Adapting to Context-Aware Knowledge in Natural Conversation for Multi-Turn Response Selection

Chen Zhang*
runchen.zc@alibaba-inc.com
Alibaba Group
Beijing, China

Hao Wang*
qiao.wh@alibaba-inc.com
Alibaba Group
Beijing, China

Feijun Jiang
feijun.jiangfj@alibaba-inc.com
Alibaba Group
Hangzhou, China

Hongzhi Yin
h.yin1@uq.edu.au
The University of Queensland
Brisbane, Austrilia

ABSTRACT

Virtual assistants aim to build a human-like conversational agent. However, current human-machine conversations still cannot make users feel intelligent enough to build a continued dialog over time. Some responses from agents are usually inconsistent, uninformative, less-engaging and even memoryless. In recent years, most researchers have tried to employ conversation context and external knowledge, e.g. wiki pages and knowledge graphs, into the model which only focuses on solving some special conversation problems in local perspectives. Few researchers are dedicated to the whole capability of the conversational agent which is endowed with abilities of not only passively reacting the conversation but also proactively leading the conversation.

In this paper, we first explore the essence of conversations among humans by analyzing real dialog records. We find that some conversations revolve around the same context and topic, and some require additional information or even move on to a new topic. Base on that, we conclude three conversation modes shown in Figure 1 and try to solve how to adapt to them for a continuous conversation. To this end, we define "Adaptive Knowledge-Grounded Conversations" (AKGCs) where the knowledge is to ground the conversation within a multi-turn context by adapting to three modes. To achieve AKGC, a model called ***MNDB*** is proposed to **m**odel **n**atural **d**ialog **b**ehaviors for multi-turn response selection. To ensure a consistent response, ***MNDB*** constructs a multi-turn context flow. Then, to mimic user behaviors of incorporating knowledge in natural conversations, we design a ternary-grounding network along with the context flow. In this network, to gain the ability to adapt to diversified conversation modes, we exploit multi-view semantical relations among response candidates, context and knowledge. Thus, three adaptive matching signals are extracted for final response selection. Evaluation results on two benchmarks indicate that ***MNDB*** can significantly outperform state-of-the-art models.

*Chen Zhang and Hao Wang contributed equally to this research.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449902

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

Response Selection, Knowledge-grounded Conversation.

ACM Reference Format:
Chen Zhang, Hao Wang, Feijun Jiang, and Hongzhi Yin. 2021. Adapting to Context-Aware Knowledge in Natural Conversation for Multi-Turn Response Selection. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449902

1 INTRODUCTION

Conversation system in open-domain settings as one of the ultimate goals of artificial intelligence has gained much attention. As a more natural interaction means, conversation system has been applied in massive chatbot platforms, such as Amazon Alexa [1] and Google Assistant [2]. To accomplish a conversation, a chatbot should be able to take the historical utterances as input, comprehend the natural language, and output the proper natural language response.

Current open-domain chatbots have made much progress, but still are not intelligent enough. They mainly suffer from three issues:

(1) They produce inconsistent and memoryless conversations lacking context information.
(2) They generate uninformative responses without considering background knowledge and human experiences.
(3) They form less-engaging conversations with passively reacting to an utterance.

Basically, existing conversation methods either model multi-turn context information to address the issue (1), or use knowledge to ground conversations for the issue (2). For example, initial models tend to produce inconsistent responses as they only model single-turn dialog behaviors and neglect the context information. Afterwards, multi-turn conversation models are proposed [19, 20, 34], but they still suffer an issue of producing less-informative responses. Recently, the knowledge-grounded models have been introduced

[1] https://developer.amazon.com/en-US/alexa.
[2] https://assistant.google.com/.

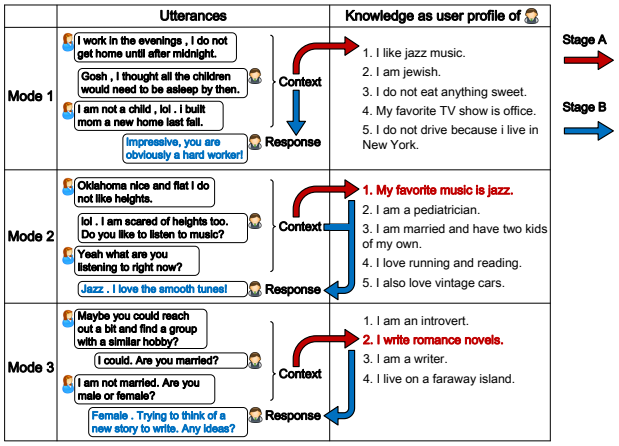


Figure 1: Examples of two-stage conversation process from Persona-Chat dataset [27]. In stage A, knowledge k is selected based on context c. In stage B, the response r follows three conversation modes. 1: r is derived from c for acknowledgment. 2: r is derived from c grounded by k to continue current topics. 3: r is derived from k to start new topics.

to address this issue [2, 7, 10, 16, 22, 28, 29]. In those models, background knowledge about conversation plays a critical role for producing more informative responses that are consistent and relevant with both the context and the knowledge. Current researches employ personal profiles [27], wiki pages [2], Knowledge Graph (KG) triplets [22], etc as the knowledge. Yet, some knowledge-grounded models [2, 3, 22, 29] simplify the problem that the multi-turn context is concatenated as a single sequence, and converted to a fixed-length vector when selecting the proper knowledge. This multi-turn information loss degrades the effect of knowledge selection and affects the consistency of conversations.

However, there is no solution that concentrates on the whole capability of a conversational agent which is endowed with abilities of not only passively reacting the conversation but also proactively leading the conversation. Besides, current models, including knowledge-grounded ones, have very limited efforts on the issue (3), i.e. they lack the ability to lead the conversation, which leads to less-engaging conversations.

Figure 1 tries to demonstrate the essence of human conversation, which shows that human produces a response in a two-stage process. First, he determines whether to use knowledge and which knowledge to use based on current context. After that, he returns a response following three conversation patterns or modes. That is, he decides to return a response derived from context without knowledge for acknowledgment (**Mode 1**), derived from knowledge-enriched context to continue the current topic (**Mode 2**), or derived from knowledge to start a new topic (**Mode 3**). Besides, human can adaptively change one mode to another based on current context and knowledge. These modes will be elaborated in Section 2. However, existing methods can not well model to adapt to these modes, so that the whole conversational capability can not be reached.

Therefore, *how to perform context-aware and adaptive knowledge grounding to achieve a continuous conversation* is a challenge. To this end, we propose "Adaptive Knowledge-Grounded Conversations" (AKGCs). In AKGCs, we should address two tasks within a multi-turn context: *1. Which part of knowledge to choose* and *2. How to adaptively incorporate knowledge*. Previous studies mainly focus on handling the first task, although their solutions of knowledge selection are not perfect enough due to context information loss. However, they have difficulties on the second task. Unlike in current methods that knowledge is used to solve partial conversation modes, the knowledge in AKGCs should be leveraged to adapt to all three modes for a continuous conversation.

To achieve AKGCs, we propose a two-stage retrieval model, called ***MNDB***, to **m**odel **n**atural **d**ialog **b**ehaviors for multi-turn response selection. Taking the multi-turn semantic information into consideration, ***MNDB*** constructs a conversation flow with each dialog turn as a node, where targeted matching networks are leveraged on each node to capture the corresponding conversation modes. Thus, the correct response is selected by linking the conversation flow via a RNN network.

In the first stage, we propose a flow-based knowledge selection network to solve the task of *which part of knowledge to choose*. Specifically, each utterance-knowledge pair is interacted for distilling matching information on each conversation node. Then, the matching information on all nodes are aggregated along with the conversation flow. To model the responses that no knowledge is to be incorporated (Mode 1), we add a sentinel denoting null knowledge into the knowledge base, then the model decides to select whether the real knowledge or the sentinel.

In the second stage, we propose a ternary-grounding response selection network that solves the task of *how to adaptively incorporate knowledge*. Particularly, unlike previous studies that knowledge is exclusively used for grounding context representations, we design three complementary grounding strategies, called *knowledge-enriched matching*, *knowledge-centered matching* and *knowledge-bridged matching*. In *knowledge-enriched matching*, for modeling the responses that the knowledge is used to supplement current topics (Mode 2), we leverage the knowledge to enrich the matching between context and response by constructing knowledge-aware representations of context and response. In *knowledge-centered matching*, for modeling the knowledge being used to support new topics (Mode 3), we perform direct matching between response and knowledge. In *knowledge-bridged matching*, we additionally use the knowledge to bridge the implicit relations between context and response for solving Mode 2 in another view. Finally, three complementary signals are extracted and again aggregated via a grounding adapter for gaining the ability of adaptively changing from one mode to another based on current context and knowledge.

