49.7%**	Human

Table 4.5: **Human evaluation results in terms of Knowledgeableness**, showing the percentage (%) of responses that human annotators considered knowledgeable.

4.4 Analysis

4.4.1 Case Study

Figure 4.5 shows the models' outputs for a randomly selected example. The example is about a battle in Afghanistan. RAM-T, RAGToken, DPRThenPoly, and TADM+Bw produce responses relevant to the given conversation history. Among the above four models, our model TADM+Bw outputs a response including the word 'science,' making us think that the model knows more about the context. Looking up relevant words in the top-1 document, we can find the words 'planes' and 'aircraft.' This is evidence that our model finds relevant documents and uses some information from these documents. Additionally, our model appears to find the context-relevant document in the top-4 document.

To help readers understand our model's properties in detail, we present more examples in Figure 4.6, Figure 4.7, and Figure 4.8. In the examples, we highlighted the relevant parts in our model's response and retrieved documents. The examples confirm to us that our models can generate responses relevant to the given conversation history and the retrieved documents.

4.4.2 Ablation Study

Table 4.6 shows the effects of removing each module from our best model. Recall@5 is the accuracy of the retriever when we assume the GT knowledge document is the main document of a conversation. The results show that Recall@5 is proportional to Ground-to-MainDoc, of which the Pearson coefficient is 0.94. Our data weighting scheme appears essential for training the model end-to-end in the CbR task because Recall@5 gets almost zeros without the weighting scheme. Utilizing the topic keywords and dual matching improves Recall@5 by 8.84% and 4.09%, respectively, Grounding metric F1 improved by 0.09% and 0.03% as well. In terms of Relevance, the BLEU score decreases by 0.32% without the data weighting scheme. In addition, the model without dual-matching improves Ent-4. We presume that it is the result of utilizing the contents of other documents rather than the main document of the conversations. Table 4.7 shows the results of the ablation of representations on top of tokens that the salient token checker selected using matching MLST. Without representations of the non-stop words, the model can retrieve the main documents better than using all union of ones with the non-stop words, named entity, and keywords extracted with TextRank (All), but the Relevance metrics such as BLEU and METEOR decrease.

Conversation history	til soldiers in the u . s . military ' s psychological warfare department used loud speakers to tempt enemy fighters in afghanistan into battles they couldn ' t wi n by calling them , " lady men " and " cowardly dogs " .
GT Knowledge	<title> psychological operations </title> psychological operations <anchor> afghanistan burning bodies incident </anchor> [edit] < p> on 1 october 2005 in gumbad , afghanistan ,soldiers from the 173rd airbor ne decided to burn the bodies of two taliban fighters killed in a firefight the pr evious day for hygienic reasons . despite islamic customs that forbid cremation
MemNet	" we need to get the same " . it was not a good deal . the guy from the other h and and the us , but they just don't
CMR-F	i was hoping it was a good idea for the rest of reddit i guess .
CMR	if they had a chance to kill the other side of the world .
RAM-T	the soviets killed a dog for <NUM> years . they were only on the ground that w as killed . it was also a
BART	and they called me the bitch .
RagToken	" i'm the guy that got shot , so i have to tell you to get your balls out ... " " i'm g onna tell you i have to ,
DPRThenPoly	we can't win a battle against the man ! " " then you should go back and fuck hi m ! "
TADM + Bw	you're so wrong . i'm sure this was done in the name of science .
Top-1 Retrieved doc	Operation Moolah | North Korea, though internationally their involvement wa s never announced. The Soviets had gone to great lengths to hide their involve ment in the war, including painting Chinese and North Korean insignia on their planes. By the end of the war, the Russians had provided half the aircraft and 5,000 pilots in support of the Communist effort against the U.N. Speculation ex ists about the origin of Operation Moolah. …
…	…
Top-4 Retrieved doc	since U.S. combat operations in Afghanistan began. Operation Mountain Lion began 15 April 2002 and involved Afghan National Army and US and Coalition Forces performing search operations in the Gardez and Khost regions. Significa nt participation by the Royal Marines in this operation was known as Operatio n Jacana. Operation Snipe began in May 2002 to search and clear a significant area in the remote

Figure 4.5: **Sample output** for a given conversation history. It compares the generated response of the competing models with our model TADM+Bw, where we assume that the GT knowledge document is a document concerning the conversation topic. We present two documents manually chosen as relevant to the responses (Top-1 and Top4) from documents retrieved in the top-5 documents.

Conversation history	START EOS til : the vietnamese are the biggest consumers of the rhino horn and they received the idea that rhino horn was an aphrodisiac from the western media's false reporting . there is no known record of its use as an aphrodisiac in traditional chinese medicine .
GT Knowledge	<title> rhinoceros </title> <h1> rhinoceros </h1> <h2> <anchor> horn trade and use </anchor> </h2> <p> rhinoceros horns , unlike those of other horned mammals , only consist of keratin , similar to human hair and nails . rhinoceros horns are used in traditional medicines in parts of asia , and for dagger handles in yemen and oman . esmond bradley martin has <TRNC>
MemNet	and now they have to keep you on the same way .
CMR-F	they do have a great day . that's why i was <NUM> % sure they did have more than that to make you a
CMR	this is why we can eat all the things that would come in .
RAM-T	there would be a lot of genetic advances between the other things , the rest is not the result to humans , it's about <NUM> %
BART	i've never heard of anyone using it for any sort of health or medical benefit , except in a certain type of medicine .
RagToken	it was a popular aphrodisiac for centuries in chinese medicine . they use it in chinese medicine all over the world .
DPRThenPoly	what if i am in the same sex with the rhino and it's my horn ? would you let me have it ?
TADM + Bw	i was under the impression this was a cure for cancer , and that rhinoceros horn is a powerful aphrodisiac . now i think back about this and feel like
Top-1 Retrieved doc	Rhino poaching in Assam | <TRNC>, it is kept by wealthy people in Vietnam as a \"peace of mind\" cure. In 1993, rhino horn was removed from the official lists of Traditional Chinese Medicine. It is now only sold in Vietnam following an unsubstantiated rumour that horn cured a high-ranking official in Vietnam of cancer. Aaranyak, a society for biodiversity conservation, on Friday urged the Assam government not to <TRNC>
...	
Top-4 Retrieved doc	Ben Sakoguchi | Japanese Americans. \"Aphrodisiac Brand\", in the collection of the Honolulu Museum of Art, is a mock orange-crate label. Is shows a rhinoceros slaughtered for its horn, which is erroneously believed to be an aphrodisiac or a cure for cancer in Traditional Chinese Medicine as \"Cornu Rhinoceri Asiatici

Figure 4.6: **Sample output** for a given conversation history. It compares the generated response of the competing models with our model TADM+Bw, where we assume that the GT knowledge document is a document concerning the conversation topic. We present two documents manually chosen as relevant to the responses (Top-1 and Top4) from documents retrieved in the top-5 documents.

Conversation history	START EOS til that medal of honor recipient thomas baker asked to left behind during a japanese counter attack on saipan due to being wounded . he only asked for 1 pistol with 8 bullets . after returning the us army found his corpse with 8 dead japanese .
GT Knowledge	doc<title> thomas baker </title> thomas baker <anchor> medal of honor citation </anchor> [edit] for conspicuous gallantry and intrepidity at the risk of his life above and beyond the call of duty at saipan , mariana islands , 19 june to 7 july 1944 . when his entire company was held up by fire from automatic weapons and small-arms fire from strongly fortified enemy positions that commanded the view of the company , sgt . baker voluntarily took a bazooka and dashed alone to within 100 yards of the enemy . <TRNC>
MemNet	i don't think anyone wants to know what he means .
CMR-F	i guess i saw this post in my book <NUM> years ago . i had a feeling of the show , i was just trying to go through a movie
CMR	" he must be in love for the nazis . "
RAM-T	the story is misleading to the article that was in the first place , which he didn't die to die in the air , and then the
BART	what was i thinking ? a medal of honor and a gun ?
RagToken	i'm surprised that he was not in the front line or anything .
DPRThenPoly	he must have wanted a new gun so badly he would give up his life for it .
TADM + IDF-Bw	i don't know how this happened , but he was shot through the chest , but the japanese soldiers still shot him .
Top-1 Retrieved doc	Hideki Tojo | Japan's unconditional surrender in 1945, U.S. general Douglas MacArthur ordered the arrest of forty alleged war criminals, including Tojo. Five American GIs were sent to serve the arrest warrant. As American soldiers surrounded Hideki's house on September 11, he shot himself in the chest with a pistol but missed his heart. As a result of this experience, the Army had medical personnel present <TRNC>
...	
Top-3 Retrieved doc	105th Infantry Regiment (United States) | found the pistol, then empty, and eight dead Japanese soldiers around Baker's body. PVT Baker received a posthumous Medal of Honor. CPT Ben L. Salomon was the 105th's regimental dentist officer but on Saipan he <TRUN>

Figure 4.7: **Sample output** for a given conversation history. It compares the generated response of the competing models with our model TADM+IDF-Bw, where we assume that the GT knowledge document is a document concerning the conversation topic. We present two documents manually chosen as relevant to the responses (Top-1 and Top3) from documents retrieved in the top-5 documents.

Conversation history	START EOS til that thalidomide , the infamous morning sickness drug that caused severe birth defects , was never approved for use in the us because of a single reviewer at the fda <TRUN> people give the fda shit for not approving drugs that they assume will be beneficial . but if the fda approves a drug that turns out to be harmful , everybody is mad . it ain't easy being a government agency !
GT Knowledge	<title> frances oldham kelsey </title> <h1> frances oldham kelsey </h1> <h2> <anchor> work at the fda and thalidomide </anchor> [edit] </h2> <p> in 196 0 , kelsey was hired by the fda in washington , d . c . at that time , she \" was o ne of only seven full-time and four young part-time physicians reviewing drugs \" [3] for the fda . one of her first assignments at the fda was to review an ap plication by richardson merrell <TRNC>
MemNet	so they get a lot of a lot of people from the us .
CMR-F	you are a racist .
CMR	i'm just going to guess that you can get the idea that if you read that correctly , it's not that the difference is a cure .
RAM-T	it's not really that hard to believe it would be more important because it is leg al to the us . also , the amount of people who have the
BART	you mean they don't give a shit about whether or not a drug is beneficial .
RagToken	you could say the same of the us military and the us government . the same thi ng happens .
DPRThenPoly	the FDA is a private company and not regulated by any government
TADM + IDF-Bw	it's the fda's job to make sure that the products that are being marketed are safe and effective . and it is . if they fail to do so ,
Top-1 Retrieved doc	SIGA Technologies | SIGA Technologies SIGA Technologies, Inc. () is a pharmaceutical company set up in 1995 that is now based in New York City. The company markets pharmaceutical solutions for smallpox, Ebola, dengue, Lassa fever, and other lethal disease-causing <TRNC>
...	
Top-5 Retrieved doc	History of the Food and Drug Administration | infancy at the start of this era, rapid advances in experimental assays for food additive and drug safety testing were made during this period by FDA regulators and others. In 1959, Senator <TRUN> uncertain efficacy of many drugs promoted by manufacturers. There was significant opposition, however, to <TRUN>

Figure 4.8: **Sample output** for a given conversation history. It compares the generated response of the competing models with our model TADM+IDF-Bw, where we assume that the GT knowledge document is a document concerning the conversation topic. We present two documents manually chosen as relevant to the responses (Top-1 and Top5) from documents retrieved in the top-5 documents.

