# Towards Implicit Content-Introducing for Generative Short-Text Conversation Systems

**Lili Yao**[1], **Yaoyuan Zhang**[1], **Yansong Feng**[1], **Dongyan Zhao**[1,2] and **Rui Yan**[1,2] *
[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]Beijing Institute of Big Data Research, Beijing, China
{yaolili,zhang_yaoyuan,fengyansong,zhaody,ruiyan}@pku.edu.cn

## Abstract

The study on human-computer conversation systems is a hot research topic nowadays. One of the prevailing methods to build the system is using the generative Sequence-to-Sequence (Seq2Seq) model through neural networks. However, the standard Seq2Seq model is prone to generate trivial responses. In this paper, we aim to generate a more meaningful and informative reply when answering a given question. We propose an implicit content-introducing method which incorporates additional information into the Seq2Seq model in a flexible way. Specifically, we fuse the general decoding and the auxiliary cue word information through our proposed hierarchical gated fusion unit. Experiments on real-life data demonstrate that our model consistently outperforms a set of competitive baselines in terms of BLEU scores and human evaluation.

## 1 Introduction

To establish a conversation system with adequate artificial intelligence is a long-cherished goal for researchers and practitioners. In particular, automatic conversation systems in open domains are attracting increasing attention due to its wide applications, such as virtual assistants and chatbots. In open domains, researchers mainly focus on data-driven approaches, since the diversity and uncertainty make it impossible to prepare the interaction logic and domain knowledge. Basically, there are two mainstream ways to build an open-domain conversation system: 1) to search pre-established database for candidate responses by query retrieval (Isbell et al., 2000; Wang et al., 2013; Yan et al., 2016; Song et al., 2016), and 2) to generate a new, tailored utterance given the user-issued query (Shang et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Mou et al., 2016; Song et al., 2016). In these studies, generation-based conversation systems have shown impressive potential. Especially, the Sequence-to-Sequence (Seq2Seq) model (Sutskever et al., 2014) based on neural networks has been extensively used in practice; the idea is to encode a query as a vector and to decode the vector into a reply. Inspired by (Mou et al., 2016), we mainly focus on the generative short-text conversation without context information.

Despite this, the performance of Seq2Seq generation-based conversation systems is far from satisfactory because its generation process is not controllable; it responses to a query according to the pattern learned from the training corpus. As a result, the system is likely to generate an unexpected reply even with little semantics, e.g, "I don't know" and "Okay" due to the high frequency of these patterns in training data (Li et al., 2016a; Mou et al., 2016). To address this issue, Li et al. (2016a) proposed to increase diversity in the Seq2Seq model so that more informative utterances have a chance to stand out. Mou et al. (2016) provided a content-introducing approach that generates a reply based on a predicted word. The word is usually enlightening and drives the generated response to be more meaningful. However, this method is to some extent rigid; it requires the predicted word to explicitly occur in the generated utterance. As shown in Table 1, sometimes, it is better to generate a semantic related sentence based on the cue word rather than including it in the reply directly.

As for such content-introducing method, there are two aspects that need to be taken into consid-

*Corresponding author: ruiyan@pku.edu.cn



*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

| | |
|---|---|
| Query | 你不觉得好丑吗(Don't you think it is ugly?) |
| Cue Word | 审美(Aesthetics) |
| Reply | 好恶心啊! (It's disgusting!) |
| Query | 先放个大招(Let me use my ultimate power.) |
| Cue Word | 技能(Skill) |
| Reply | 新技能？ (New skill?) |

Table 1: The content-introducing conversation examples.

eration. 1) How to add the additional cue words during the generation process? One of the prevailing methods is modifying the neural cell with various gating mechanisms (Wen et al., 2015a,b; Xu et al., 2016). However, we need careful operation to ensure the neuron works as expected. 2) How to display the cue words in replies? As mentioned above, the explicit content-introducing approach in (Mou et al., 2016) does not fit well with all situations.