To validate ***MNDB***, we perform extensive experiments on two public datasets: Persona-Chat [27] and Wiki-Wizard [2]. The experimental results show that ***MNDB*** outperforms baselines by a large margin. In Persona-Chat, ***MNDB*** achieves 8.0% absolute improvement on r@1 than the state-of-the-art model. In Wiki-Wizard, ***MNDB*** as a non-pre-trained model can even outperform the state-of-the-art pre-trained model by 4.1% absolute improvement.

To the best of our knowledge, this is the first attempt to propose a response selection method which models the natural conversation behaviors for the AKGC task. The major contributions are:

Table 1: Ratio of 3 response modes and their characteristics.

	Mode 1	Mode 2	Mode 3
Ratio in Persona-Chat [27]	36.0%	52.0%	12.0%
Ratio in Wiki-Wizard [2]	10.5%	79.0%	10.5%
Context consistent	✓	✓	✓
Knowledge incorporated	×	✓	✓
Knowledge starting new topic	–	×	✓

(1) We analyze the essence of natural conversations and conclude three conversation modes.
(2) We propose a two-stage model to adapt to the conversation modes for a continuous conversation.
(3) We conduct extensive experiments to evaluate our model on two benchmarks.

2 NATURAL CONVERSATION MODES

We try to investigate the essence of conversations among humans by analyzing two public datasets Persona-Chat [27] and Wiki-Wizard [2]. Then, we find that natural conversations could be classified as three modes:

- **Mode 1 (Context based response)**: The response follows the dialog context, but uses no extra knowledge.
- **Mode 2 (Context based response with knowledge)**: The response follows the context, and uses the extra knowledge that is complementary to current topic.
- **Mode 3 (Cross-context based response with knowledge)**: The response follows the context, but switches to a new topic, and uses the new knowledge that is related to the new topic.

The process that leads to the discovery of the three conversation modes will be described in Appendix B. Table 1 shows the ratio of responses that belong to such these modes [3]. More examples of the three modes are in Appendix C.

From the investigation we find that these natural conversations have some vital characteristics, as shown in bottom part of Table 1.

First, the basis of conversations is that in all modes the response r must be semantically consistent (or called "matched") with the dialog context c. Therefore, the requirement of semantic matching between r and c is necessary for all responses.

Second, the condition of knowledge grounding is optional for a response. For example, the responses in Mode 1 use no extra knowledge, because the information in the context is sufficient for returning a response; but the responses of Modes 2 and 3 use extra knowledge to ground the context information.

Third, the knowledge is to be incorporated adaptively. In Mode 2, the knowledge that contains complementary information to the current topic is used to continue the current topic, i.e. the matching between r and c is enriched by the knowledge k. However, the knowledge k in Mode 3 is introduced to start a new topic, which means that k can be rarely relevant or irrelevant with the current topic. In other words, k can be matched with r directly.

Last, human can adaptively and freely switch one mode to another based on current context and knowledge in a natural conversation, instead of following a pre-defined procedure.

[3]The ratio is obtained on 200 random samples from both datasets.

Section 4.2 will introduce the corresponding matching signals for modeling the three conversation modes adaptively.

3 PROBLEM STATEMENT

The AKGC dataset $\mathcal{D} = \left\{D_i : (y_i^k, y_i^r, C_i, K_i, R_i)\right\}_{i=1}^{|\mathcal{D}|}$, where $|\mathcal{D}|$ denotes the data size, i.e. the number of conversations. Each sample D_i has three attributes: C_i, K_i and R_i. C_i is a dialog context with N turns:$\{u_{i,1}, u_{i,2}, ..., u_{i,N}\}$, where $u_{i,j}$ denotes the j-th utterance. $K_i = \{k_{i,1}, k_{i,2}, ..., k_{i,M}\}$ represents a knowledge base, e.g. wiki pages [2] or user profiles [27], that contains M piece of knowledge, where $k_{i,j}$ means the j-th knowledge. R_i is a response candidate pool that contains L candidates: $\{r_{i,1}, r_{i,2}, ..., r_{i,L}\}$, where $r_{i,j}$ denotes the j-th candidate. Here each utterance $u_{i,j}$, knowledge $k_{i,j}$ and response $r_{i,j}$ is a sentence, see examples in Figure 1.

Besides, each sample has two labels. $y_i^k \in [1, M]$ is a knowledge label: y_i^k=j denotes that $k_{i,j}$ is the correct knowledge for grounding, e.g. red-font knowledge in Figure 1. $y_i^r \in [1, L]$ is a response label: y_i^r=j denotes that $r_{i,j}$ is the correct response of C_i, e.g. blue-font utterances in Figure 1.

Formally, we define three input sources: the context C, the knowledge base K and the response candidates R. Then, we assign two tasks to learn: *knowledge selection* as the auxiliary task and *response selection* as the main task. In *knowledge selection*, we learn $f_1(k|c) \rightarrow y^k$ that measures the matching degree between C and each k_j in K, thus a proper knowledge k_s with maximal matching degree is selected for grounding. In *response selection*, we learn $f_2(r|c, k_s) \rightarrow y^r$ that measures the matching degree between C and each r_j in R given k_s. Finally, a proper response r_s is selected.

4 *MNDB* MODEL

Figure 2 outlines the framework of our ***MNDB*** to **m**odel **n**atural **d**ialog **b**ehaviors. Based on the task definition in Section 3, ***MNDB*** can be decomposed as two stages for the two tasks. In the first stage, we propose a flow-based knowledge selection network to perform the matching between C and K, which selects a proper knowledge k_s. In the second stage, we propose a ternary-grounding response selection network to perform the matching among R, C and k_s, thus a best response r_s is selected. The two networks are linked by two conversation flows along with the direction of context.

4.1 Flow-based Knowledge Selection

Given the context C and the knowledge base K, the objective of this stage is to determine the matching score between C and each knowledge candidate k_j in K. Thus, the proper knowledge k_s can be selected for the next stage. The implementation of the knowledge selection network is shown in Figure 3. The contributions of this network towards traditional conversation models such as Transformer model [2] lies in three points. First, a conversation flow is constructed, in which deep matching between each utterance and knowledge is performed, and the dynamic matching signals are aggregated via a RNN for precise knowledge selection. Second, to model the responses that no knowledge is used, a sentinel sentence denoting null knowledge, i.e. k_0 in Figure 3, is added into the knowledge base so that the model decides to select whether the real knowledge or the sentinel. Third, after we select a knowledge

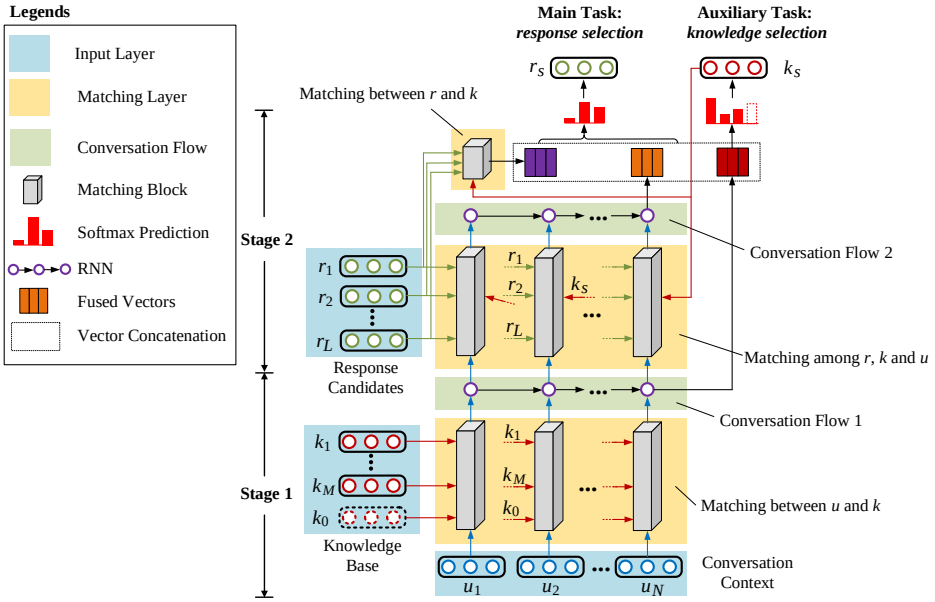


Figure 2: The *MNDB* model (Best viewed in color).

k_s, the aggregated matching signal between C and k_s, called $v^{c\text{-}k}$, is also injected to the next stage for better response selection. The knowledge selection network can be decomposed as follows:

4.1.1 Matching between Utterance and Knowledge. We take each utterance u_i and the knowledge candidate k_j as a pair and feed it to the *u-k matching block* to produce the matching vector $h_{i,j}$.The *u-k matching block* is shown in the bottom of Figure 3. It can be divided into three layers.