Model	Retrieval	G-to-M	Relevance			Diversity	
	Recall@5	F1	NIST	BLEU	METEOR	Ent-4	Div-2
TADM+Bw	**14.07%**	**0.90%**	2.86	**2.47%**	8.87%	10.34	**0.46**
W/O Topic keywords	5.23%	0.81%	2.83	2.46%	8.95%	10.35	**0.46**
W/O Dual matching	9.98%	0.87%	**2.87**	2.42%	**9.06%**	**10.37**	**0.46**
W/O data weighting	0.10%	0.63%	2.85	2.15%	8.57%	10.22	**0.46**

Table 4.6: **Results of the ablation study.** W/O Topic keywords denotes our model that does not use topic keywords as the conversation encoder input; W/O Dual matching denotes our model that has a single matching layer; W/O data weighting denotes our model trained without our data weighting scheme. G-to-M stands for Ground-to-MainDoc.

Models	Retrieval	G-to-M	Relevance			Diversity	
	Recall@5	F1	NIST	BLEU	METEOR	Ent-4	Div-2
All	12.68%	**0.89%**	**2.89**	**2.70%**	**8.93%**	10.31	0.464
W/O non-stop words	**13.79%**	0.88%	2.83	2.57%	8.91%	10.32	0.459
W/O named entity	13.43%	0.88%	2.84	2.52%	8.81%	**10.33**	0.458
W/O TextRank	11.63%	0.85%	**2.89**	2.54%	8.67%	10.28	**0.474**

Table 4.7: Comparison of TADM+IDF-Bw's performance according to representation on top of different token types for MLST. NE represents named entity.

4.4.3 Model Variations

To measure the importance of the components of our re-ranker, we change the configurations of our TADM+IDF-Bw model. Table 4.8 shows the trend that as the number of shallow layers decreases, the retrieval accuracy increases except when the number of layers is 4, somewhat supporting our decision to utilize the small number of layers for the dual matching. Table 4.9 shows the performance change depending on the number of representations used as input of MLLC. The results show that using an appropriate number of representations is vital for finding the best retrieval accuracy, grounding, and relevance metric. Table 4.10 compares the models' performance according to whether to use representations on top of the first N tokens or the last N tokens as the MLLC input, which is the basis of our design choice.

Models	Retrieval	G-to-M	Relevance			Diversity	
	Recall@5	F1	NIST	BLEU	METEOR	Ent-4	Div-2
# layers = 1	**13.33%**	0.85%	**2.94**	2.69%	8.82%	10.27	**0.474**
# layers = 2	12.68%	0.89%	2.89	**2.70%**	**8.93%**	**10.31**	0.464
# layers = 3	9.91%	0.82%	2.89	2.37%	8.66%	10.27	0.471
# layers = 4	10.69%	**0.90%**	2.87	2.54%	8.87%	**10.31**	0.473

Table 4.8: Comparisons of TADM+IDF-Bw using different dual matching layers. We report all metrics on the 2208 testset [4]

Models	Retrieval	G-to-M	Relevance			Diversity	
	Recall@5	F1	NIST	BLEU	METEOR	Ent-4	Div-2
# FirstToken = 1	11.32%	0.87%	**2.91**	2.31%	**8.97%**	**10.30**	0.470
# FirstToken = 4	**14.84%**	**0.89%**	2.87	2.39%	8.69%	10.29	**0.475**
# FirstToken = 8	11.26%	0.84%	**2.91**	**2.56%**	8.80%	10.28	0.474

Table 4.9: Comparison of TADM+IDF-Bw's performance according to the number of representations on top of the first token used in MLLC

Models	Retrieval	G-to-M	Relevance			Diversity	
	Recall@5	F1	NIST	BLEU	METEOR	Ent-4	Div-2
First 8 Tokens	**12.68%**	**0.89%**	**2.89**	**2.70%**	8.93%	10.31	0.464
Last 8 Tokens	7.73%	0.84%	2.86	2.60%	**8.94%**	**10.33**	**0.465**

Table 4.10: Comparison of TADM+IDF-Bw's performance according to whether to use representations on top of the first N tokens or the last N tokens

4.4.4 Error Analysis

To grasp the quality of the responses generated by the models and examine the errors in the responses more specifically, we inspect the 11 randomly sampled generated responses of TADM+IDF-BW and DPRThanPoly by answering some questions about metrics. Table 4.11 shows the qualitative results of the examination. Questions that correspond to the quality of the general aspect of the response include the existence of contradiction (Q2 and Q2-1), grammaticality (Q1), and relevance (Q4). The metrics that show the two highest differences are Relevance and Ground-to-MainDoc. This result is consistent with the above human and automatic evaluation results. Though metrics on grammar and co-

herence are lower than that of DPRThenPoly, this does not appear to have a considerable impact on relevance and groundness.

We present some of the failure cases of our models to grasp the properties of our models in detail. Figure 4.9 shows an example where some of the words in the quotations ("in their 20s") may be missing so that the response does not make sense. Figure 4.10 demonstrates an example where our model TADM+BW generates responses similar to the context, which can be regarded as not an interesting response. Figure 4.11 shows an example where the term '1800' in the response is not consistent with the main document.

Questions	Related Metrics	TADM+IDF-Bw	DPRThenPoly	Average
Q1. Does it contain not finished sentence?	Grammar	**13.6**%	18.2%	15.9%
Q2. Does it make sense?	Coherence within a response*	77.3%	**81.8%**	79.6%
Q2-1. Number of words that makes the response incoherent	Coherence within a response*	0.64	**0.73**	0.69
Q3. Is it appropriate response to context?	Appropriateness	**63.6**%	54.5%	59.1%
Q4. Does it make sense in the context of the conversation?	Relevance	**81.8%**	45.5%	63.7%
Q5. Can it be inferred by the main document?	Ground-to-MainDoc	**36.4%**	18.2%	27.3%
Q6. Does it contain personal opinion?	Factual correctness*	81.8%	81.8%	81.8%
Q7. Dose it contain interesting words?	Interestingness	**18.2**%	9.1%	13.7%
Q8. Does it contain non-informative statement?	Informativeness	4.6%	**0.0**%	2.3%

Table 4.11: Error case analysis of TADM+IDF-Bw and DPRThenPoly. If the item in the first column is in the form of a question, then the result is the proportion of answers as Yes; else is the Figure according to the item. * indicates that the metric is defined in Table 2.4

Conversation history	START EOS til - men in their 20s that ejaculate often may impart protections ag ainst prostate cancer later in life .
Main document of the conversation	<title> ejaculation </title> <h1> ejaculation </h1> <h2> <anchor> health issues </anchor> </h2> <p> for most men , no detrimental health effects have been determined from ejaculation itself or from frequent ejaculations , [citation ne eded] though sexual activity in general can have health or
MemNet	that doesn't really change that for it .
CMR-F	i have a feeling that there is a lot of people in the us . i had no idea what happ ened to me . the fact that he is being
CMR	and it's so easy to know why this is a thing .
RAM-T	i don't think the problem is that the vast majority of people who have been dia gnosed with them . they would think the pain is a very
BART	so ... it doesn't work that way for the men i've known ?
RagToken	i am in my 20 ' s and have already been diagnosed . i'm not sure it's a good thi ng , but i don't know if i have prostate cancer
DPRThenPoly	i had a friend who would ejaculate every morning and i thought it was cool . tu rns out he's a cancer patient .
TADM + Bw	" in their 20s " is a pretty broad term for the early stages of prostate cancer . i have friends that were only diagnosed with it when they were 30
Top-1 Retrieved doc	Tubal factor infertility | not be recognized for years after a gonorrhea, Chlamy dia or \"Mycoplasma\" infection has caused tubal damage, as the affected wo man may not have attempted to become pregnant until years later. Infertility i s the major symptom of TFI and is generally defined as a woman under 35 who has not become pregnant after 12 months without the use of contraception. T welve months is the lower reference limit for \"Time to Pregnancy\" (TTP) by t he World Health Organization.
...	
Top-5 Retrieved doc	BRCA mutation | to bear children). Infertility services can be used to preserve her eggs, if wanted. However, as the benefits to the surgery are greatest close to menopause, most women simply postpone the surgery until they have alrea dy borne as many children as they choose to. The surgery also artificially induc es menopause, which causes hot flashes, sleep disturbances, mood swings, va ginal dryness, sexual difficulties,

Figure 4.9: A example of output of TADM+BW and baselines

Conversation history	START EOS til fight club author chuck palahniuk lives with his male partner of 2 0 years
Main document of the conversation	<title> chuck palahniuk </title> <h1> chuck palahniuk </h1> <h2> <anchor> pe rsonal life </anchor> [edit] </h2> <p> as an adult , palahniuk became a mem ber of the rebellious cacophony society . he is a regular participant in their eve nts , including the annual santa rampage in portland , oregon . his participatio n in the society inspired some of the events in his writings ...
MemNet	i love how it took me a lot to me , i'm not gonna lie .
CMR-F	my dad died for <NUM> years old . i don't know . they were so much different from the rest of the world , the
CMR	what is this , you can't just sit there ?
RAM-T	he may also have a lot of people in his <NUM>s . they just had to be on his ow n and he died . he was in the
BART	it's a good man and a great writer .
RagToken	he lives with someone . that is a good thing .
DPRThenPoly	it's not uncommon for writers to get " divorced " from their partner .
TADM + Bw	til fight club author chuck palahniuk lives with his male partner
Top-1 Retrieved doc	Danny Downes | Danny Downes Daniel William Downes (born April 8, 1986) is a retired American mixed martial artist. Even after his retirement, Downes has still continued to stick around the sport of MMA. He became what some call a \"fighter turned writer\", as he now writes for various MMA media outlets, inc luding MMAjunkie.com Since turning professional in 2007, Downes has compe ted for several smaller organizations, while compiling an undefeated record an d obtaining
...	
Top-5 Retrieved doc	Eric Klopfer | in science and math - connecting game play with scientific practi ce, problem-solving, and real world issues. In 2014, Klopfer produced a series of four online courses on the edX learning platform, delivering videos, tools, a nd assignments for participants to create, implement, and evaluate projects in educational technology