In this paper, we present an implicit content-introducing method for generative conversation systems, which incorporates cue words using our proposed hierarchical gated fusion unit (HGFU) in a flexible way. Our main contributions are as follows:

- We propose the cue word GRU, another neural cell, to deal with the auxiliary information. Compared with other gating methods, our cue word GRU is more flexible.
- We focus on the implicit content-introducing method during generation: the information of the cue word will be fused into the generation process but not necessarily occur explicitly. In this way, we change the "hard" content-introducing method into a new "soft" schema.

The rest of paper is organized as follows. We start by introducing the technical background. In Section 3, we describe our proposed method. In Section 4, we illustrate the experimental setup and evaluations against a variety of baselines. Section 5 briefly reviews related work. Finally, we conclude our paper in Section 6.

## 2 Technical Background

### 2.1 Seq2Seq Model and Attention Mechanism

Seq2Seq model was first introduced in statistical machine translation; the idea is to encode a source sentence as a vector by a recurrent neural network (RNN) and to decode the vector to a target sentence by another RNN. Now, the conversational generation is treated as a monolingual translation task (Ritter et al., 2011; Shang et al., 2015). Given a query $Q = (x_1, ..., x_n)$, the encoder represents it as a context vector $C$ and then the decoder generates a response $R = (y_1, ..., y_m)$ word by word by maximizing the generation probability of $R$ conditioned on $Q$. The objective function of Seq2Seq can be written as:

$$\begin{aligned} &p(y_1, ..., y_m | x_1, ..., x_n) \\ =&p(y_1|C) \prod_{t=2}^{T} p(y_t|C, y_1, ..., y_{t-1}) \end{aligned} \tag{1}$$

To be specific, the encoder RNN calculates the context vector by:

$$h_t = f(x_t, h_{t-1}); C = h_T \tag{2}$$

where $h_t$ is the hidden state of encoder RNN at time $t$ and $f$ is a non-linear transformation which can be a long-short term memory unit (LSTM) (Hochreiter and Schmidhuber, 1997) or a gated recurrent unit (GRU) (Cho et al., 2014). In this work, we implement $f$ using GRU.

The decoder RNN generates each reply word conditioned on the context vector $C$. The probability distribution $p_t$ of candidate words at every time step $t$ is calculated as:

$$s_t = f(y_{t-1}, s_{t-1}, C); p_t = \textit{softmax}(s_t, y_{t-1}) \tag{3}$$

where $s_t$ is the hidden state of decoder RNN at time $t$ and $y_{t-1}$ is the generated word in the reply at time $t-1$.

Attention mechanisms (Bahdanau et al., 2014) have been proved effective to improve the generation quality. In Seq2Seq with attention, each $y_i$ corresponds to a context vector $C_i$; it is weighted average of all hidden states of the encoder. Formally, $C_i$ is defined as $C_i = \sum_{j=1}^{T} \alpha_{ij} h_j$, where $\alpha_{ij}$ is given by:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T} exp(e_{ik})}; e_{ij} = \eta(s_{i-1}, h_j) \tag{4}$$

where $\eta$ is usually implemented as a multi-layer perceptron (MLP) with tanh as an activation function.

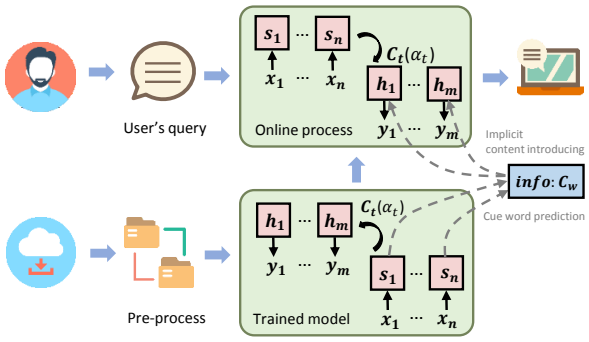


Figure 1: The architecture of our system. Based on the constructed corpus, we train our implicit content-introducing conversation system. Given a user-issued query, we first predict the cue word. Then, we incorporate the cue word into decoding process to generate a meaningful response.

### 2.2 Pointwise Mutual Information

Pointwise mutual information (PMI) (Church and Hanks, 1990) is a measure of association ratio based on the information theoretic concept of mutual information. Given a pair of outcomes $x$ and $y$ belonging to discrete random variables $\mathcal{X}$ and $\mathcal{Y}$, the PMI quantifies the discrepancy between the probability of their coincidence based on their joint distribution and their individual distributions. Mathematically:

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} \quad (5)$$

This quantity is zero if $x$ and $y$ are independent, positive if they are positively correlated, and negative if they are negatively correlated.

