Content Selection Network for Document-grounded Retrieval-based Chatbots

Yutao Zhu[1]*, Jian-Yun Nie[1], Kun Zhou[2], Pan Du[1], and Zhicheng Dou[3]

[1] Université de Montréal, Québec, Canada
[2] School of Information, Renmin University of China
[3] Gaoling School of Artificial Intelligence, Renmin University of China
yutao.zhu@umontreal.ca, {nie,pandu}@iro.umontreal.ca,
francis_kun_zhou@163.com, dou@ruc.edu.cn

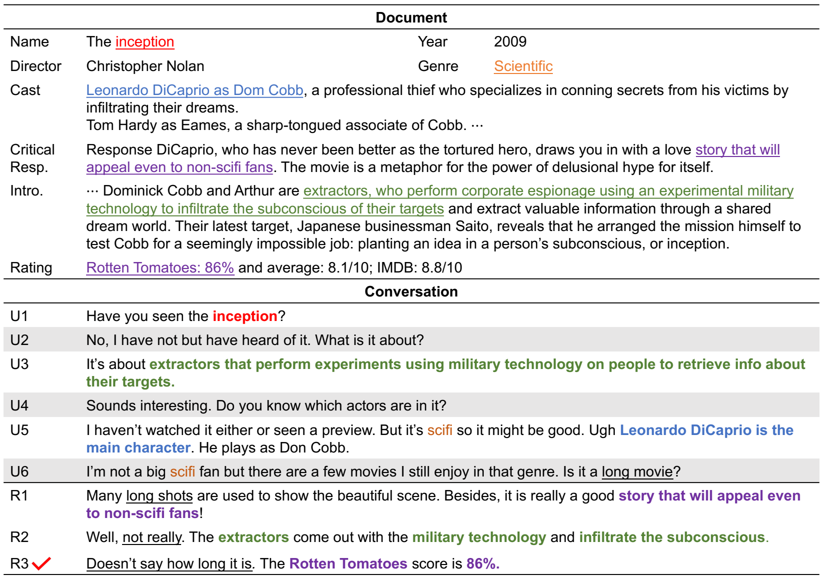
Abstract. Grounding human-machine conversation in a document is an effective way to improve the performance of retrieval-based chatbots. However, only a part of the document content may be relevant to help select the appropriate response at a round. It is thus crucial to select the part of document content relevant to the current conversation context. In this paper, we propose a document content selection network (CSN) to perform explicit selection of relevant document contents, and filter out the irrelevant parts. We show in experiments on two public document-grounded conversation datasets that CSN can effectively help select the relevant document contents to the conversation context, and it produces better results than the state-of-the-art approaches. Our code and datasets are available at `https://github.com/DaoD/CSN`.

Keywords: Content Selection · Document-grounded Dialogue · Retrieval-based Chatbots.

1 Introduction

Retrieval-based chatbots such as Microsoft XiaoIce [19] and Amazon Alexa [16] are widely used in real-world applications. Given a user input, an upstream retrieval system can provide a set of response candidates, and the retrieval-based chatbot should choose the appropriate one. This mainly relies on a matching score between the context and each candidate response. It has been found that the conversation context alone is insufficient in many cases for response selection [22,28]. In fact, human conversations are usually also grounded in external knowledge or documents: our responses are strongly related to our knowledge or information contained in the documents at hand. On Reddit, for example, people usually discuss about a document posted at the beginning of a thread, which provides the background topics and basic facts for the following conversations. On Twitter, people may also exchange opinions related to a news article. In these cases, in addition to the conversation context, the document or news article also

* Corresponding author.



Document			
Name	The inception	Year	2009
Director	Christopher Nolan	Genre	Scientific
Cast	Leonardo DiCaprio as Dom Cobb, a professional thief who specializes in conning secrets from his victims by infiltrating their dreams. Tom Hardy as Eames, a sharp-tongued associate of Cobb. ···		
Critical Resp.	Response DiCaprio, who has never been better as the tortured hero, draws you in with a love story that will appeal even to non-scifi fans. The movie is a metaphor for the power of delusional hype for itself.		
Intro.	··· Dominick Cobb and Arthur are extractors, who perform corporate espionage using an experimental military technology to infiltrate the subconscious of their targets and extract valuable information through a shared dream world. Their latest target, Japanese businessman Saito, reveals that he arranged the mission himself to test Cobb for a seemingly impossible job: planting an idea in a person's subconscious, or inception.		
Rating	Rotten Tomatoes: 86% and average: 8.1/10; IMDB: 8.8/10		
Conversation			
U1	Have you seen the **inception**?		
U2	No, I have not but have heard of it. What is it about?		
U3	It's about **extractors that perform experiments using military technology on people to retrieve info about their targets.**		
U4	Sounds interesting. Do you know which actors are in it?		
U5	I haven't watched it either or seen a preview. But it's scifi so it might be good. Ugh **Leonardo DiCaprio is the main character**. He plays as Don Cobb.		
U6	I'm not a big scifi fan but there are a few movies I still enjoy in that genre. Is it a long movie?		
R1	Many long shots are used to show the beautiful scene. Besides, it is really a good **story that will appeal even to non-scifi fans**!		
R2	Well, not really. The **extractors** come out with the **military technology** and **infiltrate the subconscious**.		
R3 ✓	Doesn't say how long it is. The **Rotten Tomatoes** score is **86%.**		