Figure 4.10: A example of output of TADM+BW and baselines

Conversation history	START EOS til that a da vinci painting " madonna of the yarwinder " was stolen from drumlanrig castle by two thieves posing as tourists who claimed to be un dercover police . it was found four years later in a lawyer ’ s office in glasgow .
Main document of the conversation	<title> madonna of the yarnwinder </title> <h1> madonna of the yarnwinder </h1> <h3> <anchor> theft and recovery </anchor> [edit] </h3> <p> in 2003 t he buccleuch madonna was stolen from drumlanrig castle by two thieves posi ng as tourists , who said \" don't worry love , we're the police . this is just pract ice \" to two tourists from new zealand as they exited through a window carryi ng the leonardo . [16] in 2007
MemNet	what about their own ?
CMR-F	that was the first time i've seen it on the <NUM>th season , and i am not going to have a good understanding of him .
CMR	it doesn't matter .
RAM-T	the only thing that would have been , but it seems to be an easy decision to m ake money . also , the only thing that came
BART	i have never heard this tale before ...
RagToken	" i'm not a criminal . i just want to make a quick buck . "
DPRThenPoly	i think i saw the movie that has this story in it somewhere on tv
TADM + IDF-Bw	i'm not sure if it is true , but in the late 1800s this painting was stolen from a h ouse in drumlanrig , near mexico . this painting
Top-1 Retrieved doc	Terringzean Castle, East Ayrshire | The Prince of Wales, including various herit age charities and the Scottish Government, purchased from the 7th Marquis, D umfries House, along with its roughly 2,000-acre (8.1 km2) estate). In 1587 He w Campbell of Terringane (sic) and others were appointed by parliament to 'vi sie' or inspect the bridges at Irvine and Ayr; in 1595 he was appointed to inspe ct the port and harbour
...	
Top-5 Retrieved doc	Summerhill House | was given to Robert Fowler who was the Master of the M eath Hounds at the time of her stay in Summerhill. The whip had been lost and had been found not long before the auction in Rahinston House. The whip was found in a mahogany presentation case with a silver crest plate bearing the Im perial Arms of Habsburg.

Figure 4.11: A example of output of TADM+IDF-BW and baselines

Chapter 5

Application: Quote Recommendation with Knowledge Ranking

5.1 Motivation

People are currently facing a significant transition era while experiencing a pandemic of COVID-19, which makes some offline activities, such as work, and education, transfer to online. Henceforth, we increasingly rely on digital devices for communications with humans, including text messages, tweets, emails, and blogs. In other words, we take part in a kind of conversation on digital devices every day.

Citing quotations such as proverbs and (famous) statements of other people in conversations can provide support, shed new perspectives, and/or add humor to one's arguments. However, it is hard for humans to come up with the right search keywords to retrieve appropriate quotes because the words in quotes have different meanings or are metaphorical from words' of our life. It will be ben-

eficial if a practical quote recommendation system provides useful quotations for a given conversation context instead of making them explore websites such as BraninyQuotes[1]. Even for a machine, recommending appropriate quotes can be challenging due to the same reason.

Prior methods for recommending quotes can be categorized into ranking-based [81, 82, 36] and generation-based methods [83, 127]. Ranking-based methods usually exploit features extracted solely from context, quote, and both context and quote, which we will call context-quote features for features from the last. The recent work [82] showed that combining context features and context-quote features with the explicit transformation between sentence-level semantic space of context and quote can achieve state-of-the-art performance. Specifically, the model learns the relationship between quotations and query turns by adopting a transformation matrix to consider context-quote features.

In this chapter, we attempt to fine-grained match between conversation and quotation using multiple representations rather than only a single sentence-level representation from the encoders to go one step further. Specifically, we study and verify the applicability of our knowledge re-ranker module TADM presented in Chapter 4 in recommending quotations for conversation. As shown in Figure 5.1, metaphorical words in quotes can correspond to words in conversations, which humans intuitively may figure out. Our matching-based knowledge selection model may learn this correspondence since it attempts to utilize representations originating from words.

For this purpose, we propose a new quote recommendation framework called ClAssification-based candidate Generation And Re-ranking (CAGAR) to adopt our knowledge ranking method. The framework consists of candidate generation and re-ranker modules, which are trained simultaneously. The candidate

[1] `https://www.brainyquote.com/` (Accessed: June 24th, 2022)

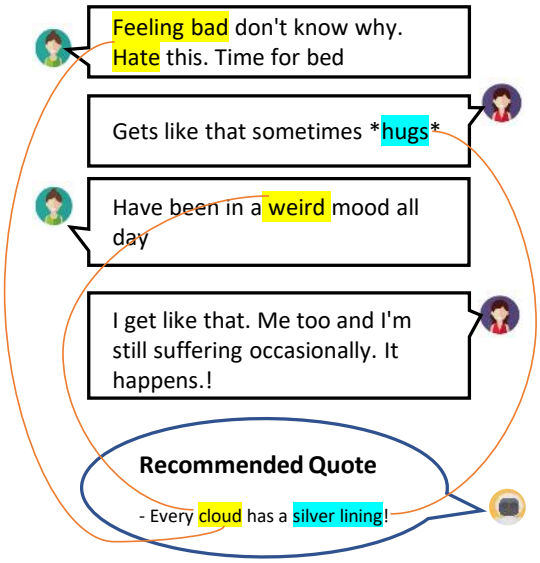

Figure 5.1: **Quote recommendation in conversation**. The quote recommendation system can recommend an appropriate quote for a given conversation history. The figure highlights the words where words in quotes correspond to ones in the conversation history. This chapter aims to learn the correspondences by utilizing our knowledge selection model.

generation modules yield a list of candidate quotes recognized as suitable for the context using the single representations of the quotes and conversations and transferring them to the re-ranker. Then, the re-ranker module finally produces the score of the quotes from candidate generation modules. Note that the candidate quotes contain appropriate quotes, but not GT ones, which are used as hard negative examples for the re-ranker module.

5.2 CAGAR: A Framework for Quote Recommendation

Figure 5.2 shows the overall structure of our quote recommendation model, CAGAR, which comprises a conversation encoder, a quote encoder, a candi-

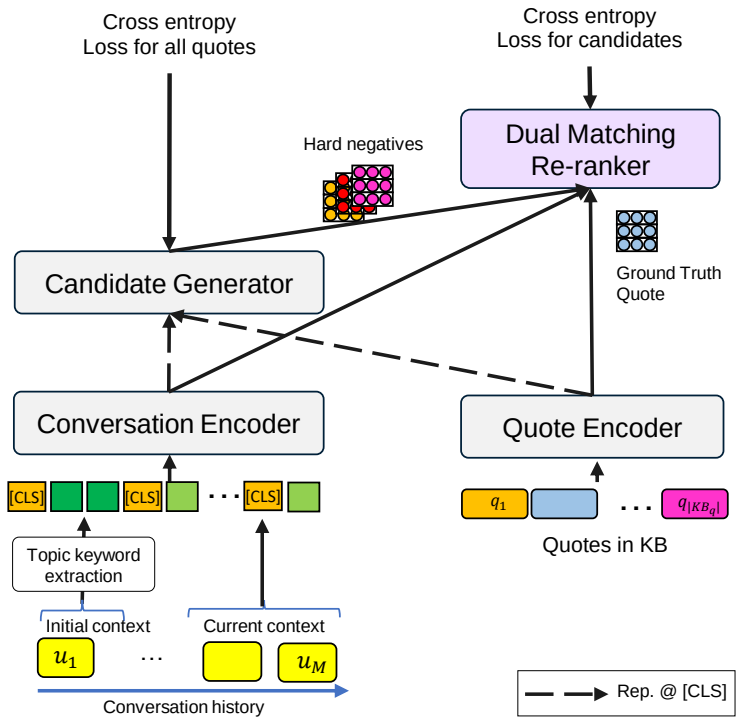

Figure 5.2: **Our quote recommendation model**. The candidate generator and dual-matching re-ranker are jointly trained to predict appropriate quotes for a given conversation history.

date generator, and a re-ranker. Although we use our dual matching knowledge re-ranker as a re-ranker module, any neural re-ranking methods that can compare the two text inputs can be adopted in the CAGAR framework. Our model outputs a ranked list of labels ys indicating quotes qs from a candidate quote set $\text{KB}_q = \{q_1, q_2, ..., q_{|KB_q|}\}$ for a given conversation history c. The conversation history c is a sequence of turns (e.g., comments or tweet) $\{u_1, u_2, ...u_{n_c}\}$, where u_i represents the i-th turn of the conversation and contains words $(w^c_1, w^c_2, ..., w^c_{|u_i|})$. Each quote q_m is a sequence of tokens $(w^q_1, w^q_2, ..., w^q_{|q_m|})$. In the following, we explain each module in detail.

5.2.1 Conversation Encoder

We use DistillBERT [128] for conversation modeling. DistillBERT is a pre-trained Transformer encoder retaining BERT's language understanding capabilities with a light model size. DistillBERT is known that it can reduce the size of a BERT model by 40% while retaining 97% of BERT's performance. The input of our conversation encoder is a concatenation of topic keywords of the conversation history and local context, where each of them starts with the [CLS] token. The concatenation of tokens in the input conversation can allow the encoder to conduct self-attention between the tokens in different turns, considerably enhancing performance. The form of input is similar to the method described in Section 4.2.2 where the topic keywords are a sequence of tokens ([CLS], $w_1^k, ..., w_{p-1}^k$) extracted from the initial turns of the conversation history. The local context is a sequence of the latest $L - p$ tokens starting with [CLS], where L is the maximum input size of DistillBERT. We use the latest turns with [CLS] delimiter before the response for the local context. We assign different segment embeddings for the local context and topic keywords. Finally, conversation encoder produces representation $\{\boldsymbol{r}_1^c, ..., \boldsymbol{r}_L^c\}$ for conversation history c.

5.2.2 Quote Encoder

The quote encoder processes all quotes in the quote KB using a 2-layer Transformer encoder. The input to this module is a sequence of tokens appended [CLS] to each quote. We do not share the token/segment/position embeddings with the conversation encoder since the words in conversation and quote can have different meanings. Also, the words' meaning in quotes can differ from the corpus used in pre-trained models such as BERT. For this reason, we determine to train the encoder from scratch without using an encoder pre-trained with a

web corpus that usually contains metaphoric words rarely. The quote encoder finally outputs a set of representations $\{\boldsymbol{r}_1^q, ..., \boldsymbol{r}_{|q_m|}^q\}$.