## 3 Implicit Content-Introducing Conversation System

Figure 1 provides an overview of our system architecture. We crawl conversational data from social media which are publicly available. After filtering and cleaning procedures, we establish the conversational parallel dataset, which consists of a large number of aligned $\langle query - reply \rangle$ pairs. Based on the entire set, we first predict the cue word for the given query in Subsection 3.1. Next, we propose the new implicit content-introducing process, which explores when to incorporate the predicted cue word in Subsection 3.2 and how to apply such information in Subsection 3.3.

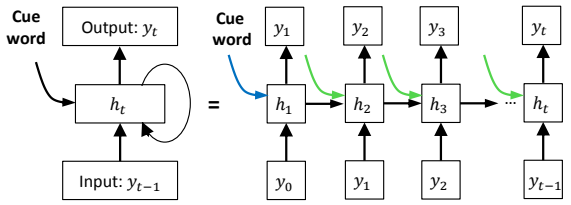


Figure 2: The information fusion patterns. The local information initialization is presented by the blue arrowhead and the global information inception includes both the blue arrowhead and the green arrowhead.

### 3.1 Cue Word Prediction

In computational linguistics, PMI has been used for finding collocations and associations between words. As mentioned in Mou et al. (2016), it is an appropriate statistic for cue words prediction, which is also adopted in this paper to predict a cue word $C_w$ for the given query. Formally, given a query word $w_q$ and a reply word $w_r$, the PMI is computed as:

$$\text{PMI}(w_q, w_r) = \log \frac{p(w_q|w_r)}{p(w_q)} \quad (6)$$

Then, we choose the cue word $C_w$ with highest PMI score against the query words $w_{q1}, ..., w_{qn}$ during the prediction, i.e., $C_w = \text{argmax}_{w_r} \text{PMI}(w_{q1}, ..., w_{qn}, w_r)$, where

$$\begin{aligned} \text{PMI}(w_{q1}, ..., w_{qn}, w_r) &\approx \log \frac{\prod_i p(w_{qi}|w_r)}{\prod_i p(w_{qi})} \\ = \sum_i \log \frac{p(w_{qi}|w_r)}{p(w_{qi})} &= \sum_i \text{PMI}(w_{qi}, w_r) \end{aligned} \quad (7)$$

The approximation is based on the independence assumptions of both the prior distribution $p(w_{qi})$ and posterior distributions $p(w_{qi}|w_r)$. Even the two assumptions may not be true, we use them in a pragmatic way so that the word-level PMI is additive for a whole utterance. PMI penalizes a common word by dividing its prior probability; hence, it prefers a word which is most "mutually informative" with the query.

### 3.2 Information Fusion Patterns

To implant the specific information in conversation system, we consider two types of information fusion patterns, namely 1) Local information initialization 2) Global information inception.

**Local information initialization.** In the local pattern, we fuse the cue word $C_w$ as the auxiliary

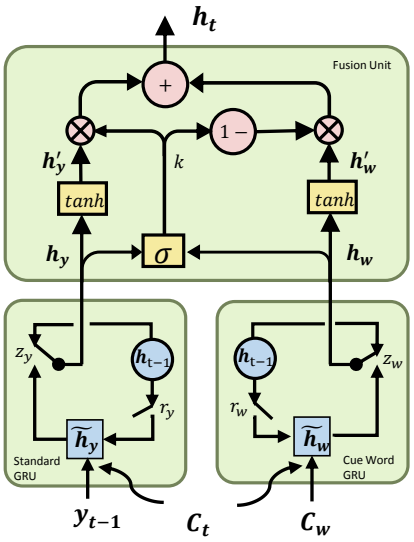


Figure 3: The structure of a HGFU. The bottom of two GRUs deal with corresponding input source, i.e., the last generated word $y_{t-1}$ and the cue word $C_w$. After that, fusion unit combines the output of two GRUs to compute current hidden state $h_t$.