Fig. 1: An example in CMUDoG dataset. The words in color correspond to those in the document. R3 is the ground-truth response.

provides useful background information to guide response selection. A conversation that does not take into account the background information may lead to off-topic responses. This paper deals with the problem of document-grounded conversation - conversation based on a given document [28,1,15,29,4].

The task of document-grounded response selection is formulated as selecting a good response from a candidate pool that is consistent with the context and relevant to the document. Several existing studies have shown that leveraging the background document can significantly improve response selection [28,29,4]. Generally, the common strategy is selecting the response based on a combination of context-response matching and document-response matching. The latter can boost the responses that are related to the document content. However, a good response does not need to be related to the whole content of the document, but to a small part of it. The selection of the relevant part of the document is crucial.

The problem can be illustrated by an example from CMUDoG [30] in Fig. 1. In this dataset, a movie-related wiki article is used as the grounding document. We can see that the conversation is highly related to the document. R1, R2, and R3 are three candidate responses for U6, and R3 is the desired response. The wrong response R1 could be highly scored because it shares several key words with the document (*i.e.*, document-response matching score is high). However, R1 is not an appropriate response in the current context, which asks about the length of the movie. This example shows that a correct response is well grounded in the document not because it corresponds to the document content, but because it corresponds to the part relevant to the conversation context. Therefore, a first challenge is to select the part of the document content relevant to the current conversation context. R2 looks like a proper response to U6, yet it conveys similar information as U3, which makes the dialogue less informative. This response could be selected if we use the whole conversation history as conversation context - the response could have a high context-response matching score. In fact, the current context in this example is about the length of the movie. The previous utterances in the history are less relevant. This case illustrates the need to well calibrate and model the current conversation context.

The two key problems illustrated by the above example (R1 and R2) are not well addressed in previous studies: (1) They usually perform a soft selection of document content by assigning attention weights to them [29,4]. Even though the less relevant parts could be assigned lower weights, the cumulative weight of many irrelevant parts could be large, so that they collectively influence the response selection in a wrong direction. We believe that a key missing element is a proper (hard) selection of the document content that fits the current conversation context, instead of a (soft) weighting. The hard selection of document content is motivated by the following observation: although the whole conversation can cover many aspects described in the grounding document, each of the step is related to only a small part of the document content. For example, in our conversation about a movie, we could discuss about an actor in one step. The selection of such a small part of the content is crucial. This observation advocates a hard selection rather than a soft weighting used in the previous studies. (2) The existing studies usually use the entire context to determine the weights of parts (sentences) of the document content. This strategy fails to distinguish the current conversation context from the ones in the history. As a result, a past round of conversation could be mistaken as the current one, leading to a redundant response as illustrated by the R2 example.

In this paper, we propose a **Content Selection Network** (CSN) to tackle these problems. **First**, we use a modified *gate mechanism* to implement the document content selection according to the conversation context, before using it to match with the response candidate. The content relevant to the current conversation step will be assigned a higher weight and pass the selection gate, while the irrelevant parts will be blocked. We use the gate mechanism to select sentences or words. **Second**, as the topic usually evolves during the conversation, we determine the current conversation context by focusing on the most recent utterances, rather than on the whole conversation history. To this end, we design a decay mechanism for the history to force the model focusing more on the current dialogue topic. The selected document contents and the conversation context are finally combined to select the candidate response.

The main contributions of this paper are: (1) We propose a content selection network to explicitly select the relevant sentences/words from the document to complement the conversation context. Our experiments show that this is a much more effective way to leverage the grounding document than a soft weighting. (2) We show that document-grounded conversation should focus on the topics in the recent state rather than using the whole conversation context. On two public datasets for document-grounded conversation, our method outperforms the existing state-of-the-art approaches significantly.

2 Related Work

Retrieval-based Chatbots Existing methods for open-domain dialogue can be categorized into two groups: retrieval-based and generation-based. Generation-based methods are mainly based on the sequence-to-sequence (Seq2seq) architecture with attention mechanism and aim at generating a new response for conversation context [18,17,8,26,2]. On the other hand, retrieval-based methods try to find the most reasonable response from a large repository of conversational data [10,25,21,27]. We focus on retrieval-based methods in this paper. Early studies use single-turn response selection where the context is a single message [7,6], while recent work considers all previous utterances as context for multi-turn response selection [25,31,21,27]. In our work, we also consider the whole conversation history (but with decaying weights).
Document-grounded Conversation Multiple studies have shown that being grounded in knowledge or document can effectively enhance human-machine conversation [3,11,28,29]. For example, a Seq2seq model is first applied to generate responses based on both conversation history and external knowledge [3]. An approach using a dually interactive matching network has been proposed, in which context-response matching and document-response matching are performed separately using a shared structure [4]. This model achieved state-of-the-art performance on persona-related conversation [28]. Recently, Zhao et al. [29] proposed a document-grounded matching network that lets the document and the context to attend to each other so as to generate better representations for response selection. Through the attention mechanism, different parts (sentences) of the document are assigned different weights and will participate in response selection to different extents. However, even though one may expect the noise contents (for the current step) be assigned with lower weights, they can still participate in response selection.