5.2.3 Candidate Generator

Our candidate generator yields top-k non-GT quotes and the scores indicating how much the quote fits the given conversation to provide them as hard negative examples to our re-ranker. To this end, the candidate generator is trained to predict the probability of recommending the GT quote $\hat{q}$ using a single representation of [CLS] of inputs, which is the same method with the Bi-encoder [118]. Following the previous work [82] that uses features from context, quote, and both context and quote, we first form a feature vector $\boldsymbol{v}$ as follows:

$$\boldsymbol{v} = [\boldsymbol{s}; \boldsymbol{r}_1^c; P(\boldsymbol{r}_f^c)] \tag{5.1}$$

where $\boldsymbol{s}$ is $Q \times P(\boldsymbol{r}_f^c)$, Q is quote representation matrix of which row represents representation $\mathbf{r}^q$ at [CLS] token and P is projection function. Regarding the context features, we preliminarily confirmed that features from the conversation solely could enhance the model's performance through experiments[2].

Then, we define the final scores for all the quotes as

$$p_{cg}(q = i) = \frac{exp(y_i)}{\sum_{k=1}^{n_q} exp(y_k)}, \tag{5.2}$$

where $\boldsymbol{y} = W\boldsymbol{v} + \mathbf{b}$. Here W and $\boldsymbol{b}$ are learnable parameters. Our candidate generator outputs the best non-GT quotes according to p_{cg} and gives the re-ranker them.

Interestingly, our method can be seen as knowledge distillation which transfers the results of light layers to heavy layers as negative examples. Its direction

[2]We provide our preliminary experimental results using various ranking-based methods in Appendix A

of transferring knowledge is the opposite of knowledge distillation's direction, which transfers knowledge from a large model to a smaller one.

5.2.4 Re-ranker

Recently, Wang et al. [82] presented a visualization that shows evidence that the Transformer encoder can identify relevant words via the self-attention score of relevant tokens. This visualization suggests that the method utilizing sentence-level representation for measuring similarity between conversation and quote sentences can learn the relationship between words to some extent.

Inspired by the above facts, we attempt to learn correspondences between words in the context and ones in quotes by considering the interaction between representations focusing on the words that have the potential to be matched to the others. To this end, we adopt the TADM module as a re-ranker without any modifications. Specifically, our TADM collects and matches the representations corresponding to salient keywords from the conversation encoder and quote encoder using a MLST, which is a 1-layer Transformer. The re-ranker finally outputs the conditional probability $p_{rr}(\hat{q}|c, \mathbf{q}_{neg}, \hat{q})$. While applying the TADM re-ranker to our CAGAR framework, what differs from the model in Chapter 4 is that we can use cross-entropy loss because the GT quotes are available.

5.2.5 Training and Inference

We train the model end-to-end for the candidate generator and re-ranker to be enhanced simultaneously. The training objective of our model is as follows:

$$\mathcal{L} = -\log p_{cg}(\hat{q}|c, KB_q) - \boldsymbol{\lambda} \log p_{rr}(\hat{q}|c, \mathbf{q}_{neg}, \hat{q}) \tag{5.3}$$

where $\boldsymbol{\lambda}$ is trainable weight. For inference, we calculate all the re-ranking scores of the quotes and add them to p_{cg} to get the final scores for all the quotes.

5.3 Experiments

5.3.1 Experimental Setup

Datasets We use Reddit [83] and Twitter [36] datasets. Both datasets contain English conversations containing quotes and the quote database, where conversation history is texts appearing before a quote. On Reddit data, a comment corresponds to a turn of conversation. On Twitter dataset, a tweet is considered the turn of conversation. Table 5.1 and Table 5.2 show the statistics of the two datasets.

Items	Reddit	Twitter
# of quotes	1,111	400
# of conversations	44,539	222,408
avg. turn # per conversation	4.25	3.89
avg. # of words per turn	71.8	13.7

Table 5.1: **Statistics of datasets**. "avg." refers to average. # represents number.

# of Turns in Context	1	2	3	4	5	6	7	8	9	10	11
Reddit	3	1356	986	617	420	308	210	144	102	0	0
Twitter	3336	6781	3505	2250	1315	965	627	499	421	739	636

Table 5.2: **Number of examples per the number of turns in the context of the test sets**

Experiment Setting We initialize the conversation encoder with the parameters of `distilbert-base-uncased` which Huggingface[3] officially provides. For the Reddit and Twitter datasets, we use a Transformer quote encoder of 3-layers, 200 hidden units, two attention heads, and 2,048 hidden units of position-wise feed-forward network with embedding layer initialized with 200-

[3] `https://huggingface.co/` (Accessed: May 15th, 2022)

dimensional Glove embeddings [129]. We train all models using Adam optimizer [113] with an initial learning rate of 1e-4 and early stop using loss on the validation set with the patience of 5. The batch size is set to 32. We use dropout with 0.2 probability and L2 regularization of 3e-4 weight. For the baselines, we choose their hyperparameters according to the authors' recommendation.

Evaluation Metrics To evaluate the models, we use MAP (Mean Average Precision), P@1 (Precison@1), P@3 (Precison@3), and NG@5 (normalized DiscountedCumulative Gain@5).

1) **P@k:** The precision at k is the proportion of examples whose GT quotes appeared in respective top-k ranked quotes. The P@k for a test set D_{test} can be computed by

$$P@k = \frac{\text{\# of GT quotes in top-}k\text{ results}}{|D_{test}|} \tag{5.4}$$

2) **MAP:** MAP with the test set D_{test} is calculated by

$$MAP(Q) = \frac{1}{|D_{test}|} \sum_{j=1}^{|D_{test}|} Precision(R_j) \tag{5.5}$$

where R_j is the set of ranked quotes from the top results until you get to GT quote q_j and $Precision$ is one is divided by rank of the GT quote in the R_j.

3) **NDCG@k:** NDCG@k for the test set can be calculated as:

$$NDCG@k = \frac{1}{|D_{test}|} \sum_{e \in D_{test}} NDCG(e, k) \tag{5.6}$$

where $NDCG(c, k)$ is calculated as:

$$NDCG(e, k) = Z_k \sum_{i=1}^{k} \frac{2^{r(e,i)} - 1}{log(i + 1)} \tag{5.7}$$

Here i is the rank of the i-th candidate quote in the ranked list, $r(e, i) = 1$ if the i-th quote is the GT quote of a test instance e and $i < k$, otherwise $r(e, i) = 0$.

Z_k is a normalization constant to make the perfect ranked list get a NDCG score of 1.

Baselines We compare our model with the following comparative quote recommendation models for conversation.

1) **BERT**: A model that encodes the conversation history by BiLSTM on top of BERT representations followed by a prediction layer. We determine this architecture to allow the model to accept all the turns in the conversation.

2) **RBT**: The state-of-the-art model [82] in the Reddit dataset and Weibo dataset [83]. The model introduces a transformation matrix mapping the query representations to quotation representations.

3) **RBT+DistillBERT**: RBT, which uses our conversation encoder that does not use the topic keywords. We use a vector from the encoder corresponding to the first [CLS] to represent the conversation history. As a representation of the last turn, we use the encoder's output corresponding to [SEP] before the turn.

5.3.2 Experimental Results

Main results Table 5.3 shows comparison results on Reddit and Twitter datasets. From the table, we can observe several interesting facts. First, our model outperforms the baselines by a large margin, which means that the combination of our hard negative example generator and re-ranker enhances the model that depends on only the single sentence-level representation. Second, adopting our conversation encoder with the pre-trained DistillBERT (RBT+DistillBERT) shows significant performance gain by the model that does not use that one (RBT). Third, the model that applies a large pre-trained model for turn representation (BERT) can get the performance that matches the state-of-the-art model (RBT) in terms of P@1 in the Reddit dataset. Figure 5.3 shows the re-

Model	Reddit				Twitter			
	MAP	P@1	P@3	NG@5	MAP	P@1	P@3	NG@5
BERT	23.9	18.7	25.2	23.7	25.4	18.8	26.4	25.1
RBT	26.5	18.8	28.2	26.3	29.6	21.1	31.7	29.6
RBT+DistillBERT	37.2	29.9	39.8	37.5	34.1	26.3	36.6	34.3
Our Model	**39.8**	**31.8**	**42.9**	**40.3**	**38.4**	**29.6**	**41.4**	**38.7**

Table 5.3: **Main comparison results on Reddit and Twitter datasets (in %).** NG@5 represent NDCG@5

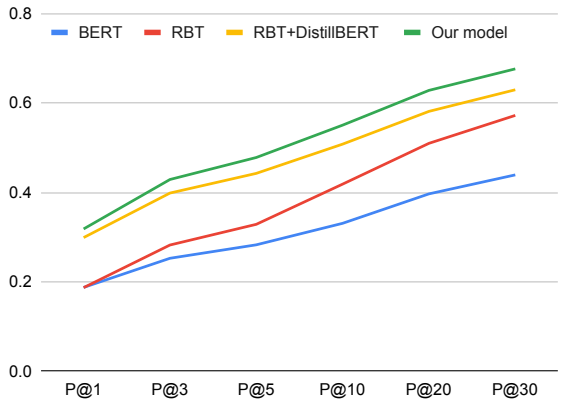

Figure 5.3: **Comparison results of P@k in the Reddit dataset**. Our model outperforms the baselines in terms of P@k, where k=1, 3, 5, 10, 20, and 30

sults of P@k while increasing the recommendation list size. Although overall precision decreases as k decreases, our method consistently outperforms other methods.

5.4 Analysis

5.4.1 Ablation Study

Table 5.4 shows impacts of our model's submodules. The first notable observation is that the performance of "W/O candidate generation" is extremely low. This result suggests that using the hard negative examples is essential for training the whole model. Second, using both topic keywords and matching with

Model	MAP	P@1	P@3	NG@5
Our Model	**39.8**	**31.8**	**42.9**	**40.3**
W/O topic keywords	39.7	31.6	42.8	40.2
W/O re-ranker	39.1	31.0	42.2	39.7
W/O candidate generation	0.8	0.0	0.2	0.3

Table 5.4: **Results of ablation study (in %)**. "W/O topic keywords" is a model that uses context input without topic keywords. "W/O re-ranker" is a model that uses only a candidate generator. "W/O candidate generation" is a model that depends solely on dual matching re-ranker.

representations on top of salient tokens also help to improve the performance. We can observe that the effect of applying dual matching in the CbR task is more effective than using topic keywords in the quote recommending task.