1. Encoding Layer. Giving a u containing $|u|$ terms and a k containing $|k|$ terms, we represent their word embeddings as $E^u = [e^{u,1}, ..., e^{u,|u|}]$ and $E^k = [e^{k,1}, ..., e^{k,|k|}]$. $\forall i \in \{1, ..., |u|\}$ and $\forall j \in \{1, ..., |k|\}$, $e^{u,i}$ and $e^{k,j}$ denotes the i-th term of u and the j-th term of k, respectively. $E^u \in \mathbb{R}^{|u| \times d}$ and $E^k \in \mathbb{R}^{|k| \times d}$, where d denotes the dimension of word embedding. The embedding of the sentinel knowledge k_0 is denoted as $E^{k_0} = [e^0]$, where $e^0 \in \mathbb{R}^{1 \times d}$ is randomly initialized and learned in the training.

Then, we feed E^u and E^k into a self attention layer to produce the representations U^u and U^k. The self attention layer uses the single-head attention module defined in Transformer [21], which consists of a scaled dot-product attention component and a feed-forward component with two non-linear projections, where a residual connection and a row-wise normalization are applied to the result of each projection. For ease of presentation, we denote the whole attention layer as $f_{ATT}(\cdot, \cdot, \cdot)$, and U^u and U^k are calculated as:

$$U^u = f_{ATT}(E^u, E^u, E^u), \quad U^k = f_{ATT}(E^k, E^k, E^k), \tag{1}$$

where $U^u \in \mathbb{R}^{|u| \times d}$ and $U^k \in \mathbb{R}^{|k| \times d}$.

2. Interaction Layer. In this layer, U^u and U^k are interacted with each other for information alignment and enrichment via a cross attention module. Specifically, the single-head attention module [21] is again used here to obtain knowledge-aware utterance representation Z^u and utterance-aware knowledge representation Z^k:

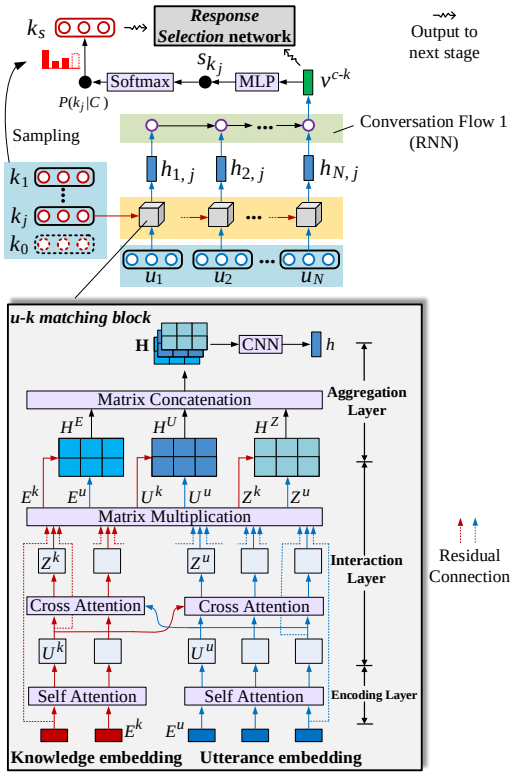


Figure 3: Knowledge selection network.

$$Z^u = f_{ATT}(U^u, U^k, U^k), \quad Z^k = f_{ATT}(U^k, U^u, U^u), \tag{2}$$

where $Z^u \in \mathbb{R}^{|u|\times d}$ and $Z^k \in \mathbb{R}^{|k|\times d}$.

Based on that, we construct three interaction matrices by matrix multiplication operations:

$$H^E = \frac{E^k \cdot (E^u)^{\mathrm{T}}}{\sqrt{d}}, \quad H^U = \frac{U^k \cdot (U^u)^{\mathrm{T}}}{\sqrt{d}}, \quad H^Z = \frac{Z^k \cdot (Z^u)^{\mathrm{T}}}{\sqrt{d}}, \tag{3}$$

where $H^E, H^U, H^Z \in \mathbb{R}^{|k|\times|u|}$.

3. Aggregation Layer. The three matrices are then concatenated into a 3-D matching tensor $\mathbf{H} \in \mathbb{R}^{|k|\times|u|\times 3}$:

$$\mathbf{H} = H^E \oplus H^U \oplus H^Z, \tag{4}$$

where $\oplus$ denotes a matrix concatenation operation. Inspired by [20], we use a CNN network $f_{CNN}(\cdot)$ to extract matching signals from $\mathbf{H}$ and output a d-dimensional matching vector h:

$$h = f_{CNN}(\mathbf{H}). \tag{5}$$

Here, for each knowledge candidate k_j given the context C, we can obtain a list of matching signals $(h_{1,j}, ..., h_{N,j})$ where $h_{i,j}$ denotes the signal from u_i and k_j.

4.1.2 Signal Aggregation on Conversation Flow. We feed the list of matching signals $(h_{i,j}, ..., h_{N,j})$ to a RNN network to construct the conversation flow, where each matching signal $h_{i,j}$ between k_j and u_i, is considered as a conversation node:

$$v_{i,j} = f_{RNN}(h_{i,j}, v_{i-1}), \tag{6}$$

where $v_{i,j}$ is the i-th hidden state, and $v_{0,j}$ is randomly initialized.

We consider the last hidden state $v_{N,j}$ as the aggregated matching signal that represents the semantic relatedness between k_j and C. For the selected knowledge k_s below, its aggregated matching signal $v_{N,s}$ is denoted as $v^{c\text{-}k}$, and will be sent to the next stage for boosting response selection, see Section 4.2.3.

The matching score s_{k_j} is then calculated by feeding $v_{N,j}$ to a MLP layer with a ReLU activation: $s_{k_j} = ReLU(v_{N,j} \cdot \mathbf{w_k} + b_k)$, where $\mathbf{w_k} \in \mathbb{R}^{d\times 1}$ and $b_k \in \mathbb{R}$ are trainable parameters.

Here, we obtain the matching scores between C and each knowledge candidate in K as $(s_{k_0}, s_{k_1}, ..., s_{k_M})$, where s_{k_0} denotes the matching score of the sentinel knowledge k_0.

4.1.3 Knowledge Sampling. We use a softmax layer to transform each matching score s_{k_j} to its probability form as $P(k_j|C)$:

$$P(k_j|C) = \frac{exp(s_{k_j})}{\sum_{i=0}^{M} exp(s_{k_i})}. \tag{7}$$

Based on the above distributions, we output the selected knowledge k_s to the next stage by weighted-summing the word embeddings of all knowledge pieces in a knowledge base [4].

$$E^{k_s} = \sum_{j=0}^{M} P(k_j|C) \cdot E^{k_j}. \tag{8}$$

[4]Note that we have also tried to use Gumbel-Softmax re-parametrization [5] for knowledge sampling. But it fails to improve performance.

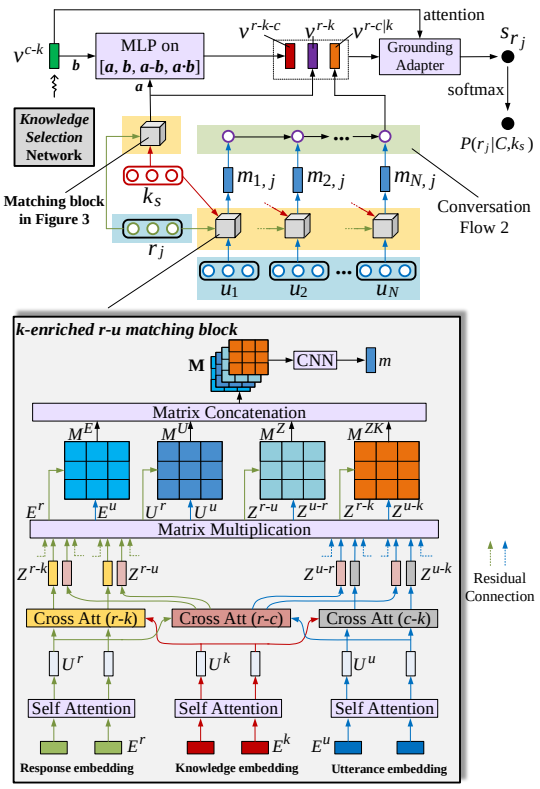


Figure 4: Response selection network (Best viewed in color).