Our work differs from the existing studies in that we explicitly model the document content selection process and prevent the irrelevant contents from participating in response selection. In addition, we also define the current conversation context by focusing more on recent utterances in the history rather than taking the whole history indistinctly. These ideas will bring significant improvements compared to the existing methods.

3 Content Selection Network

3.1 Problem Formalization

Suppose that we have a dataset $\mathcal{D}$, in which each sample is represented as (c, d, r, y), where $c = \{u_1, \ldots, u_n\}$ represents a conversation context with $\{u_i\}_{i=1}^{n}$ as utterances; $d = \{s_1, \ldots, s_m\}$ represents a document with $\{s_i\}_{i=1}^{m}$ as sentences;

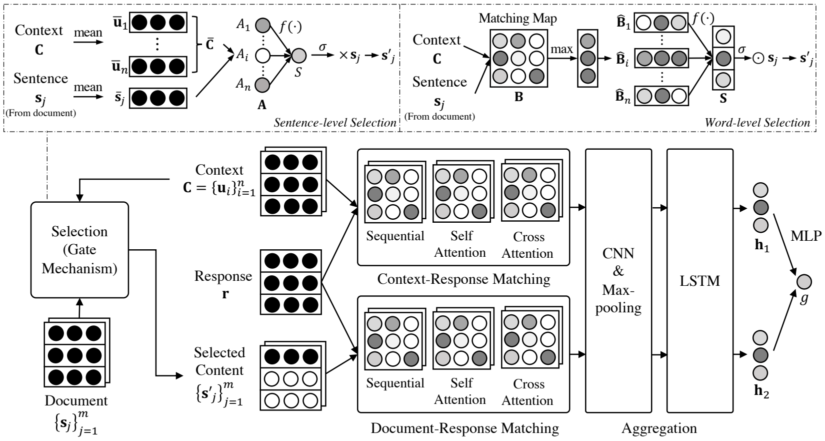


Fig. 2: The structure of CSN.

r is a response candidate; $y \in \{0,1\}$ is a binary label, indicating whether r is a proper response. Our goal is to learn a matching model g from $\mathcal{D}$, such that for a new context-document-response triplet (c,d,r), $g(c,d,r)$ measures the degree of suitability of a response r to the given context c and the document d.

3.2 Model Overview

We propose a content selection network (CSN) to model $g(\cdot,\cdot,\cdot)$, which is shown in Fig. 2. Different from the previous work [25,21,4] which uses the whole document contents, we propose a selection module with a gate mechanism to select the relevant parts of document content based on the context. Then, the context-response matching and the document-response matching are modeled based on the sequential, self-attention, and cross-attention representations. Finally, CNNs and RNNs are applied to extract, distill, and aggregate the matching features, based on which the response matching score is calculated.

3.3 Representation

Consider the i-th utterance $u_i = (w_1^u, \cdots, w_L^u)$ in the context, the j-th sentence $s_j = (w_1^s, \cdots, w_L^s)$ in the document, and the response $r = (w_1^r, \cdots, w_L^r)$, where L is the number of words[4]. CSN first uses a pre-trained embedding table to map each word w to a d_e-dimension embedding $\mathbf{e}$, *i.e.*, $w \Rightarrow \mathbf{e}$. Thus the utterance u_i, the sentence s_j, and the response r are represented by matrices $\mathbf{E}^{u_i} = (\mathbf{e}_1^{u_i}, \cdots, \mathbf{e}_L^{u_i})$, $\mathbf{E}^{s_j} = (\mathbf{e}_1^{s_j}, \cdots, \mathbf{e}_L^{s_j})$, and $\mathbf{E}^r = (\mathbf{e}_1^r, \cdots, \mathbf{e}_L^r)$, respectively. Then, CSN encodes the utterances, sentences and responses by bi-directional long short-term memories (BiLSTM) [5] to obtain their sequential representations: $\mathbf{u}_i = \mathrm{BiLSTM}(\mathbf{E}^{u_i})$, $\mathbf{s}_j = \mathrm{BiLSTM}(\mathbf{E}^{s_j})$, $\mathbf{r} = \mathrm{BiLSTM}(\mathbf{E}^r)$. Note that the

[4] To simplify the notation, we assume their lengths are the same.

parameters of these BiLSTMs are shared in our implementation. The whole context is thus represented as $\mathbf{C} = [\mathbf{u}_1, \cdots, \mathbf{u}_n]$. With the BiLSTM, the sequential relationship and dependency among words in both directions are expected to be encoded into hidden vectors.