5.4.2 Case Study

Table 5.5 demonstrates an example that our model succeeds in predicting the GT quote at the top-1 position of the results. The given conversation discusses a situation where conflicts between politicians arise. The recommended quote at the top-1 position shows that our model can recommend the quote of which word includes the metaphorical word 'enemy.' The recommended quotes such as #2, #3, and #4 are regarding politics. We cannot say that the bottom-ranked quotes #8 and #10 perfectly fit the local context, but these are also about conflicts between humans.

5.4.3 Impact of Length of Context

As we confirmed in Section 3.4.3, the performance of knowledge selection can decrease as the context length (the number of turns) increases. We also examine the context length's effects on performance in the quote recommendation task. Figure 5.4 shows the performance MAP of each model according to the number of turns in the conversation history. We excluded the result whose number

Context	labor with a seat majority election now
	if abbott doesn't want to go down without a fight and causes a nuclear war within the liberal party essentially distracting them from running the country all the senate would have to do is block supply and force a double dissolution
	i agree they could do this but they wont. this government is tearing itself apart anyone seeking re-election in the senate will be pragmatic and pass reasonable legislation and resist anything unpopular. the government can only go from bad to worse at this point as napoleon so famously put it
Outputs	1 - *Never interrupt your enemy when he is making a mistake.
	2 - Politics is the art of the possible.
	3 - The tree of liberty must be refreshed from time to time with the blood of patriots and tyrants.
	4 - In politics, nothing happens by accident. If it happens, you can bet it was planned that way.
	5 - The one pervading evil of democracy is the tyranny of the majority, or rather of that party, not always the majority, that succeeds, by force or fraud, in carrying elections.
	6 - Facts are stubborn things.
	7 - In the councils of government, we must guard against the acquisition of unwarranted influence, whether sought or unsought, by the military-industrial complex. The potential for the disastrous rise of misplaced power exists and will persist.
	8 - I hope we shall crush in its birth the aristocracy of our monied corporations which dare already to challenge our government to a trial by strength, and bid defiance to the laws of our country.
	9 - When the going gets weird, the weird turn pro.
	10 - We shall defend our island, whatever the cost may be, we shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender.

Table 5.5: **An example of our systems' output**. The model recommended top-10 quotes for the given conversation history where each row corresponds to a turn. * denotes GT quote. The model recommended the GT quote as the first rank.

of turns is one because its test size is too small[4] to evaluate the performance reliably. While there exists a trend wherein the accuracy of each model slightly decreases as the turn proceeds, our model shows an improvement by 0.02 on average compared to RBT+DistillBERT. This means that applying dual matching

[4]Table 5.2 reports the size as 3

with the topic keywords extracted from the beginning of the conversation, i.e., the differentiation of our model compared with RBT+DistillBERT can alleviate the decrease in performance as the context length increases.

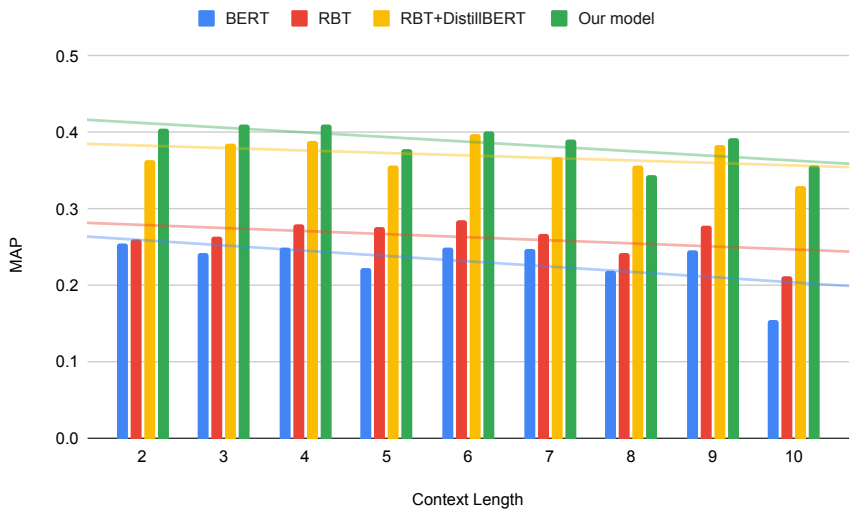

Figure 5.4: **Performance (MAP) depending on context length**

5.4.4 Impact of Training Set Size per Quote

Depending on the quote, the frequency of use of the quote in conversations can vary. Consequently, for some quotes, a small number of training instances can be collected in building the training dataset, affecting the recommendation performance of the rare quotes. Figure 5.5 shows the model's performance depending on the number of quotes in the training set. Our model shows higher accuracy than other models, even for quotes with a frequency of about 200 or less. In particular, there is an improvement of 0.04 on average for quotes with a frequency of 50 or less.

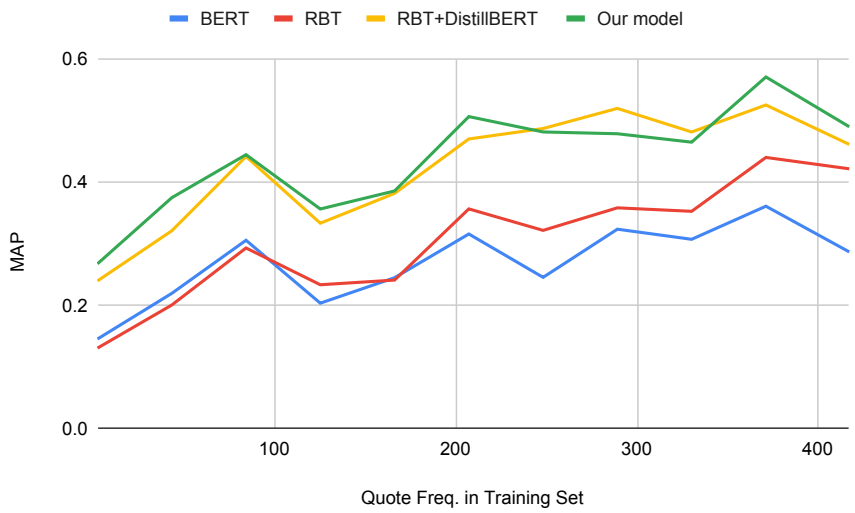

Figure 5.5: **Performance (MAP) depending on the training set size per quote**

Chapter 6

Conclusion

In this dissertation, we studied two methods to find appropriate knowledge from unstructured KB for KGCs. In particular, we explored various conversation characteristics, i.e., sequence of turns, topic, and local context of conversations, to find documents (knowledge) suitable for a given conversation to generate knowledge grounded responses. We showed that our methods could help the model generate relevant, informative, and interesting responses via extensive automatic and human experiments. Moreover, we showed that our knowledge ranking method could be applied to challenging task quote recommendations. We expect our methods to be adopted in social chatbots, task-oriented chatbots such as personal assistants, smart speakers, or recommendation bots in e-commerce.

Our methods have advantages in that our model can identify appropriate knowledge for a long conversation history without complicated training methods. In contrast, many existing KGC models focus on complicated learning methods such as posterior-prior distribution matching [27, 70], dual learning

[130], and reinforcement learning [131].

6.1 Contributions and Limitations

We summarize our contributions and the limitations of our work for Chapter 3, Chapter 4, and Chapter 5 in which we introduced our methods.

Chapter 3 We introduced two novel knowledge selection strategies, Match-Reduce and Reduce-Match, that extensively apply text-matching techniques while considering the sequential structure of a conversation. We performed experiments with various implementations in two publicly available datasets and showed that our best models could outperform the state-of-the-art models. Our best model based on Match-Reduce outperformed the baselines in the WoW dataset, and the one based on Reduce-Match did in the CMU_DoG dataset in terms of knowledge selection accuracy and text generation metrics. We reported interesting findings through in-depth analysis: 1) While the Match-Reduce strategy conducting sentence-level match outperforms the Reduce-Match overall, which distills the features before the match, the Reduce-Match strategy can show remarkable performance in a specific circumstance, i.e., GRU Agg. $\rightarrow$ Dot product Match method shows the best result in the CMU_DoG dataset, which shows the importance of extensive exploration of the specific methods. 2) There exist a trend wherein the knowledge selection accuracy drops as the turn proceeds, which may necessitate methods to filter irrelevant features accurately. 3) Our best matching-based methods have a limitation in examples with specific dialogue acts, e.g., context including the question.

In order to train our model, pseudo labels should be created if a GT knowledge sentence does not exist. It would be difficult to create meaningful labels when the ratio of grounded responses whose words overlap the knowledge be-

cause the pseudo label generation method is a term-matching-based method. One should explore different methods, such as ones based on distributional representation to address this issue.

Chapter 4 We proposed a novel retrieval-augmented KGC model considering both the conversation topic and local context and a training method based on our new data weighting scheme. Automatic and human evaluation results with the CbR task show that our model achieves state-of-the-art performances in terms of utilizing external knowledge (i.e., grounding) and general aspect quality of the response. The automatic evaluation results regarding groundness show that our models produce responses more reflective of relevant documents in the KB. Additionally, the results indicate that our model yields responses that are more relevant to the context and contain more diverse words than other models. The human evaluation results show that our best model beats other models in terms of knowledgeability, interestingness, and relevance. The advantage of our TADM re-ranker is to maintain topical consistency in a long conversation without processing all the tokens in the conversation history, contrasted with recent works on Transformer architecture such as Longformer [132] of which aim is to accept lengthy input. Our experiment results show that only extracting keywords is sufficient in finding the main reference document. Also, we observed data weighting scheme is a workhorse for training E2E retrieval-augmented model in noisy real-world conversations.

Our method has limitations in that it can learn undesirable properties that should not be learned if trained with conversations in the wild, such as sexist or racist utterances for commercial service. Moreover, our model has a somewhat high latency in responding to the input because our model has to retrieve the document and encode the retrieved after the input conversation is accepted.

The latency issue is also a common limitation of retrieval-augmented models such as in [2] and [25].

Chapter 5 We applied the TADM module of Chapter 4 to the quote recommendation task, which is considered difficult due to the metaphorical expressions in the quote. To this end, we proposed a novel quote recommendation framework called CAGAR that can provide hard negative examples to our TADM re-ranker. Our experiment results show that our model outperforms the new state-of-the-art models on Twitter and Reddit datasets.

Our method mines hard negative examples based on a classification model which uses all quotes in the quote KB as input. Consequently, if the size of quote KB is large, it will be difficult for the candidate generator to operate due to the limited computation time and memory. To this end, approximation methods substituting the classification-based candidate generator are necessary to utilize a small number of hard negative candidates. We can exploit the K-NN search used in Chapter 4 as an approximation method.