4.2 Ternary-Grounding Response Selection

In this section, we propose a ternary-grounding response selection network. Given the context C, the selected knowledge k_s and the response candidate pool R, the objective is to determine the matching score between C and each response candidate r_j in R, with the grounding knowledge k_s.

Figure 4 demonstrates the response selection network. Based on the investigation in Section 2, we extract three complementary matching signals for each response candidates r_j. The first signal $v^{r\text{-}c|k}$ called *knowledge-enriched matching* is from the matching between C and r_j enriched by k_s for modeling the responses in Modes 1 and 2, i.e. the knowledge is used to continue the current topic. Note that the knowledge could be a sentinel denoting null knowledge. The second signal $v^{r\text{-}k}$ *knowledge-centered matching* is from the matching between k_s and r_j for modeling the responses in Mode 3 that the knowledge is incorporated to start a new topic. The third signal $v^{r\text{-}k\text{-}c}$ *knowledge-bridged matching* is to measure the distance between $v^{r\text{-}k}$ (the semantic relatedness between r_j and k_s) and $v^{c\text{-}k}$ (the semantic relatedness between C and k_s), where k_s is used to bridge the implicit relations between C and r_j. It can be considered to model Mode 2 in another perspective.

4.2.1 Knowledge-Enriched Matching between Response and Context. We take each utterance u_i, the selected knowledge k_s and the response candidate r_j as a triple and feed it to the *k-enriched r-u matching block* to produce the matching vector $m_{i,j}$. The *k-enriched*

r-u matching block is shown in the bottom of Figure 4. Here we briefly introduce it when it handles a triple (u, k, r).

First, it uses the same encoding layer as in Section 4.1.1. So, we can get the word embeddings of u, k and r as E^u, E^k and E^r, respectively. Similarly, they are converted to U^u, U^k and U^r by a self attention layer in Section 4.1.1.

Second, the pairs of (U^u, U^k), (U^r, U^k) and (U^u, U^r) are mutually interacted with each other via three cross attention modules defined in Section 4.1.1:

$$Z^{u\text{-}k} = f_{ATT}(U^u, U^k, U^k),\ Z^{r\text{-}k} = f_{ATT}(U^r, U^k, U^k),$$
$$Z^{u\text{-}r} = f_{ATT}(U^u, U^r, U^r),\ Z^{r\text{-}u} = f_{ATT}(U^r, U^u, U^u), \quad (9)$$

where $Z^{u\text{-}k}$ and $Z^{u\text{-}r} \in \mathbb{R}^{|u|\times d}$, $Z^{r\text{-}k}$ and $Z^{r\text{-}u} \in \mathbb{R}^{|r|\times d}$, and $|r|$ denotes the length of r. Here $Z^{u\text{-}k}$ and $Z^{r\text{-}k}$ is the knowledge-enriched representation of u and r, respectively.

Third, we construct four matching matrices by matrix multiplication operations:

$$M^E = \frac{E^u \cdot (E^r)^{\mathrm{T}}}{\sqrt{d}}, \qquad M^U = \frac{U^u \cdot (U^r)^{\mathrm{T}}}{\sqrt{d}},$$
$$M^Z = \frac{Z^{u\text{-}r} \cdot (Z^{r\text{-}u})^{\mathrm{T}}}{\sqrt{d}}, \qquad M^{ZK} = \frac{Z^{u\text{-}k} \cdot (Z^{r\text{-}k})^{\mathrm{T}}}{\sqrt{d}}, \quad (10)$$

where M^E, M^U, M^Z and $M^{ZK} \in \mathbb{R}^{|u|\times|r|}$. Unlike in Eq. (3), here M^{ZK} means that the interaction between u and r is enriched by k.

Then, we copy the operations in Eqs. (4, 5) on the four matrices, and get a list of matching signals $(m_{1,j}, ..., m_{N,j})$ for each response candidate r_j given the context C and the knowledge k_s.

Last, like in Eq. (6), we again use a RNN network to aggregate these signals and obtain the signal $v_j^{r\text{-}c|k} \in \mathbb{R}^d$.

4.2.2 Knowledge-Centered Matching between Response and Knowledge. We consider two strategies to obtain the signal $v_j^{r\text{-}k} \in \mathbb{R}^d$.

The first strategy is to feed each response candidate r_j and the knowledge k_s to the *u-k matching block* in Section 4.1.1, i.e. by getting the vector h in Eq. (5) with r_j replacing u_i, thus $v_j^{r\text{-}k} = h$. We call it the context-independent matching.

The second strategy is to append r_j to the tail of C as $C' = (u_1, ..., u_N, r_j)$, and feed C' and k_s to the entire knowledge selection network, i.e. by getting the vector $v_{N+1,s}$ in Eq. (6) with C' replacing C, thus $v_j^{r\text{-}k} = v_{N+1,s}$. We call it the context-dependent matching.

4.2.3 Knowledge-Bridged Matching between Response and Context. The signal $v^{r\text{-}c|k}$ in Section 4.2.1 represents the explicit similarity between r_j and C given the grounding knowledge k_s. Here we additionally propose the signal $v^{r\text{-}k\text{-}c}$ that represents the implicit similarity between r_j and C from another complementary perspective: we take k_s as an bridge and measure the distance between $v^{r\text{-}k}$ and $v^{c\text{-}k}$.

Here we explain the intuition that underlies the different designs of *knowledge-bridged matching* and *knowledge-enriched matching*.

- The *knowledge-enriched matching*, i.e., using knowledge to enrich the representation of context and response, is derived from current sota models as a common technique. It measures the explicit similarity between response and context given the grounding knowledge.
- The *knowledge-bridged matching*, on the other hand, is firstly proposed by us which can be regarded as a complementary way to use knowledge. It takes the knowledge as an anchor/bridge and measures the implicit similarity between response and context by calculating the distance between two similarities: the similarity of response-knowledge pair and the similarity of context-knowledge pair. It implies that if the distance between the response-knowledge pair is similar to the distance between the response-knowledge pair, then the response and the context are similar as well.

Formally, we combine $v^{r\text{-}k}$ and $v^{c\text{-}k}$ by concatenating the difference and element-wise product vectors with the original vectors, and then feed the fused vector to a MLP layer with ReLU activation:

$$v_j^{r\text{-}k\text{-}c} = ReLU([v_j^{r\text{-}k}; v_s^{c\text{-}k}; v_j^{r\text{-}k} - v_s^{c\text{-}k}; v_j^{r\text{-}k} \odot v_s^{c\text{-}k}] \cdot \mathbf{W_r} + \mathbf{b_r}), \quad (11)$$

where $\mathbf{W_r} \in \mathbb{R}^{4d\times d}$ and $\mathbf{b_r} \in \mathbb{R}^d$ are trainable parameters, $[\cdot;\cdot]$ denotes the vector concatenation operation, and $\odot$ denotes the element-wise production operation.

4.2.4 Response Selection by Fusing Signals. To mimic the human behaviors of switching different conversation modes adaptively, we calculate the matching score s_{r_j} by proposing an attention based grounding adapter *ga* that fuses the three signals based on current context and knowledge: $s_{r_j} = ReLU(ga \cdot \mathbf{w_o} + b_o)$, where $\mathbf{w_o} \in \mathbb{R}^{d\times 1}$ and $b_o \in \mathbb{R}$ are trainable parameters. The adapter $ga \in \mathbb{R}^d$ is calculated by:

$$ga = w_j^{r\text{-}k} \cdot v_j^{r\text{-}k} + w_j^{r\text{-}c|k} \cdot v_j^{r\text{-}c|k} + w_j^{r\text{-}k\text{-}c} \cdot v_j^{r\text{-}k\text{-}c}, \quad (12)$$

where $w_j^{r\text{-}k}, w_j^{r\text{-}c|k}$ and $w_j^{r\text{-}k\text{-}c}$ denote the importance of each signal and $w_j^{r\text{-}k} + w_j^{r\text{-}c|k} + w_j^{r\text{-}k\text{-}c} = 1$. The importance is derived from measuring the similarity between each signal, e.g. $v_j^{r\text{-}k}$, and the knowledge-aware representation of the last utterance, i.e. $v^{c\text{-}k}$ in Section 4.1.2. For example, $w_j^{r\text{-}k} \propto exp\left[v^{c\text{-}k} \cdot (v_j^{r\text{-}k})^{\mathrm{T}}\right]$ and $w_j^{r\text{-}c|k} \propto exp\left[v^{c\text{-}k} \cdot (v_j^{r\text{-}c|k})^{\mathrm{T}}\right]$.