3.4 Content Selection

In document-grounded conversation, the document usually contains a large amount of diverse information, but only a part of it is related to the current step of the conversation. To select the relevant part of document contents, we propose a content selection phase by a *gate mechanism*, which is based on the relevance between the document and the context. We design the gate mechanism at two different levels, *i.e.*, sentence-level and word-level, to capture relevant information at different granularities. If the sentences/words in the document are irrelevant to the current conversation, they will be filtered out. This is an important difference from the traditional gating mechanism, in which elements are assigned different attention weights, but no element is filtered out. We use the conversation context to control the gate, which contains several previous turns of conversation. Along the turns, the conversation topic gradually changes. The most important topic is that of the most recent turn, while more distant turns are less important. To reflect this fact, we design a decay mechanism on the history to assign a higher importance to the recent context than to the more distant ones. The selection process is automatically trained with the whole model in an end-to-end manner.

Sentence-level Selection Let us first explain how document sentences are selected according to conversation context. Considering the context $c = (u_1, \cdots, u_n)$ and the j-th sentence s_j in the document, CSN computes a score for the sentence s_j by measuring its matching degree with the current dialogue context. In particular, CSN first obtains the sentence representations of the context c and the sentence s_j by mean-pooling over the word dimension of their sequential representations:

$$\bar{\mathbf{C}} = \underset{dim=2}{\text{mean}}(\mathbf{C}), \quad \bar{\mathbf{s}}_j = \underset{dim=1}{\text{mean}}(\mathbf{s}_j), \tag{1}$$

where $\bar{\mathbf{C}} \in \mathbb{R}^{n \times 2d}$ and $\bar{\mathbf{s}}_j \in \mathbb{R}^{2d}$. Then CSN computes a sentence-level matching vector $\mathbf{A}$ by cosine similarities:

$$\mathbf{A} = \cos(\bar{\mathbf{C}}, \bar{\mathbf{s}}_j). \tag{2}$$

We can treat $\mathbf{A} \in \mathbb{R}^n$ as a similarity array $\mathbf{A} = [A_1, \cdots, A_n]$ and compute a matching score S for the sentence s_j by fusing the similarity scores:

$$S = f(A_1, A_2, \cdots, A_n). \tag{3}$$

The fusion function $f(\cdot)$ can be designed in different ways, which will be discussed later. After obtaining the matching scores for sentences, we select the relevant sentences and update their representations as follows:

$$S' = S \times (\sigma(S) \geq \gamma), \quad \mathbf{s}'_j = S' \times \mathbf{s}_j, \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid function and γ is a hyperparameter of the gate threshold. By this means, we will filter out a sentence s_j if its relevance score is below γ. The filtering is intended to remove the impact of clearly irrelevant parts of document content.

Word-level Selection In the sentence-level selection, all words in a sentence are assigned the same weights. We can further perform a selection of words by computing a score for each word in the sentence. Specifically, CSN constructs a word-level matching map through the attention mechanism as follows:

$$\mathbf{B} = \mathbf{v}^\top \tanh(\mathbf{s}_j^\top \mathbf{W}_1 \mathbf{C} + \mathbf{b}_1), \tag{5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{2d \times 2d \times h}$, $\mathbf{b}_1 \in \mathbb{R}^h$ and $\mathbf{v} \in \mathbb{R}^{h \times 1}$ are parameters. $\mathbf{B} \in \mathbb{R}^{n \times L \times L}$ is the word-alignment matrix between the context and the document sentence. Then, to obtain the most important matching features between s_j and each utterance in the context, CSN conducts a max-pooling operation as follows:

$$\hat{\mathbf{B}} = \max_{dim=3} \mathbf{B}, \tag{6}$$

where $\hat{\mathbf{B}} \in \mathbb{R}^{n \times L}$, and it can be represented in an array form as $\hat{\mathbf{B}} = [\hat{\mathbf{B}}_1, \cdots, \hat{\mathbf{B}}_n]$. The element $\hat{\mathbf{B}}_i \in \mathbb{R}^L$ contains L local matching signals for all words in the document sentence s_j with respect to the utterance u_i. Thereafter, CSN applies a fusion function to combine these local matching signals and obtains a global matching vector:

$$\mathbf{S} = f(\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \cdots, \hat{\mathbf{B}}_n). \tag{7}$$

$\mathbf{S} \in \mathbb{R}^L$ thus contains L global matching scores for all words in $\mathbf{s}_j$ to the whole context. In the next step, CSN selects the relevant words in the document and updates the document representation as follows:

$$\mathbf{S}' = \mathbf{S} \odot (\sigma(\mathbf{S}) \geq \gamma), \quad \mathbf{s}_j' = \mathbf{S}' \odot \mathbf{s}_j, \tag{8}$$

where $\odot$ is the element-wise product. Different from the sentence-level matching score S' in Equation 4, the word-level matching score $\mathbf{S}'$ is a vector containing weights for different words.