6.2 Future Works

Integrating KGC model and QA model As the name of the *knowledge grounded conversation* suggests, we can expect the KGC models should be able to respond to users' questions regarding external knowledge well as a fundamental functionality. Furthermore, question and non-question type utterances are mixed in the conversation, as shown in Table 3.6. However, the model's performance when the input is question-type utterances is unsatisfactory. We propose to study the conversational system that integrates the capabilities of the KGC model and QA system as future works. The recent work [133], which integrates open-domain DS into task-oriented DS, can be a promising direction

for solving this problem. Another promising direction can be defining an additional classifier to discriminate the query type and integrate them with two pre-trained models, a KGC and a QA model.

Resolving hallucination problem Hallucination problem [25, 24] indicates that the model generates text that is not factually correct. To address this issue, we propose to study the hallucination problem in training the models with KGCs in the wild. The hallucination problem in the KGC model can get worse if we train the model with KGC datasets in the wild because the users often make responses that are not grounded. However, if we resolve the problem, we can open up more opportunities to exploit the richness of conversations in the wild. In this regard, several current studies are at ongoing status involving building the benchmark [134] and studies on the models using crowd-sourced KGC [135]. One of the promising methods to cope with this issue will be training the KGC model based on the signals from a fact verification model using a dialogue fact-checking dataset [134].

Developing KGC system for children Previous studies on KGC utilized data collected from conversations between adult participants. Thus the level of language and knowledge in the dataset fit adults. Thus the KGC systems trained with such datasets will have limitations in that they have no basic functionalities giving linguistic or common sense to children in children's language or children's perspective. To end this, we propose to develop KGC systems for educating children. KGC system dedicated to children will be more beneficial for those lacking reading and comprehension and enable them to learn knowledge themselves without detailed guidance from an educator. Conducting fundamental research activities such as building datasets, surveys, and

exploring methods are needed. For example, to build the dataset, we can collect knowledge grounded conversations occuring while parents read books to their children. Moreover, we can use some conversation corpus from TV programs for children, such as [136].

Exploiting knowledge in various types In Chapter 1, we presented the advantages of unstructured text compared to structured data. However, not all knowledge in the world is in the form of unstructured text. Knowledge can be represented by various data types such as language, graphs, images, and tables. The ultimate goal of knowledge grounded conversation will be to build a system that can communicate with knowledge in the form of various types. Therefore, one of the most important research directions for KGC will be developing a model that can generate a response grounded on the various types of knowledge. The model presented in [33] is a good example of this direction of research—this study attempts to represent a specific type of knowledge through a dedicated representation method for each data type. In future works, we may first need to explore representation methods for KGC models by exploiting the results of multi-modal representation research shown in [137], including joint representation, coordinated representation, and encoder-decoder models. After that, we naturally need to develop a model which can be operated in real-world conversations where knowledge in various types in one dataset is utilized, where building knowledge grounded conversation datasets should also be needed.

Appendix A

Preliminary Experiments for Quote Recommendations

In order to find out the usefulness of the feature from the conversation solely, we explored various fundamental IR or machine learning algorithms with conversation and composition data. The algorithms include term-matching-based similarity-based methods, random forest, and neural network algorithms. The following session explains the methods, why we chose each method for the experiments, and the experimental results.

A.1 Methods

A.1.1 Matching Granularity Adjustment

In this section, we discuss methods to deal with the contexts of quotes when measuring relevance between a query and a set of contexts of quotes, which we call matching granularity adjustment. As the words in quotes differ from the words in context, the prior models in quote/citation recommendation [77, 78, 138] measure the relevance between query and contexts of a quote. More

specifically, all of them attempt to examine the individual context of a quote to the query. A drawback of this approach is that it suffers from sparsity problem that words in query do not match the individual context of the correct quote. In order to alleviate this sparsity problem, we explore methods to adjust the matching unit of contexts to the given query. We believe that more semantics can be exploited if the contexts of a quote are treated collectively.

Firstly, we experiment a method called context clustering, which groups the context by context cluster representing (latent) topics. In the collected dataset, we observed that there exist many quotes that can be used in multiple contexts with different topics. For example, the quote All work and no play makes jack a dull boy can be used in very different situations such as "overworking in workplace" or "educating children." Thus when dealing with a query about a specific topic, we need to consider the contexts related to it among different topics of quote. In context clustering, we first cluster the contexts of each quote. Then, we exploit the context clusters to measure the relevance of a quote. For context clustering, we adopt the affinity propagation clustering algorithm, which is known to perform better than others in short text clustering [139]. Based on context clustering, we propose a scoring function given the query.

$$sim_{cmax}(q,t) = max(sim(q, CC_t^{(j)})) \tag{A.1}$$

where $CC_t^{(j)}$ is jth context cluster of quote t and sim is cosine similarity with their TF-IDF vector representation.

In order to solve the sparsity problem, we experiment another method called context lumping to adjust the matching granularity. In context lumping, we simply concatenate all the context of each quote and make it a matching unit to the query. Then the lumped context of the quote is compared to a query with cosine similarity with TF-IDF vector representation. In context clustering and

lumping, quotes are sorted by the proposed similarities in descending order.

A.1.2 Random Forest

In the datasets, we observe that some simple rules, such as checking whether the given context contains certain words are reliable cursors to its correct label. For example, in the Twitter dataset, given that a context contains the keywords 'invite', 'join', 'come over,' or any of the morphemes, there is 40.2% probability that the context is labeled with the proverb "the more the merrier". From this observation, we explore the possibility of adopting a tree-based classification algorithm for the quote recommendation task.

Among various decision tree algorithms, Random Forest (RF) [140] is an ensemble learning method that has had notable success in various fields due to its resilience to over-fitting and tendency to exhibit low variance and bias. Random Forest constructs n_{tree} decision trees by training each tree with samples of a random subset of features. The method can populate each decision tree with the most discriminating quotes at each state and aggregate the results by voting. In the case of our dataset, we view contexts as 'documents' and use bag-of-words TF-IDF as features for each context. Then, we train the Random Forest Classifier using the vectors of TF-IDFs and their correct labels, i.e., quote.

A.1.3 Convolutional Neural Network

Word matching-based methods such as context-aware relevance model [77] and citation translation model [78] have difficulty exploiting n-gram features because of sparsity problem, so they only use unigram-based features. However, n-gram features are crucial because many phrases are meaningful only when the terms stay together. For example, if a phrasal verb "give up" lose its meaning when it

is tokenized into “give” and “up”. Unlike matching-based methods, CNN based approach can exploit important n-gram features in the context by learning the parameters of fixed size filters for each n-gram. Generally, CNN comprises several pairs of a convolution layer and max-pooling layer, which capture the local patterns from the training example and down-sample extracted features to prevent overfitting. When CNN is applied to natural language sentences, it captures the significant local semantics, i.e., n-gram, of sentences.

We adopted a single-layer CNN, mainly inspired by [141] which reports that a simple CNN model performs as well as a complex one with several convolutions-pooling layers to capture distinguished n-gram features in contexts of quotes. Our CNN model takes context in the form of a list of word embedding vectors and maps that context of vector representation with a single convolution and pooling layer. After that, the vector representation is fully connected to the softmax layer to compute the probability of candidate quotes and rank the quotes.

A.1.4 Recurrent Neural Network

We use a recurrent neural network (RNN) to tackle our quote recommendation problem from the perspective of language modeling, which means that we treat each quote as a special token or word and compute the probability of it for a given context. While none of the above approaches uses order information of words in the context, RNN based approach can recursively model such a sequence of words. We use a long LSTM [73] which is a recurrent neural network consisting of three gates (forget, input, output) that control the networks to learn long-term dependencies without loss of information. The input vector of each time step passes through the three gates and updates latent vectors, which LSTM is retaining. In our model, we recurrently feed LSTM with a sequence of

words in the form of a list of word embedding vectors. The output of the LSTM layer is passed to a fully connected layer with softmax activation to compute the probability of target quotes to be recommended.

A.2 Experiments

A.2.1 Baselines and Implementation Details

We compare our approaches with three state-of-the-art recommendation approaches: popularity-based method (Popularity), cosine similarity-based method (Cosine similarity), learning-to-recommend quote (LRQ) [138], context-aware relevance model (CRM) [77], and citation translation model (CTM) [78]. These methods are described in detail below. Among the methods, popularity-based and cosine similarity-based methods are used to conduct control experiments to reveal the different levels of difficulties of the datasets.

LRQ exploits an existing learning-to-rank framework for quote recommendation with quote-based features, quote-query similarity features, and context-query similarity features.

CRM recommends quotes according to an average of the squared cosine similarities between the contexts of each quote and the query.

CTM recommends quotes according to the probability that the query context would be translated into the quote.

Popularity ranks the quotes according to their frequency in contexts of the training set.

Cosine similarity ranks the quote by examining the individual context of the quote with the given query using bag-of-words representation.

We implement these methods and set the parameters to optimum as specified in the respective papers of the methods. Specifically, we truncate each half-context (pre-context or post-context) of length longer than 150 charac-

ters for LRQ, 50 words for CRM and one sentence for CTM, respectively, as the respective authors suggested in the papers. For our approaches, we set the length of half-context to its optimal value which shows best result in validation dataset: 1) 150 characters of pre-context and post-context with word truncation for context clustering and context lumping, 2) 50 words for RF, and 3) 30 words of pre-context for CNN and RNN. As stated in the introduction, we used pre-context and post-context as a query for Gutenberg and Blog datasets and pre-context as a query for Twitter dataset. Hyperparameters of single algorithms are set by using a validation set.

A.2.2 Datasets

We have collected 439,655 quotes from three sources: Wikiquote[1], Oxford Concise Dictionary of Proverbs[2], and Library of Quotes[3]. For the context data, we searched blocks of texts that contain these quotes from three different sets of corpus: 2 million tweet threads from Twitter ($\sim$ 2015.11.15), 20GB of electronic books from the Project Gutenberg Database[4], and 190GB of ICWSM spinn3r 2009 blog dataset[5]. In the tweet corpus, to extract dialogs only, we selected threads where only two users are involved. Next, we chose the top 400 quotes from each corpus according to the number of contexts, to reflect the characteristics of the quotes that frequently appeared in the different corpora. Finally, we generate three datasets: Twitter dataset, Gutenberg dataset, and Blog dataset Table A.1 shows the number of contexts for each quote in each dataset, which describes the average, maximum, and minimum number of contexts for each quote and its standard deviation of them. From Table A.1, we see that the

[1] `https://en.wikiquote.org/` (Accessed: May 5th, 2016)
[2] Oxford University Press, 1998
[3] `http://www.libraryofquotes.com/` (Accessed: May 5th, 2016)
[4] `http://www.gutenberg.org/` (Accessed: May 5th, 2016)
[5] `http://icwsm.cs.umbc.edu/data/icwsm2009/` (Accessed: May 5th, 2016)

most frequently appeared quotes from each corpus cover a large range of quotes of varying frequencies, helping us deal with the situation recommending quotes by using a small number of contexts as well as a large number of contexts. We divide the dataset into the proportion of 8:1:1 as the training, validation, and test set. We create test sets by hiding the quotes with which the contexts are paired.