So, we obtain the matching scores between C and all response candidates as $(s_{r_1}, ..., s_{r_L})$. Next, we use a softmax layer to transform each matching score s_{r_j} to its probability form $P(r_j|C, k_s) = \frac{exp(s_{r_j})}{\sum_{i=1}^{L} exp(s_{r_i})}$.

Last, we select the final response r_s from candidates that $r_s = argmax_j \left[P(r_j|C, k_s)\right]$.

4.3 Model Training

To train the ***MNDB*** model, we define the following loss functions:

$$\mathcal{L}_K = -log \prod_{(C,r^+,k^+)} P(k^+|C),\ \mathcal{L}_R = -log \prod_{(C,r^+,k^+)} P(r^+|C, k^+), \quad (13)$$

where (C, r^+, k^+) denotes a positive instance that k^+ is the correct knowledge for grounding C, and r^+ is the correct response of C given k^+.

The two losses $\mathcal{L}_K$ and $\mathcal{L}_R$ are for the tasks of *knowledge selection* and *response selection*, respectively. Then, we jointly optimize the two losses when learning our ***MNDB*** model: $\mathcal{L}_{total} = \mathcal{L}_K + \mathcal{L}_R$. [5]

5 EXPERIMENTS

5.1 Experiment Settings

Evaluation Tasks. We evaluate our model on two tasks. The first task *knowledge selection* is to investigate the ability to select proper knowledge. The second task *full AKGC* is to evaluate the model to select proper responses for the whole conversation capability.

Datasets. The first benchmark is Persona-Chat [27] collected by two workers to chat according to their respective profiles as knowledge. We use it in the exact way reported in [28]. For example, two data versions are used: Original Persona and Revised Persona that the knowledge, i.e. profiles, has and does not have word overlap with utterances, respectively. Person-Chat is only used for the *full AKGC* task as it only contains response labels. The second benchmark Wiki-Wizard [2] is a open-domain conversation dataset grounded by Wikipedia knowledge. When using Wiki-Wizard we exactly follow the settings defined in the original paper [2]. For example, knowledge is provided in two settings: given the ground truth knowledge, or where the model needs to predict which knowledge to use. Besides, the test set is split into two subsets: Test Seen and Test Unseen, where the former contains certain overlapping topics with the training set with new dialogs about those topics, and the latter contains some topics never seen before in train or validation. Wiki-Wizard is used for both two tasks.

Baselines. We compare some retrieval-based baselines, most of which have been reported in the papers [2, 28]. Note that generation-based ones such as [6–8, 10, 14, 16] are not compared because our model is retrieval-based.

(1) **Profile Memory** and **KV Profile Memory** [27]. They use a memory network or key-value memory network to encode context, then perform attention over knowledge to enrich context representation, and measure cosine distance between context and response.
(2) **Transformer** [9]. It uses the Transformer [21] to encode conversations. Then like Profile Memory, it performs attention over knowledge to enrich context representation, and measure cosine distance between context and response.
(3) **DGMN** [28]. Document-grounded matching network by hierarchical attention, that is state-of-the-art on Persona-Chat dataset.
(4) **BoW MemNet** [17]. It is equivalent to Profile Memory but using bag-of-words representation, as reported in [2].
(5) **Pre-trained Transformer** [9]. The baseline (2) pre-trained on a large-scaled conversation data Reddit [9].
(6) **Transformer MemNet** [2]. A Transformer based memory network, that is pre-trained on Reddit data [9], with multi-task learning on SQuAD data [15]. It is the state-of-the-art model on Wiki-Wizard [2].

There are two major differences between the baselines and our model. First, all baselines except DGMN encode multi-turn context into a fixed-length vector. Second, all baselines use knowledge to merely enrich the context representation and then perform matching between context and response.

[5]For Persona-Chat dataset [27] in Section 5, our model is trained by only optimizing $\mathcal{L}_R$, because no knowledge label is provided.

Table 2: Performance on the *knowledge selection* task on Wiki-Wizard [2].

Indicators(%)	Test Seen		Test Unseen	
	r@1	F1	r@1	F1
BoW MemNet [2]	23.0	36.3	8.9	22.9
Transformer [2]	22.5	33.2	12.2	19.8
Pretrained Transformer [2]	24.5	36.4	23.7	35.8
Transformer MemNet[2]	46.8	51.1	32.4	41.1
MNDB-D	48.9	55.1	36.0	42.7
MNDB-I	**50.8**	**57.6**	**36.3**	**45.7**

Note that the performance of some baselines, i.e. (1-3) on Persona-Chat and (4-6) on Wiki-Wizard have been reported in [2, 28], so we do not need to re-implement them. For the baselines (1-3) on Wiki-Wizard and (4-6) on Persona-Chat, we re-implement them and tune their parameters individually for a fair comparison. The baselines (1) and (3) ares only for the *full AKGC* task because they can not select knowledge [28].

Here our model has two variants: ***MNDB-D*** using the context-dependent matching strategy for obtaining the *knowledge-centered matching* signal $v^{r\text{-}k}$, and ***MNDB-I*** using the context-independent strategy, see Section 4.2.2. Appendix A will describe the implementation details, e.g. hyper-parameter configurations of our model.

Evaluation metrics. We use the same metrics just as in [2, 28] for easy comparison. Specifically, for both tasks on Wiki-Wizard, we use r@1 and F1 [2] to evaluate the models. On Persona-Chat, $r@k$ is selected as evaluation metrics where k=1, 2 and 5 [28].

5.2 Evaluation on Knowledge Selection Task

The task is to select the gold knowledge from M candidates (M=32). The results are shown in Table 2. Our model achieves the best performance compared with all baselines. Such significant superiority is due to our flow-based network defined in Section 4.1, where each utterance-knowledge pair is matched along with the conversation flow. Thus, the multi-turn semantic information is preserved.

However, other baselines including pre-trained ones in Table 2 get much lower indicators. The reason behind lies in that they condense the multi-turn context representations into a fixed-length vector before attending the knowledge for selection. Therefore, the information loss of multi-turn context is inevitable, which is vital for accurate knowledge selection. Note that in Tables 2,3 and 4, numbers in bold mean that our improvement over the best baseline is statistically significant (t-test, p-value < 0.01).

5.3 Evaluation on Full AKGC TASK

The task is to select the correct response from L candidates. For Wiki-Wizard, L = 100. For Persona-Chat, L = 20. The results are shown in Tables 3 and 4. It is clear that on both datasets, our model achieves the best performance. Especially on Wiki-Wizard, our model even outperforms pre-trained models.

Such an evident improvement over baseline comes from two aspects. First, our model has better knowledge selection ability as proved in Section 5.2, which boosts the subsequent performance of

Table 3: Performance on the *full AKGC* task on Wiki-Wizard [2]. The baselines without reference are implemented by us.

	Predicted Knowledge				Gold Knowledge	
	Test Seen		Test Unseen		Seen	Unseen
Indicators(%)	r@1	F1	r@1	F1	r@1	r@1
Profile Memory	67.1	14.8	32.2	10.4	80.1	58.8
KV Profile Memory	68.3	15.0	33.1	10.7	81.4	60.5
Transformer	70.6	14.2	50.8	12.3	87.7	69.4
DGMN	82.4	15.3	60.2	11.3	89.5	71.8
BoW MemNet [2]	71.3	15.6	33.1	12.3	84.5	66.7
Pretrained Transformer [2]	80.2	15.1	63.6	12.2	89.2	77.4
Transformer MemNet [2]	87.4	15.4	69.8	12.4	92.3	83.1
MNDB-D	89.6	16.1	74.4	**13.1**	92.7	**86.6**
MNDB-I	**92.1**	**16.3**	**76.2**	**13.4**	**93.2**	86.5

Table 4: Performance on the *full AKGC* task on Persona-Chat [27].