Fusion Function The fusion function $f(\cdot)$ in Equation (3) and (7) is used to aggregate the matching signals with each utterance in the context. Our fusion strategies attribute different weights to the utterances in the conversation history. Two different functions are considered: (1) Linear combination – the weight of each matching signal is learned during the model training. Ideally, an utterance containing more information about the conversation topic will contribute more to the selection of document content. (2) Linear combination with decay factors. This method assumes that the topic gradually changes along the conversation and the response is usually highly related to the most recent topic in the context. Therefore, we use a decay factor $\eta \in [0, 1]$ on the utterances in the context to decrease their importance when they are far away. The matching scores are then computed as:

$$A_i = A_i * \eta^{n-i}, \quad \text{(sentence-level)} \qquad \hat{\mathbf{B}}_i = \hat{\mathbf{B}}_i * \eta^{n-i}. \quad \text{(word-level)} \tag{9}$$

The decay factor η is a hyperparameter. Note that when $\eta = 1$, it degenerates to the normal linear combination.

3.5 Matching and Aggregation

The next problem is to select the appropriate response by leveraging the selected document parts. Following a recent study [4], CSN uses a dually interactive matching structure (as shown in Fig. 2) to determine context-response matching and document-response matching, where the two kinds of matching features are modeled by the same structure.

Based on the recent work [25,31,27] that constructs different matching feature maps, in addition to using the sequential representations of the sentences, CSN also uses matching on both self-attention and cross-attention representations. Given the sequential representations of the context $\mathbf{C} = [\mathbf{u}_1, \cdots, \mathbf{u}_n]$, the document $\mathbf{D} = [\mathbf{s}'_1, \cdots, \mathbf{s}'_m]$, and the response candidate $\mathbf{r}$, CSN first constructs a word-word similarity matrix $\mathbf{M}_1$ by dot product and cosine similarity:

$$\mathbf{M}_1^{cr} = \mathbf{C}\mathbf{H}_1\mathbf{r}^\top \oplus \cos(\mathbf{C}, \mathbf{r}), \qquad \mathbf{M}_1^{dr} = \mathbf{D}\mathbf{H}_1\mathbf{r}^\top \oplus \cos(\mathbf{D}, \mathbf{r}), \tag{10}$$

where $\mathbf{H}_1 \in \mathbb{R}^{2d \times 2d}$ is a parameter, and $\oplus$ is the concatenation operation.

To better handle the gap in words between two word sequences, CSN applies the attentive module, which is similar to that used in Transformer [23]. The input of an attentive module consists of three sequences, namely query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$). The output is a new representation of the query and is denoted as $f_{\text{ATT}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ in the remaining description.

At first, CSN uses the attentive module over the word dimension to construct multi-grained representations, which is formulated as:

$$\hat{\mathbf{C}} = f_{\text{ATT}}(\mathbf{C}, \mathbf{C}, \mathbf{C}), \qquad \hat{\mathbf{D}} = f_{\text{ATT}}(\mathbf{D}, \mathbf{D}, \mathbf{D}), \qquad \hat{\mathbf{r}} = f_{\text{ATT}}(\mathbf{r}, \mathbf{r}, \mathbf{r}). \tag{11}$$

The second similarity matrix $\mathbf{M}_2$ is computed based on these self-attention representations:

$$\mathbf{M}_2^{cr} = \hat{\mathbf{C}}\mathbf{H}_2\hat{\mathbf{r}}^\top \oplus \cos(\hat{\mathbf{C}}, \hat{\mathbf{r}}), \qquad \mathbf{M}_2^{dr} = \hat{\mathbf{D}}\mathbf{H}_2\hat{\mathbf{r}}^\top \oplus \cos(\hat{\mathbf{D}}, \hat{\mathbf{r}}). \tag{12}$$

Then, another group of attentive modules (cross-attention) is also applied to represent semantic dependency between the context, the document, and the response candidate:

$$\tilde{\mathbf{C}} = f_{\text{ATT}}(\mathbf{C}, \mathbf{r}, \mathbf{r}), \qquad \tilde{\mathbf{r}}^c = f_{\text{ATT}}(\mathbf{r}, \mathbf{C}, \mathbf{C}), \tag{13}$$

$$\tilde{\mathbf{D}} = f_{\text{ATT}}(\mathbf{D}, \mathbf{r}, \mathbf{r}), \qquad \tilde{\mathbf{r}}^d = f_{\text{ATT}}(\mathbf{r}, \mathbf{D}, \mathbf{D}). \tag{14}$$