Datasets	Avg	Std dev	Max	Min
Twitter	556	971	10764	15
Gutenberg	89	122	1366	14
Blog	230	543	5923	24

Table A.1: number of contexts for each quote in datasets

A.2.3 Results and Discussions

Results of experiments are listed in Table A.2. P@5 and the improvement ratio of each algorithm over the best baseline in each dataset are denoted. Even without rank aggregation, the individual algorithms (context lumping and CNN) outperform baselines in all datasets. Surprisingly, the simple method, context lumping, is the best performer in Gutenberg and Blog dataset, which beats LRQ up to 35%. Context clustering outperforms CRM and Cosine similarity, which does not collectively treat the context of the quote. These better results of context lumping and context clustering show the effectiveness of adjusting context matching granularity. One can observe that performance of the baseline Cosine similarity in the Twitter dataset is worse than the ones in Gutenberg and Blog datasets. This means that sparsity problem is more serious in the Twitter dataset, where the tweet contains more infrequent words than others. In the Twitter dataset, deep learning algorithms (CNN and RNN) outperform CTM by up to 43%. From this result, we can see that deep learning algorithms

can mitigate such serious sparsity problem because it is not based on word matching. Results of RF show that it is competitive with the CTM algorithm. In fact, RF outperforms CNN in our preliminary experiments on a top-100 Twitter dataset. However, in a large dataset, generalization of the algorithm is not made as expected, an area for future investigation.

Table A.2: **Results of P@k of different methods.** * indicates that each of our algorithms outperform the best baseline algorithm with statistically significant increase at $p < 0.01$ in two-tailed t-tests

Approaches	Twitter	Gutenberg	Blog
Context clustering	0.190	0.299	0.494
Context lumping	0.286*	0.409*	0.521*
RF	0.244	0.246	0.470
CNN	0.390*	0.326*	0.506
RNN	0.389*	0.294	0.473
LRQ	0.196	0.302	0.494
CRM	0.119	0.237	0.382
CTM	0.273	0.257	0.441
Popularity	0.156	0.111	0.223
Cosine similarity	0.196	0.248	0.469

Bibliography

[1] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 654–664, 2017.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*, (Online), pp. 7871–7880, Association for Computational Linguistics, July 2020.

[4] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, "Grounded response generation task at dstc7," in *AAAI Dialog System Technology Challenges Workshop*, 2019.

[5] H.-Y. Shum, X. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *arXiv preprint arXiv:1801.01957*, 2018.

[6] J. McCarthy, M. Minsky, N. Rochester, and C. Shannon, "The dartmouth summer research project on artificial intelligence," *Artificial intelligence: past, present, and future*, 1956.

[7] N. Chomsky, *Aspects of the Theory of Syntax*. The MIT Press, MIT Press, 1969.

[8] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, p. 36–45, jan 1966.

[9] R. S. Wallace, "The anatomy of alice," in *Parsing the turing test*, pp. 181–210, Springer, 2009.

[10] T. Winograd, "Procedures as a representation for data in a computer program for understanding natural language," 1972.

[11] G. Güzeldere and S. Franchi, "Dialogues with colorful "personalities" of early ai," *Stanford Humanities Review*, vol. 4, no. 2, pp. 161–169, 1995.

[12] "History of natural language processing." `https://en.wikipedia.org/wiki/History_of_natural_language_processing`. Accessed: 2022-06-19.

[13] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[15] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[19] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Trans. Inf. Syst.*, vol. 38, apr 2020.

[20] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 110–119, Association for Computational Linguistics, June 2016.

[21] X. Gao, S. Lee, Y. Zhang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "Jointly optimizing diversity and relevance in neural response generation," *arXiv preprint arXiv:1902.11205*, 2019.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[24] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 1906–1919, Association for Computational Linguistics, July 2020.

[25] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, (Punta Cana, Dominican Republic), pp. 3784–3803, Association for Computational Linguistics, Nov. 2021.

[26] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley, "A knowledge-grounded neural conversation model," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018.

[27] B. Kim, J. Ahn, and G. Kim, "Sequential latent knowledge selection for knowledge-grounded dialogue," in *Int. Conference on Learning Representations*, 2020.

[28] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention," in *Proceedings of the Twenty-Seventh Int. Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 4623–4629, Int. Joint Conferences on Artificial Intelligence Organization, 7 2018.

[29] B. P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, and J. J. McAuley, "Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions," in *EMNLP*, 2020.

[30] H. Su, X. Shen, S. Zhao, X. Zhou, P. Hu, R. Zhong, C. Niu, and J. Zhou, "Diversifying dialogue generation with non-conversational text," *arXiv preprint arXiv:2005.04346*, 2020.

[31] S. Yavuz, A. Rastogi, G.-L. Chao, and D. Hakkani-Tur, "{D}eep{C}opy: Grounded Response Generation with Hierarchical Pointer Networks," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, (Stockholm, Sweden), pp. 122–132, Association for Computational Linguistics, sep 2019.

[32] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, "Learning to select knowledge for response generation in dialog systems," in *Proceedings of the*

Twenty-Eighth Int. Joint Conference on Artificial Intelligence, IJCAI-19, pp. 5081–5087, Int. Joint Conferences on Artificial Intelligence Organization, 7 2019.

[33] A. Fan, C. Gardent, C. Braud, and A. Bordes, "Augmenting transformers with knn-based composite memory for dialog," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 82–99, 2021.

[34] Y. Ahn, S.-g. Lee, and J. Park, "Exploiting text matching techniques for knowledge-grounded conversation," *IEEE Access*, vol. 8, pp. 126201–126214, 2020.

[35] Y. Ahn, S.-g. Lee, J. Shim, and J. Park, "Retrieval-augmented response generation for knoweldge grounded conversation in the wild," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, submitted.

[36] H. Lee, Y. Ahn, H. Lee, S. Ha, and S.-g. Lee, "Quote recommendation in dialogue using deep neural network," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, (New York, NY, USA), p. 957–960, Association for Computing Machinery, 2016.

[37] Y. Ahn, H. Lee, H. Jeon, S. Ha, and S.-g. Lee, "Quote recommendation for dialogs and writings.," in *CBRecSys@ RecSys*, pp. 39–42, 2016.

[38] L. Ma, M. Li, W.-N. Zhang, J. Li, and T. Liu, "Unstructured text enhanced open-domain dialogue system: A systematic survey," *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 1, pp. 1–44, 2021.

[39] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," in *Int. Conference on Learning Representations*, 2018.

[40] L. Qin, M. Galley, C. Brockett, X. Liu, X. Gao, B. Dolan, Y. Choi, and J. Gao, "Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 5427–5436, Association for Computational Linguistics, jul 2019.

[41] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, (Melbourne, Australia), pp. 2–7, Association for Computational Linguistics, July 2018.

[42] T. Adewumi, F. Liwicki, and M. Liwicki, "State-of-the-art in open-domain conversational ai: A survey," 2022.

[43] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–32, 2020.

[44] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, (USA), p. 583–593, Association for Computational Linguistics, 2011.

[45] X. Zhou, D. Dong, H. Wu, S. Zhao, D. Yu, H. Tian, X. Liu, and R. Yan, "Multi-view response selection for human-computer conversation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Lan-*

guage Processing, (Austin, Texas), pp. 372–381, Association for Computational Linguistics, Nov. 2016.

[46] J. Gao, P. Pantel, M. Gamon, X. He, and L. Deng, "Modeling interestingness with deep neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 2–13, Association for Computational Linguistics, Oct. 2014.

[47] M. Henderson, I. Casanueva, N. Mrkšić, P.-H. Su, T.-H. Wen, and I. Vulić, "ConveRT: Efficient and accurate conversational representations from transformers," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 2161–2174, Association for Computational Linguistics, Nov. 2020.

[48] J. Song, K. Zhang, X. Zhou, and J. Wu, "Hka: A hierarchical knowledge attention mechanism for multi-turn dialogue system," in *ICASSP 2020 - 2020 IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3512–3516, 2020.

[49] W. Zhu, K. Mo, Y. Zhang, Z. Zhu, X. Peng, and Q. Yang, "Flexible end-to-end dialogue system for knowledge grounded conversation," *CoRR*, vol. abs/1709.04264, 2017.

[50] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, and D. Yin, "Knowledge diffusion for neural dialogue generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 1489–1498, Association for Computational Linguistics, July 2018.

[51] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2," in *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, (Online), pp. 583–592, Association for Computational Linguistics, July 2020.

[52] P. Budzianowski and I. Vulić, "Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, (Hong Kong), pp. 15–22, Association for Computational Linguistics, Nov. 2019.

[53] M. Caballero, "A brief survey of question answering systems," *International Journal of Artificial Intelligence & Applications (IJAIA)*, vol. 12, no. 5, 2021.

[54] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: retrieval-augmented language model pre-training," *CoRR*, vol. abs/2002.08909, 2020.

[55] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 6769–6781, Association for Computational Linguistics, Nov. 2020.

[56] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, *et al.*, "Building watson: An overview of the deepqa project," *AI magazine*, vol. 31, no. 3, pp. 59–79, 2010.

[57] J. L. Austin, *How to do things with words*. Oxford university press, 1975.

[58] O. Vinyals and Q. V. Le, "A neural conversational model," *CoRR*, vol. abs/1506.05869, 2015.

[59] B. Mitra and N. Craswell, *An introduction to neural information retrieval*. Now Foundations and Trends, 2018.

[60] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1715–1725, Association for Computational Linguistics, Aug. 2016.

[61] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152, IEEE, 2012.

[62] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[63] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[64] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[65] K. Zhou, S. Prabhumoye, and A. W. Black, "A dataset for document grounded conversations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 708–713, Association for Computational Linguistics, Oct.-Nov. 2018.

[66] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra, "Towards Exploiting Background Knowledge for Building Conversation Systems," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 2322–2332, Association for Computational Linguistics, 2018.

[67] P. Ren, Z. Chen, C. Monz, J. Ma, and M. de Rijke, "Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation," in *The 34th AAAI Conference on Artificial Intelligence*, 2020.

[68] Z. Tian, W. Bi, D. Lee, L. Xue, Y. Song, X. Liu, and N. L. Zhang, "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* (D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, eds.), pp. 650–659, Association for Computational Linguistics, 2020.