	Original Persona			Revised Persona		
Indicators(%)	r@1	r@2	r@5	r@1	r@2	r@5
Profile Memory[27]	50.9	60.7	75.7	35.4	48.3	67.5
KV ProfMem [27]	51.1	61.8	77.4	35.1	45.7	66.3
Transformer [9]	54.2	68.3	83.8	42.1	56.5	75.0
DGMN [28]	67.6	80.2	92.9	58.8	62.5	87.7
BoW MemNet	57.4	64.7	80.4	48.9	51.3	75.2
Pretrained Transformer	65.8	74.1	86.6	50.5	57.4	81.6
Transformer MemNet	67.0	79.4	90.4	56.4	61.7	84.3
MNDB-D	**75.6**	**86.9**	95.7	**73.6**	83.0	**95.2**
MNDB-I	74.5	85.8	**95.9**	72.1	**83.9**	95.2

Table 5: Ablation analysis of our model on r@1.

	Persona-Chat		Wiki-Wizard (Predicted k)	
Datasets	Original	Revised	Seen	Unseen
MNDB-I	**74.5**	**72.1**	**92.1**	**76.2**
[a] No $v^{r\text{-}c|k}$	31.0	25.1	83.9	67.4
[b] No $v^{r\text{-}k}$	71.2	70.3	88.9	72.2
[c] No $v^{r\text{-}k\text{-}c}$	71.6	68.7	90.6	74.4
[d] No ga	72.2	70.8	90.2	75.1

a Without the *knowledge-enriched matching* signal in Section 4.2.1.
b Without the *knowledge-centered matching* signal in Section 4.2.2.
c Without the *knowledge-bridged matching* signal in Section 4.2.3.
d Without the *grounding adapter* in Section 4.2.4 and instead fusing the three signals by concatenating them.

response selection. Second, our model uses knowledge in a diversified manners by proposing three novel matching signals, which will be further validated in the ablation study section, and thus can adapt to corresponding conversation modes.

More interestingly, in a detailed perspective we find that for Wiki-Wizard in Table 3, our model gains more improvement (e.g. 92.1% $v.s.$ 87.4% on r@1) on the Predicted Knowledge setting, where models must select relevant knowledge from M knowledge candidates. While on the Gold Knowledge setting that the gold knowledge is already provided, the improvement of our model is moderate, especially on the Seen test set (93.2% $v.s.$ 92.3% on r@1). Note that the Gold Knowledge setting can be considered as a upper bound for models that will not be reached in real-world scenarios. Furthermore, our model achieves higher performance gain on the Unseen Test than that on the Seen Test. This phenomena indicate that our model is more powerful in real-world applications, where gold knowledge will NOT be provided in advance and the training data can not *see* the topics appeared in the test set.

Similar phenomena can also be found on Persona-Chat. In Table 4, our model outperforms the strong model DGMN by a much larger margin on the Revised Persona setting (e.g. 73.6% $v.s.$ 58.8% on r@1) than that on the Original Persona setting. Therefore, it is clear that our model is consistently robust on different public datasets.

Last, we find that the performance of ***MNDB-D*** is better than ***MNDB-I*** on Persona-Chat, but is worse on Wiki-Wizard. We will discuss this discrepancy in Section 5.5.

5.4 Ablation Study

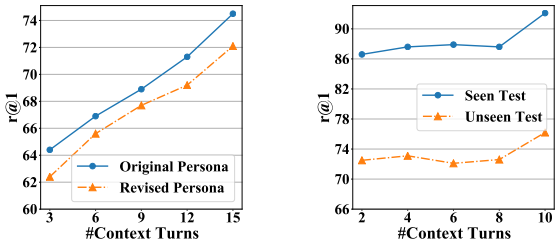
To further understand the superiority of our model, we conduct a model ablation analysis on both datasets. As different indicators are positively correlated, we only report r@1 in Table 5.

First, all the ablated model parts, i.e. the three matching signals and the grounding adapter, contribute to our final model. Second, the contribution rate of different parts can be ranked as $v^{r\text{-}c|k} > v^{r\text{-}k} > v^{r\text{-}k\text{-}c} > ga$. Specifically, $v^{r\text{-}c|k}$ as the *knowledge-enriched matching* signal is the most important signal for response selection, which is consistent with the statement in Section 2 that in all response modes the response must be semantically consistent with the context.

By contrast, without either $v^{r\text{-}k}$ or $v^{r\text{-}k\text{-}c}$, our model is still able to achieve good results, which means that $v^{r\text{-}k}$ and $v^{r\text{-}k\text{-}c}$ are not as decisive as $v^{r\text{-}c|k}$, but can improve the final performance. Note that in Table 1 the tiny ratio of Mode 3, i.e. 10%, may also answer why $v^{r\text{-}c|k}$ corresponding to Modes 1 and 2 dominates the performance gain. Last, the grounding adapter ga can further promote the model performance because it endows the model with the ability of adaptively switching between different modes. Note that all baselines only model the *knowledge-enriched matching* signal, so they can not well adapt to diversified conversations.

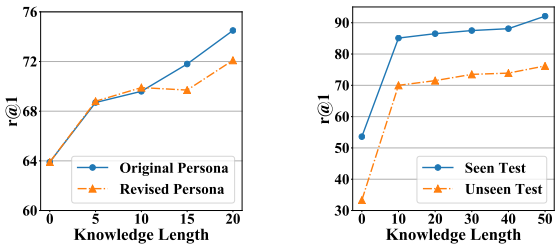
Interestingly, without $v^{r\text{-}c|k}$, r@1 on Persona-Chat drops drastically (74.5% $v.s.$ 31.0%), but drops moderately on Wiki-Wizard (92.1% $v.s.$ 83.9%). We discuss it in Section 5.5.

More detailed case studies about the capacity on adapting to three conversation modes can be found in Appendix C.



(a) Persona-Chat (b) Wiki-Wizard

Figure 5: Performance of *MNDB-I* with different N.



(a) Persona-Chat (b) Wiki-Wizard

Figure 6: Performance of *MNDB-I* with different $|k|$.

5.5 Discussions

We test the sensitivity of our model on two hyper-parameters: *#Context Turns* , i.e. the number of utterance turns, and *Knowledge Length*, the number of terms in a knowledge sentence. Then we analyze the discrepancy of performance between two datasets.

Performance Analysis on #Context Turns. We study how *#Context Turns* will affect the performance of our model. The results are shown in Figure 5. First, on Persona-Chat our model is good at handling contexts with long turns, as it gets better performance on longer context. On Wiki-Wizard, however, the gain of increasing *#Context Turns* is marginal, although the overall trend is up. It means that on Wiki-Wizard, the context information is not as significantly important as on Persona. The explanation is below.

Performance Analysis on Knowledge Length. We study the relationship between the performance of our model and *Knowledge Length*. As shown in Figure 6, when no knowledge is provided, i.e. *Knowledge Length* is equal to zero, the performance of the model on both benchmarks drops sharply. It indicates that knowledge plays a vital role on the AKGC task. Then, as *Knowledge Length* increases, e.g. $|k| \geq 5$ on Persona-Chat and $|k| \geq 10$ on Wiki-Wizard, the performance quickly reaches a high level, that means that our model can effectively incorporates the information in knowledge. Note that the average *Knowledge Length* of Persona-Chat and Wiki-Wizard is 7.3 and 23.4, respectively.

Discrepancy between Two Datasets. In the experiments above, the performance of our model variants on two datasets are both superior to all baselines. However, there are some discrepancies:

(1) In Table 3, ***MNDB-I*** obtains higher indicators than ***MNDB-D*** on Wiki-Wizard. But in Table 4 ***MNDB-I*** obtains lower indicators than ***MNDB-D*** on Persona-Chat.
(2) In Table 5, the *knowledge-enriched matching* signal called $v^{r\text{-}c|k}$ seems to be more important on Persona-Chat than that on Wiki-Wizard.
(3) In Figure 5, the performance of our model on Persona-Chat is more sensitive to *#Context Turns* than that on Wiki-Wizard.

Towards these discrepancies, the reason behind might be that in Persona-Chat, the conversations are more like multi-turn chit-chats between two equal persons such as friends. However, the conversations in Wiki-Wizard are more like single-turn QAs [25, 26] between a teacher and a student. Thus, the context semantics in Persona-Chat is more rich and utterances in different turns are tightly connected within the dialog. On the other hand, different turns in Wiki-Wizard are relatively independent with each other.