Next, CSN also constructs a similarity matrix $\mathbf{M}_3$ as:

$$\mathbf{M}_3^{cr} = \tilde{\mathbf{C}}\mathbf{H}_3\tilde{\mathbf{r}}^{c\top} \oplus \cos(\tilde{\mathbf{C}}, \tilde{\mathbf{r}}^c), \qquad \mathbf{M}_3^{dr} = \tilde{\mathbf{D}}\mathbf{H}_3\tilde{\mathbf{r}}^{d\top} \oplus \cos(\tilde{\mathbf{D}}, \tilde{\mathbf{r}}^d). \tag{15}$$

The above matching matrices are concatenated into two matching cubes:

$$\mathbf{M}^{cr} = \mathbf{M}_1^{cr} \oplus \mathbf{M}_2^{cr} \oplus \mathbf{M}_3^{cr}, \qquad \mathbf{M}^{dr} = \mathbf{M}_1^{dr} \oplus \mathbf{M}_2^{dr} \oplus \mathbf{M}_3^{dr}. \tag{16}$$

Then CSN applies a CNN with max-pooling operation to extract matching features from $\mathbf{M}^{cr}$ and $\mathbf{M}^{dr}$. The output feature maps are flattened as matching vectors. As a result, we obtain two series of matching vectors: (1) between the context and the response $\mathbf{v}^{cr} = [\mathbf{v}^{u_1}, \cdots, \mathbf{v}^{u_n}]$; and (2) between the selected document and the response $\mathbf{v}^{dr} = [\mathbf{v}^{s_1}, \cdots, \mathbf{v}^{s_m}]$.

Finally, CSN applies LSTMs to aggregate these two series of matching vectors into two hidden vectors (the last hidden states of the LSTMs):

$$\mathbf{h}_1 = \mathrm{LSTM}(\mathbf{v}^{cr}), \quad \mathbf{h}_2 = \mathrm{LSTM}(\mathbf{v}^{dr}). \tag{17}$$

These vectors are concatenated together and used to compute the final matching score by an MLP with a Sigmoid activation function:

$$g(c, d, r) = \sigma\big(\mathrm{MLP}(\mathbf{h}_1 \oplus \mathbf{h}_2)\big). \tag{18}$$

CSN learns $g(c, d, r)$ by minimizing the following cross-entropy loss with $\mathcal{D}$:

$$\mathcal{L}(\theta) = -\sum_{(y,c,d,r)\in\mathcal{D}} [y \log(g(c, d, r)) + (1 - y) \log(1 - g(c, d, r))]. \tag{19}$$

4 Experiments

4.1 Dataset

We conduct experiments on two public datasets.

PersonaChat [28] contains multi-turn dialogues with user profiles. The goal is to generate/retrieve a response that corresponds to the user profile, which is used as a grounding document [28]. This dataset consists of 8,939 complete dialogues for training, 1,000 for validation, and 968 for testing. Response selection is conducted at every turn of a dialogue, and the ratio of the positive and the negative samples is 1:19 in training, validation, and testing sets, resulting in 1,314,380 samples for training, 156,020 for validation, and 150,240 for testing. Positive responses are real human responses while negative ones are randomly sampled from other dialogues. To prevent the model from taking advantage of trivial word overlap, the revised version of the dataset modified the persona profiles by rephrasing, generalizing, or specializing sentences, making the task much more challenging. We use "revised" and "original" to indicate the different versions of the dataset.

CMUDoG [30] is designed specifically for document-grounded conversation. During the conversation, the speakers are provided with a movie-related wiki article. Two scenarios are considered: (1) Only one speaker has access to the article thus she should introduce the movie to the other; (2) Both speakers have access to the article thus they have a discussion. We use the dataset provided by [29], where the data of both scenarios are merged because the size of each dataset is relatively small. Notice that the model is only asked to select a response for the user who has access to the document. The ratio of the positive and the

Table 1: Experimental results on all datasets.

	PersonaChat-Original			PersonaChat-Revised			CMUDoG		
	R@1	**R@2**	**R@5**	**R@1**	**R@2**	**R@5**	**R@1**	**R@2**	**R@5**
Starspace	49.1	60.2	76.5	32.2	48.3	66.7	50.7	64.5	80.3
Profile	50.9	60.7	75.7	35.4	48.3	67.5	51.6	65.8	81.4
KV Profile	51.1	61.8	77.4	35.1	45.7	66.3	56.1	69.9	82.4
Transformer	54.2	68.3	83.8	42.1	56.5	75.0	60.3	74.4	87.4
DGMN	67.6	81.3	93.3	56.7	73.0	89.0	65.6	78.3	91.2
DIM	75.5	87.5	96.5	68.3	82.7	94.4	59.6	74.4	89.6
CSN-sent	77.5	88.8	96.8	70.1	83.4	95.1	**70.1**	82.5	**94.3**
CSN-word	**78.1**	**89.0**	**97.1**	**71.3**	**84.2**	**95.5**	69.8	**82.7**	94.0

negative is 1:19 in training, validation, and testing sets. This results in 723,180 samples for training, 48,500 for validation, and 132,740 for testing.