[69] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "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)*, (Doha, Qatar), pp. 1724–1734, Association for Computational Linguistics, Oct. 2014.

[70] X. Chen, F. Meng, P. Li, F. Chen, S. Xu, B. Xu, and J. Zhou, "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)*, (Online), pp. 3426–3437, Association for Computational Linguistics, Nov. 2020.

[71] C. Meng, P. Ren, Z. Chen, Z. Ren, T. Xi, and M. d. Rijke, "Initiative-aware self-supervised learning for knowledge-grounded conversations," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 522–532, 2021.

[72] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3377–3390, 2020.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[74] Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "Joint retrieval and generation training for grounded text generation," *arXiv preprint arXiv:2105.06597*, 2021.

[75] C. Meng, P. Ren, Z. Chen, C. Monz, J. Ma, and M. de Rijke, "Refnet: A reference-aware network for background based conversation," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[76] W. Wang, W. Gao, S. Feng, L. Chen, and D. Wang, "Adaptive posterior knowledge selection for improving knowledge-grounded dialogue gen-

eration," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1989–1998, 2021.

[77] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proceedings of the 19th international conference on World wide web*, pp. 421–430, 2010.

[78] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach, "Recommending citations: translating papers into references," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1910–1914, 2012.

[79] B. Shaparenko and T. Joachims, "Identifying the original contribution of a document via language modeling," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, (New York, NY, USA), p. 696–697, Association for Computing Machinery, 2009.

[80] J. Tan, X. Wan, and J. Xiao, "A neural network approach to quote recommendation in writings," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 65–74, 2016.

[81] J. Tan, X. Wan, and J. Xiao, "A neural network approach to quote recommendation in writings," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, (New York, NY, USA), p. 65–74, Association for Computing Machinery, 2016.

[82] L. Wang, X. Zeng, and K.-F. Wong, "Quotation recommendation and interpretation based on transformation from queries to quotations," in

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), (Online), pp. 754–758, Association for Computational Linguistics, Aug. 2021.

[83] L. Wang, J. Li, X. Zeng, H. Zhang, and K.-F. Wong, "Continuity of topic, interaction, and query: Learning to quote in online conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 6640–6650, Association for Computational Linguistics, Nov. 2020.

[84] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou, *et al.*, "On evaluating and comparing open domain dialog systems," *arXiv preprint arXiv:1801.03625*, 2018.

[85] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, "Incremental Transformer with Deliberation Decoder for Document Grounded Conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 12–21, Association for Computational Linguistics, jul 2019.

[86] R. Tanaka, A. Ozeki, S. Kato, and A. Lee, "Context and knowledge aware conversational model and system combination for grounded response generation," *Computer Speech & language*, vol. 62, p. 101070, 2020.

[87] Y.-C. Tam, "Cluster-based beam search for pointer-generator chatbot grounded by knowledge," *Computer Speech & language*, vol. 64, p. 101094, 2020.

[88] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 1–14, Association for Computational Linguistics, Aug. 2017.

[89] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the Int. Workshop on Paraphrasing*, 2005.

[90] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of NAACL-HLT*, 2018.

[91] R. B. Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, "The second pascal recognising textual entailment challenge," in *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, vol. 7, 2006.

[92] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The third PASCAL recognizing textual entailment challenge," in *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pp. 1–9, Association for Computational Linguistics, 2007.

[93] W. Wu and R. Yan, "Deep chit-chat: Deep learning for chatbots," in *Proceedings of the 42nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, (New York, NY, USA), p. 1413–1414, Association for Computing Machinery, 2019.

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[95] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

[96] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with gumble-softmax," in *Int. Conference on Learning Representations (ICLR 2017)*, OpenReview. net, 2017.

[97] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[98] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "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)*, pp. 1724–1734, 2014.

[99] Y. Ji and J. Eisenstein, "Discriminative improvements to distributional sentence similarity," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 891–896, 2013.

[100] Z. Wang, W. Hamza, and R. Florian, "Bilateral multi-perspective matching for natural language sentences," in *Proceedings of the Twenty-Sixth Int. Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 4144–4150, 2017.

[101] S. Yoon, J. Shin, and K. Jung, "Learning to rank question-answer pairs using hierarchical recurrent encoder with latent topic clustering," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 1575–1584, Association for Computational Linguistics, June 2018.

[102] G. Zhou, P. Luo, R. Cao, F. Lin, B. Chen, and Q. He, "Mechanism-aware neural machine for dialogue response generation," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[103] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Recurrent neural network-based sentence encoder with gated attention for natural language inference," in *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pp. 36–40, 2017.

[104] D. Yogatama, P. Blunsom, C. Dyer, E. Grefenstette, and W. Ling, "Learning to compose words into sentences with reinforcement learning," in *5th Int. Conference on Learning Representations (ICLR 2017)*, Int. Conference on Learning Representations, 2017.

[105] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zhou, "A sequential matching framework for multi-turn response selection in retrieval-based chatbots," *Computational Linguistics*, vol. 45, no. 1, pp. 163–197, 2019.

[106] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu, "Multi-turn response selection for chatbots with deep attention matching network," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Mel-

bourne, Australia), pp. 1118–1127, Association for Computational Linguistics, July 2018.

[107] Y. Wu, W. Wu, C. Xu, and Z. Li, "Knowledge enhanced hybrid neural network for text matching," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[108] C. Yuan, W. Zhou, M. Li, S. Lv, F. Zhu, J. Han, and S. Hu, "Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 111–120, Association for Computational Linguistics, nov 2019.

[109] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22Nd Int. Conference on World Wide Web*, WWW '13, (New York, NY, USA), pp. 1445–1456, ACM, 2013.

[110] X. Zhao, W. Wu, C. Tao, C. Xu, D. Zhao, and R. Yan, "Low-resource knowledge-grounded dialogue generation," in *International Conference on Learning Representations*, 2019.

[111] Y. Park, J. Cho, and G. Kim, "A Hierarchical Latent Structure for Variational Conversation Modeling," in *NAACL*, 2018.

[112] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

[113] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[114] D. Jurafsky and E. Shriberg, *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13.* ., 1997.

[115] F. Galetzka, C. U. Eneh, and D. Schlangen, "A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 565–573, European Language Resources Association, May 2020.

[116] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[117] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404–411, 2004.

[118] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring," in *International Conference on Learning Representations*, 2020.

[119] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[120] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa,

K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.

[121] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[122] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, (Prague, Czech Republic), pp. 228–231, Association for Computational Linguistics, June 2007.

[123] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, (San Francisco, CA, USA), p. 138–145, Morgan Kaufmann Publishers Inc., 2002.

[124] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, "Generating informative and diverse conversational responses via adversarial information maximization," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[125] X. Liu, Y. Shen, K. Duh, and J. Gao, "Stochastic answer networks for machine reading comprehension," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 1694–1704, Association for Computational Linguistics, July 2018.

[126] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 889–898, Association for Computational Linguistics, July 2018.

[127] Y. Liu, B. Pang, and B. Liu, "Neural-based Chinese idiom recommendation for enhancing elegance in essay writing," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 5522–5526, Association for Computational Linguistics, July 2019.

[128] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019.

[129] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[130] C. Meng, P. Ren, Z. Chen, W. Sun, Z. Ren, Z. Tu, and M. d. Rijke, "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

'20, (New York, NY, USA), p. 1151–1160, Association for Computing Machinery, 2020.

[131] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 3377–3390, Association for Computational Linguistics, Nov. 2020.

[132] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[133] T. Zhao, A. Lu, K. Lee, and M. Eskenazi, "Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, (Saarbrücken, Germany), pp. 27–36, Association for Computational Linguistics, Aug. 2017.

[134] P. Gupta, C.-S. Wu, W. Liu, and C. Xiong, "Dialfact: A benchmark for fact-checking in dialogue," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3785–3801, 2022.

[135] H. Rashkin, D. Reitter, G. S. Tomar, and D. Das, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 704–718, 2021.

[136] M. Nikolaus, A. Alishahi, and G. Chrupała, "Learning english with peppa pig," *arXiv preprint arXiv:2202.12917*, 2022.

[137] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[138] J. Tan, X. Wan, and J. Xiao, "Learning to recommend quotes for writing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, Feb. 2015.

[139] A. Rangrej, S. Kulkarni, and A. V. Tendulkar, "Comparative study of clustering techniques for short text documents," in *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, (New York, NY, USA), p. 111–112, Association for Computing Machinery, 2011.

[140] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[141] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.

초록

지식 기반 대화 모델은 대화 기록과 외부 지식 이 두 가지 모두에 관련된 응답을 생성하는 것을 목표로 한다. 지식 기반 대화 모델의 가장 중요한 부분 중 하나는 응답의 기반을 제공하는 지식을 찾는 것이다. 지식 기반 모델이 주어진 문맥에 부적합한 지식을 찾는 경우 관련성이 떨어지거나 지식이 부족한 응답이 생성될 수 있다. 이 문제를 해결하기 위해 이 논문에서는 지식 기반 대화를 위해 대화 여러 특성을 활용하여 지식을 선정하는 지식 선택 모델과 지식 순위 모델을 제시한다.

구체적으로 본 논문에서는 다중 턴 대화에서의 순차적 구조 또는 응답 이전 문맥과 대화의 주제를 활용하는 새로운 두 가지 방법을 제시한다. 첫 번째 방법으로써 본 논문은 두 가지 지식 선택 전략을 제안한다. 제안하는 전략 중 하나는 지식과 대화 턴 간의 순차적 매칭 특징을 보존하는 방법이고 다른 전략은 대화의 순차적 특성을 인코딩하여 지식을 선택하는 방법이다. 두 번째로 본 논문은 대화의 주제 키워드와 응답 바로 이전의 문맥을 모두 활용하여 적절한 범위의 관련 문서들로 검색 결과를 구성하는 새로운 지식 순위 모델을 제안한다. 마지막으로 지식 순위 모델의 적응성 검증을 위해 정답 인용구와 의미적으로 유사하지만 정답은 아닌 인용구의 집합을 인용구 순위 모델에 제공하는 인용구 추천 프레임워크를 제안한다. 제안된 지식 선택 및 순위 모델을 기반으로 하는 지식 기반 대화 모델이 경쟁 모델보다 외부 지식 및 대화 문맥과의 관련성 측면에서 우수하다는 것을 사람 간의 대화 데이터를 이용한 다수의 실험을 통해 검증하였다.

주요어: 지식 기반 대화, 오픈 도메인 대화 시스템, 의미 매칭, 지식 선택, 지식 랭킹, 신경망

학번: 2012-30214