6 RELATED WORK

Traditional conversation models can be classified as retrieval-based and generation-based. The retrieval-based models select a response from a response candidate pool by matching the relations between the dialog history and candidates, such as [19, 20, 30, 32, 34] that perform multi-turn matching, i.e. each utterance is matched with the response. The generation-based models synthesize a response based on sequence to sequence networks, such as [1, 11, 18, 23, 31]. In real-world scenarios, retrieval-based models are superior to generation-based ones ion response fluency and diversity [28].

However, traditional models suffer from producing less-informative responses because they are unaware of background knowledge. To this end, knowledge-grounded conversation models have been proposed, benefited from the open source of massive knowledge-grounded conversation datasets, such as Persona-Chat [27], Wiki-Wizard [2], Holl-E [12], CMUDoG [33] and Topical-Chat [4]. Those models can also be grouped into generation-based models [6–8, 10, 13, 14, 16, 24, 29] and retrieval-based models [2, 3, 22, 28]. Among them, [22] proposed a proactive conversation dataset and corresponding methods that explicit conversation goals are provided to guild the model to shift the conversation topic. [16] introduced a Global-to-Local mechanism for adaptive knowledge selection. [8] fused two types of knowledge: Knowledge Graphs (KGs) and texts, and formulated knowledge selection as a multi-hop graph reasoning process on KGs for explainable response generation. [28] proposed a document-grounded matching model (DGMN) which achieved state-of-the-art performance on Persona-Chat [27] as a non-pre-trained retrieval model. DGMN adopts the attention-based text matching mechanism to calculate the semantic similarity between two sentences, so that the representations of utterances and knowledge are mutually grounded. Then, the enriched representations of utterances and knowledge are matched with each response candidate. [2] proposed a Transformer Memory Networks, pre-trained on Reddit data [9] and with multi-tasking on SQuAD data [15], and achieved state-of-the-art results on Wiki-Wizard [2].

In short, previous studies fused conversation context and external knowledge into the model which only focus on solving some special conversation problems in local views. However, few researches were dedicated to the conversation capability in a global view. For instance, they fail to adaptively incorporate the knowledge to mimic human dialog behaviors in which different conversation patterns can be switched freely. To this end, we define and solve

"Adaptive Knowledge-Grounded Conversations" (AKGCs), where the context-aware knowledge is incorporated for handling diversified natural conversation patterns.

7 CONCLUSION

We developed a neural model, called ***MNDB***, to **m**odel **n**atural **d**ialog **b**ehaviors for multi-turn response selection. In ***MNDB***, we investigated the essence of natural conversation and summarized some key conversation patterns. Based on that, ***MNDB*** embraced two novel networks for better response selection. In the first stage, a flow-based matching network was proposed for accurately selecting knowledge. Then in the second stage, a ternary-grounding response selection network was presented, in which three complementary matching signals were extracted and fused for final response selection. Evaluation on two benchmarks demonstrated the advantage of our model.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (61672501, 71329201, 61502466) and Alibaba Group through Alibaba Innovative Research Program.

REFERENCES

[1] Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, and Jie Zhou. 2020. Bridging the Gap between Prior and Posterior Knowledge Selection for Knowledge-Grounded Dialogue Generation. In *EMNLP*.
[2] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241* (2018).
[3] Jatin Ganhotra, Siva Sankalp Patel, and Kshitij Fadnis. 2019. Knowledge-incorporating ESIM models for response selection in retrieval-based dialog systems. In *AAAI Workshops*.
[4] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019* (2019), 1891–1895.
[5] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
[6] Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *ACL*.
[7] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911* (2019).
[8] Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with reasoning on augmented graph. In *EMNLP*.
[9] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984* (2018).
[10] Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. RefNet: A reference-aware network for background based conversation. *arXiv preprint arXiv:1908.06449* (2019).
[11] Chuan Meng, Pengjie Ren, Zhumin Chen, weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A Dual Knowledge Interaction Network for Knowledge-grounded Conversation. In *SIGIR*.
[12] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. *arXiv preprint arXiv:1809.08205* (2018).
[13] Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *ACL*.
[14] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *ACL*.
[15] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).
[16] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-Local knowledge selection for background based conversation. In *AAAI*.
[17] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *NIPS*. 2440–2448.
[18] Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! Learning towards effective responses with multi-Head attention mechanism.. In *IJCAI*. 4418–4424.
[19] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *WSDM*. ACM, 267–275.
[20] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *ACL*. 1–11.
[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*. 5998–6008.
[22] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goals. In *ACL*.
[23] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*.
[24] Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge Graph grounded goal planning for open-domain conversation generation. In *AAAI*.
[25] Chen Zhang, Hao Wang, Liang Zhou, Yijun Wang, and Can Chen. 2019. Machine Comprehension-Incorporated Relevance Matching. In *ICDM*.
[26] Chen Zhang, Xuanyu Zhang, and Hao Wang. 2019. A Machine Reading Comprehension-based Approach for Featured Snippet Extractio. In *ICDM*.
[27] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243* (2018).
[28] Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A document-grounded matching network for response selection in retrieval-based chatbots. In *IJCAI*.
[29] Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. [n.d.]. Low-resource knowledge-grounded dialogue generation.
[30] Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-Guided Neural Conversational Model. In *AAAI*.
[31] Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts. In *AAAI*.
[32] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards Persona-Based Empathetic Conversational Models. In *EMNLP*.
[33] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358* (2018).
[34] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*. 1118–1127.

A IMPLEMENTATION DETAILS

A.1 General Configuration of Our Model

We implement our ***MNDB*** model with TensorFlow [6]. In the experiments, the word embeddings are initialized by using the pre-trained Glove word embeddings [7], and are fixed during training. The vocab size of Persona-Chat and Wiki-Wizard is 20,467 and 307,709, respectively. All the out-of-vocab (OOV) terms are mapped to an <UNK> token, whose embedding is trainable with random initialization. The dimension of word embeddings and the hidden size d are set to 300. We use the Adam optimizer ($\beta 1$=0.9, $\beta 2$=0.999, and ϵ=10^{-8}) [8], with a batch size of 32, a learning rate of 0.0005, a learning decay rate of 0.9 and a decay step of 2,000 (the minimal learning rate is set to 0.00005). The dropout rate is set to 0.2. We use gradient clipping with a maximum gradient norm of 10. When implementing the CNN layer in Section 4.1.1, we follow [20] and set the window size of convolution and pooling kernels as (3, 3), and the strides as (1, 1) and (3, 3) respectively. The number of convolution kernels is 32 in the first layer and 16 in the second layer. We adopt the GRU (Gated

[6] https://www.tensorflow.org/.
[7] http://nlp.stanford.edu/data/wordvecs/glove.840B.300d.zip.
[8] https://arxiv.org/abs/1412.6980.

Recurrent Unit) network to implement the RNN layer in Section 4.1.2. Early stopping on validation data is adopted as a regularization strategy. We observe that the model usually converges within 3~5 epochs over the entire training data. We follow the original data split for training, validation and test. In the experiments, the training process on Persona-Chat takes 1.8 hours per epoch on a Tesla P40 GPU. The training process on Wiki-Wizard takes 0.8 hours per epoch on a P40 GPU. The predicting calculation on a single sample containing 20 response candidates takes averagely 54 ms on a P40 GPU.

A.2 Configuration on Persona-Chat

Basically, we follow the same configuration and hyper-parameters setting reported in [28] for a fair comparison. The number of knowledge M per context, i.e. the persona profiles, is set to 5. The length of each knowledge sentence k, each utterance u, and each response candidate r is set to 20. Padding zeros are performed if the actual length is less than 20. Besides, we truncate the sentence with more than 20 terms and only keep the latest 20 ones. We set the length of a context, N, to 15, i.e. only the latest 15 turns of utterances are included. For each utterance, 19 negative response candidates and 1 positive one are provided by the publishers. The average number of context turns and knowledge length is 7.4 and 7.3, respectively. Following [28], there are10,907 conversations in Persona-chat, which we divide into 8,939 for train, 1,000 for validation and 968 for test. More benchmark details can be found in [28] and [27].

A.3 Configuration on Wiki-Wizard

The number of knowledge M per context is set to 32. For each utterance proposed by the wizard speaker, we randomly sample 32 knowledge as candidates from the wiki pages provided in Wiki-Wizard [9]. The length of each knowledge sentence k, each utterance u, and each response candidate r is set to 50, 20, 20, respectively. Like on Persona-Chat, truncation or zero-padding is applied when necessary. We set the length of a context, N, to 10. For each utterance proposed by the wizard speaker in the test set, 99 negative response candidates and 1 positive one are provided by the publishers. However, no negative candidates is provided for the training set, so we randomly sample 19 utterances from the whole dataset as negative candidates. The average number of context turns and knowledge length is 9.1 and 23.4, respectively. Following [2], in Wiki-Wizard, 201,999 utterance-response pairs of conversations are obtained and 166,787/17,715/17,497 of them are used for train/validation/test. The code for calculating evaluation metrics, i.e. r@1 and F1, is based on the Parlai platform.[9] More benchmark details can be found in [2].