Following previous work [29], we employ recall at position k as evaluation metrics (R@k), where $k = \{1, 2, 5\}$. For a single sample, if the only positive candidate is ranked within top k positions, then R@$k = 1$, otherwise, R@$k = 0$. The final value is the average over all test samples. Note that R@1 is equivalent to hits@1 that is used in related work [28,4].

4.2 Baseline Models

We compare CSN using sentence-level and word-level selection (denoted as CSN-sent and CSN-word respectively) with the following models:

(1) Starspace [24] concatenates the document with the context as a long sentence and learns its similarity with the response candidate by optimizing the embeddings using the margin ranking loss and k-negative sampling. Matching is done by cosine similarity of the sum of word embeddings.

(2) Profile Memory Network [28] uses a memory network with the context as input, then performs attention over the document to find relevant sentences. The combined representation is used to select the response. This model relies on the attention mechanism to weigh document contents.

(3) Key-value (KV) Profile Memory Network [28] uses dialogue histories as keys and the next dialogue utterances as values. In addition to the memory of the document, this model has a memory of past dialogues that can influence the response selection.

(4) Transformer [23] is used in [11] as an encoder for the context, document, and response. The obtained representations are input to a memory network to conduct matching in the same way as in Profile Memory Network.

(5) DGMN [29] is the state-of-the-art model on the CMUDoG dataset. It employs a cross attention mechanism between the context and document and obtains a context-aware document representation and a document-aware context representation. The two representations and the original context representation

are all matched with the response representation. The three matching features are finally combined to output the matching score.

(6) DIM [4] is the state-of-the-art model on the PersonaChat dataset. It applies a dually interactive matching structure to model the context-response matching and document-response matching respectively. DIM conducts representation, matching, and aggregation by multiple BiLSTMs, and the final matching features are used to compute the matching score by an MLP.

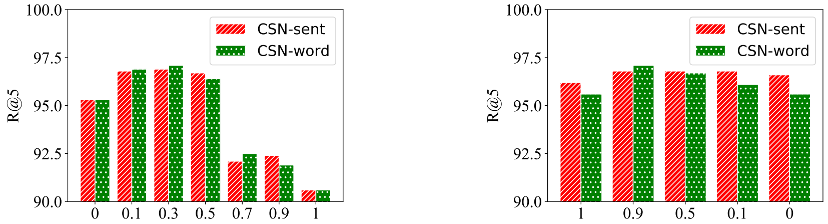
4.3 Implementation Details

We use PyTorch [13] to implement the model. A 300-dimensional GloVe embedding [14] is used on all datasets. On PersonaChat, another 100-dimensional Word2Vec [12] embedding provided by [4] is used. Dropout [20] with a rate of 0.2 is applied to the word embeddings. All hidden sizes of the RNNs are set as 300. Two convolutional layers have 32 and 64 filters with the kernel sizes as [3, 3] and [2, 2]. AdamW [9] is employed for optimization with a batch size of 100. The initial learning rate is 0.001 and is decayed by 0.5 when the performance on the validation set is not increasing.

4.4 Experimental Results

The experimental results are shown in Table 1. The results on all three datasets indicate that our CSN outperforms all baselines, including DGMN and DIM, which are two state-of-the-art models. On the PersonaChat dataset, both CSN-word and CSN-sent achieve statistically significant improvements (p-value ≤ 0.05) compared with DIM, which is the best model on this dataset. In general, CSN-word performs better than CSN-sent, indicating the word-level selection is more able to select fine-grained document contents than the sentence-level selection. This comparison also confirms our intuition that it is advantageous for document-grounded conversation to rely on fine-grained information from the document. On CMUDoG, the two document content selection strategies work equally well. We explain this by the fact that the grounding document is longer in this dataset, and there is no obvious reason that one level of selection can determine more relevant parts than another. Nevertheless, both selection strategies show clear advantages over the baseline methods without selection.

Compared with other baselines that represent the whole document as a single vector, DGMN, DIM and our CSN consider fine-grained matching between parts of the document and response. We can see that these models achieve clearly better performances, confirming the necessity to use parts of the document rather than the whole document. However, DGMN and DIM only assign attention weights to sentences according to the context, without eliminating low-weighted ones. In contrast, our CSN model filters out all the irrelevant parts. In so doing, we expect the model not to be influenced by clearly irrelevant parts. As we can see in the experimental results, CSN achieves significantly higher performance than DGMN and DIM on all the datasets, confirming the usefulness of explicit selection (and filtering) of document contents.



(a) Effect of γ.

(b) Effect of η.

Fig. 3: Performance of different γ and η settings on original PersonaChat.

Effect of Content Selection The hyperparameter γ in Equation (4) and (8) controls how much the document content is selected. We test the effect of this hyperparameter on the original PersonaChat dataset. Fig. 3a shows that if γ is too small or too large, too much or too little information from the document may be selected. In particular, when $\gamma = 0$ – the whole document content is kept, the performance drops a lot. This strategy is comparable to that used in the existing models DIM and DGMN based on attention. We see again the usefulness of explicit document content filtering. On the other hand, when $\gamma = 1$, *i.e.*, no document content is selected, it degenerates to non document-grounded response selection and the performance also drops sharply. The best setting of γ is around 0.3 for both CSN-sent and CSN-word, which retains an appropriate amount of relevant document content for response matching.

Effect of Decaying Factor The decay factor η works as prior knowledge to guide the model focusing more on the recent utterances. A lower η means the previous utterances have less contribution in the selection of the document. "$\eta = 1$" corresponds to the model with a normal linear combination (the first kind of fusion function). Based on the results, we can see that our decaying strategy ($\eta = 0.9$) performs the best. This confirms our assumption that focusing more on the recent topic of the conversation is helpful. However, when $\eta = 0$, only the last utterance in the history is used and the performance is lower. This illustrates the necessity of using a larger context.

5 Conclusion and Future Work

In this paper, we proposed a document content selection network to select the relevant content to ground the conversation. We designed a gate mechanism that uses conversation context to retain the relevant document contents while filtering out irrelevant parts. In addition, we also use a decay factor on the conversation history to focus on more recent utterances. Our experiments on two large-scale datasets for document-grounded response selection demonstrated the effectiveness of our model. We showed that both document content selection (and filtering) and the use of decay factor contributed in increasing the effectiveness of response selection. As a future work, it would be interesting to study if the selection can be done at topic level, in addition to sentence and word levels.

References

1. Arora, S., Khapra, M.M., Ramaswamy, H.G.: On knowledge distillation from complex networks for response prediction. 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. 3813–3822. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
2. Cai, D., Wang, Y., Bi, W., Tu, Z., Liu, X., Shi, S.: Retrieval-guided dialogue response generation via a matching-to-generation framework. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1866–1875. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
3. Ghazvininejad, M., Brockett, C., Chang, M., Dolan, B., Gao, J., Yih, W., Galley, M.: A knowledge-grounded neural conversation model. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 5110–5117 (2018)
4. Gu, J.C., Ling, Z.H., Zhu, X., Liu, Q.: Dually interactive matching network for personalized response selection in retrieval-based chatbots. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1845–1854. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
6. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2042–2050 (2014)
7. Ji, Z., Lu, Z., Li, H.: An information retrieval approach to short text conversation. CoRR **abs/1408.6988** (2014)
8. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: 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. pp. 110–119. Association for Computational Linguistics, San Diego, California (Jun 2016)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019)
10. Lowe, R., Pow, N., Serban, I., Pineau, J.: The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 285–294. Association for Computational Linguistics, Prague, Czech Republic (Sep 2015)
11. Mazaré, P.E., Humeau, S., Raison, M., Bordes, A.: Training millions of personalized dialogue agents. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2775–2779. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. pp. 3111–3119 (2013)
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. pp. 8024–8035 (2019)
14. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014)
15. Qin, L., Galley, M., Brockett, C., Liu, X., Gao, X., Dolan, B., Choi, Y., Gao, J.: Conversing by reading: Contentful neural conversation with on-demand machine reading. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5427–5436. Association for Computational Linguistics, Florence, Italy (Jul 2019)
16. Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., Pettigrue, A.: Conversational AI: the science behind the alexa prize. CoRR **abs/1801.03604** (2018)
17. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. pp. 3776–3784 (2016)
18. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1577–1586. Association for Computational Linguistics, Beijing, China (Jul 2015)
19. Shum, H., He, X., Li, D.: From eliza to xiaoice: challenges and opportunities with social chatbots. Frontiers Inf. Technol. Electron. Eng. **19**(1), 10–26 (2018)
20. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
21. Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., Yan, R.: One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1–11. Association for Computational Linguistics, Florence, Italy (Jul 2019)
22. Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., Zhao, D.: How to make context more useful? an empirical study on context-aware neural conversational models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 231–236. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 5998–6008 (2017)
24. Wu, L.Y., Fisch, A., Chopra, S., Adams, K., Bordes, A., Weston, J.: Starspace: Embed all the things! In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 5569–5577 (2018)
25. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 496–505. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)
26. Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., Ma, W.: Topic aware neural response generation. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. pp. 3351–3357 (2017)
27. Yuan, C., Zhou, W., Li, M., Lv, S., Zhu, F., Han, J., Hu, S.: 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 International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 111–120. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
28. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: 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. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
29. Zhao, X., Tao, C., Wu, W., Xu, C., Zhao, D., Yan, R.: A document-grounded matching network for response selection in retrieval-based chatbots. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. pp. 5443–5449 (2019)
30. Zhou, K., Prabhumoye, S., Black, A.W.: A dataset for document grounded conversations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 708–713. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)
31. Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W.X., Yu, D., Wu, H.: 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). pp. 1118–1127. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)