B DISCOVERY OF CONVERSATION MODES

We propose the three modes by a three-stage process:

(1) First, three domain experts are incorporated to supervise this work, including one research professor focused on the conversation field, and two senior engineers from industries working on the virtual assistant design. Regular research discussion meetings and unscheduled idea exchange meetings with these experts have held since 2019, with the topic of how to make the AI conversation agents more intelligent. During these communications, the experts summarize some common conversation patterns based on their knowledge and experiences.
(2) After obtaining the initial conversation modes and their descriptions, we employ and train a crowd-sourcing team with 10 workers to annotate the conversation corpus, including Person-Chat, Wiki-Wizard, and the datasets collected from real-world dialog scenarios, e.g. the data from voice-powered AI speakers. The annotators are required to examine each conversation turn in the corpus, and then to determine the correct conversation mode to which each turn belongs.
(3) Based on the labeled data, lastly, the experts refine their proposed ideas and produce the final definition of the three conversation modes.

C CASE STUDY

As shown in Table 6, we list some examples from the test set of Persona-Chat and Wiki-Wizard to demonstrate the superiority of our model, which belong to various conversation modes mentioned in Section 2. Note that we remain some typos in Table 6, which had appeared in the original dataset. We detailedly explain the examples in Table 6 as follows.

First of all, all the six examples in Table 6 can be correctly predicted by our full ***MNDB-I*** model.

Second, the two examples of Mode 1 are from bad cases of the ablated model "No $v^{r\text{-}c|k}$" in Table 5, which is not equipped with the *knowledge-enriched matching* signal. It means that the explicitly matching between the response r and the context C is vital for solving Mode 1.

Third, the two examples of Mode 2 are from bad cases of the ablated model "No $v^{r\text{-}k\text{-}c}$" in Table 5, which does not incorporate the *knowledge-bridged matching* signal. It shows that the implicitly matching signal between r and C via the knowledge k as a bridge can enhance the performance of model on handling Mode 2.

Last, the two examples of Mode 3 are from bad cases of the ablated model "No $v^{r\text{-}k}$" in Table 5, which does not use the *knowledge-centered matching* signal. It indicates that the direct matching between r and k provides necessary signals for addressing Mode 3, where knowledge is used to start new topics. Note that all baselines in the experiments failed to address such two examples of Mode 3.

[9] We use the script from Parlai (https://parl.ai/).

Table 6: Examples of conversation modes in Wiki-Wizard [2] and Persona-Chat [27]. Mode 1: No knowledge is used. Mode 2: Knowledge is used to support current topics. Mode 3: Knowledge is used to start new topics. Sentences in bold denote the responses or the corresponding knowledge used for grounding.

	Example of Mode 1 in Persona-Chat	**Example of Mode 1 in Wiki-Wizard**
Person 1:	Hi, do you like bbq?	
Person 2:	Hello yes I love bbq.	
Person 1:	I love restaurants with bbq, they are a lot of fun.	
Person 2:	My poor toes are so webbed.	
Person 1:	Really? Afterwards, I go and play racquetball to loose some calories.	
Person 2:	Cool I like to paint.	
Person 1:	What do you do as a profession? I work as administrative assistant.	
Person 2:	I dig in people trash for things to build.	I love Bentley's a great British manufacturer and marketer of luxury cars.
Person 1:	Sounds unique but that is more of a hobby instead of a profession.	Me too, they are so stylish! How much does a Bentley cost these days?
Person 2:	True. I paint for a living.	Maybe $180,000.00. I like the Bentley Turbo R, whats your favourite model?
Person 1:	Good thing I am going to retire in 5 years, no more working for this guy!	I love all of them, but one of my favorites is the Flying Spur.
Person 2:	Awesome I have about thirty more years.	Nice. The R in the Turbo R stands for road holding to set it apart from the predecessor. Just a fun fact for you.
Person 1:	Well they go by fast, tell me more about yourself.	That's interesting! I wonder what their top speed is?
Person 2:	I am short fat and ugly. You?	I think its 155miles per hour.
Person 1:	I am tall dark and handsome.	Wow that is fast! What is your favorite color? I like the alpine green metallic.
Person 2:	**Lucky you to be so sexy.**	**Wow thats mine favourite colour also. How weird is that.**
Knowledge of Person 2:	I have webbed toes. I am an artist. My job is cleaning out cages at a research facility. I am five feet tall. I use other peoples trash for my projects.	Blue is one of the three primary colours of pigments in painting and traditional colour theory, as well as in the RGB colour model. It lies between violet and green on the spectrum of visible light. The eye perceives blue when observing light with a dominant wavelength between approximately 450 and 495 nanometres...
	Example of Mode 2 in Persona-Chat	**Example of Mode 2 in Wiki-Wizard**
Person 1:	Hello. How are you doing today?	
Person 2:	Doing well! Washing dishes after work. What about you?	
Person 1:	Thankfully not doing that! My best friend does those. He is a robot.	
Person 2:	Can not say that I have a robot friend! Would help at my job though.	
Person 1:	I do not work. I am in school homeschooled, actually.	
Person 2:	So you get to learn while in pajamas! neat!	
Person 1:	Indeed! Where are you from? I am from california.	
Person 2:	Nebraska. So there is not much to do here. I enjoy bird hunting here.	
Person 1:	Do you bird hunt a lot?	I work as a lifeguard at the beach.
Person 2:	Only when my company does not have me repairing properties, Yes!	I know a lifeguard watches out for safety and rescue, did you work at the ocean or a pool?
Person 1:	You have got to love that work!	I worked at a pool.
Person 2:	I do. Been doing it my whole life.	You have to be a strong swimmer to be a lifeguard, you must be able to swim really well.
Person 1:	Fantastic! What are your hobbies? My love is to cosplay.	Lifeguard.
Person 2:	Well, Fishing as well! Except I am not allowed to eat shellfish.	Are you required to have cpr training for your job?
Person 1:	Why Can you not eat shellfish?	Yes, I had to pass a swim test to be a lifeguard.
Person 2:	**I get sick and have a severe allergic reaction.**	**It must be stressful being responsible for the safety of people.**
Knowledge of Person 2:	I am a handyman. **I am allergic to shellfish.** I work for a company that rents properties. I like to go hunting.	**A lifeguard is responsible for the safety of people in an area of water, and usually a defined area immediately surrounding or adjacent to it, such as a beach next to an ocean or lake.** The effects of stress on memory include interference with ... During times of stress, the body reacts by secreting stress hormones...
	Example of Mode 3 in Persona-Chat	**Example of Mode 3 in Wiki-Wizard**
Person 1:	Hello how is your evening?	
Person 2:	Sehr gut. bene. Yes I am able to speak german, italian, some french.	
Person 1:	Awesome. I speak english and spanish.	
Person 2:	Great! Lets agree to speak english? I am a plumber.	
Person 1:	Cool. I am in high school but no school till we get back home.	
Person 2:	What grade are you in now? Do you like school?	My Mom is in telecommunications. Technology such as radio, telephone.
Person 1:	Started 10th, school is ok. Where you live.	What does she do?
Person 2:	I live in pittsburgh in the country and belong to th club.	Specifically fiber-optic communication! They transmit information by sending pulses of light through an optical fiber.
Person 1:	I like science and want to be an orthodontist ever since I got my braces.	Is that basically the internet?
Person 2:	Oh, braces. You must have beautiful teeth. my son needs braces.	Optical fiber is used by telecommunications companies! So yes, internet but also telephone signals and cable television signals!
Person 1:	Getting there. They hurt though.	That's cool! What else can you tell me?
Person 2:	**I am sorry. Are you a boy scout? My son is.**	**Well, early means of communication were signal flags, beacons, telegraphs, snd heliographs!**
Knowledge of Person 2:	**I volunteer in my sons boy scout troop.** I can speak four languages fluently. I can do many celebrity impressions. I am an award winning th member. I work as a plumber.	**Early means of communicating over a distance included visual signals, such as beacons, smoke signals, semaphore telegraphs, signal flags, and optical heliographs.** Telecommunication is the transmission of signs, signals, messages... Telecommunication occurs when the exchange of information between...