.

information only in the beginning of decoding. We describe this kind of pattern by the blue arrowhead in Figure 2. Recurrent neural networks(RNNs) such as gated recurrent units (GRUs) have the ability to keep the information from the beginning to the end to some extent. Therefore, the cue word added on the first step of the neural networks can still influence the generation of the later steps.

**Global information inception.** However, we observe that, although the network is capable of deciding what to keep in the cell state to affect the later generation, the influence of the added information in the beginning of decoding is becoming weaker and weaker over time. Therefore, to provide the model a broader and more flexible space for learning, we propose a global information inception pattern, which fuses the cue word $C_w$ as the auxiliary information at every step of decoding. This process is presented by both the blue arrowhead and the green arrowheads in Figure 2.

### 3.3 Hierarchical Gated Fusion Unit

In this subsection, we propose our Hierarchical Gated Fusion Unit (HGFU), which incorporates cue words into the generation process and relaxes the constraint from the "hard" content-introducing method into a new "soft" schema. Figure 3 provides an overview of the structure of a HGFU. As seen, the framework consists of three components: the standard GRU, the cue word GRU, and the fusion unit. Among them, standard GRU and cue word GRU take the last generated word $y_{t-1}$ and cue word $C_w$ respectively as the decoder GRU's input; the fusion unit combines the hidden states of both GRUs to predict the next word $y_t$. In the following, we will illustrate these components in detail.

#### 3.3.1 Standard GRU

We adopt the standard gated recurrent unit (GRU) with the attention mechanism at the decoder part. Let $h_{t-1}$ be the last hidden state, $y_{t-1}$ be the embedding of the last generated word, and $C_t$ be the current attention-based context. The current hidden state of the general decoding, $h_y$, is defined as:

$$
\begin{aligned}
r_y &= \sigma(W_r y_{t-1} + U_r h_{t-1} + U_{cr} C_t + b_r) \\
z_y &= \sigma(W_z y_{t-1} + U_z h_{t-1} + U_{cz} C_t + b_z) \\
\widetilde{h_y} &= \tanh(W_h y_{t-1} + U_h(r_y \circ h_{t-1}) + U_{ch} C_t + b_h) \\
h_y &= (1 - z_y) \circ h_{t-1} + z_y \circ \widetilde{h_y}
\end{aligned}
\tag{8}
$$

where $W$'s $\in \mathbb{R}^{dim \times E}$ and $U$'s $\in \mathbb{R}^{dim \times dim}$ are weight matrices; $b$'s $\in \mathbb{R}^{dim}$ are bias terms; $E$ denotes the word embedding dimensionality and $dim$ denotes the number of hidden state units. This general decoding process is presented by the "Standard GRU" in Figure 3.

#### 3.3.2 Cue word GRU

To generate more meaningful and informative replies, we introduce cue words as the additional information during generation. Naturally, the key point lies in how to incorporate such information. One of the prevailing methods is modifying the neural cell by various gating mechanisms. However, these approaches are designed specially for a specific scenario, and not effective as expected when they are employed to other tasks. To tackle this issue, we propose the cue word GRU, another independent neural cell, to deal with the auxiliary information. Since this neural cell can be replaced easily by other units, it greatly improves the flexibility and reusability.

Given the last hidden state $h_{t-1}$ , the additional cue word $C_w$ and the current attention-based context $C_t$, the new hidden state of the auxiliary de-

coding $h_w$ is computed by following equations:

$$
\begin{aligned}
r_w &= \sigma(W_r C_w + U_r h_{t-1} + U_{cr} C_t + b_r) \\
z_w &= \sigma(W_z C_w + U_z h_{t-1} + U_{cz} C_t + b_z) \\
\widetilde{h_w} &= \tanh(W_h C_w + U_h(r_w \circ h_{t-1}) + U_{ch} C_t + b_h) \\
h_w &= (1 - z_w) \circ h_{t-1} + z_w \circ \widetilde{h_w}
\end{aligned}
\tag{9}
$$

where $W$'s and $U$'s are weights and $b$'s are bias terms like those in the standard GRU. Note that the standard GRU does not share parameter matrixes with the cue word GRU. The "Cue word GRU" in Figure 3 describes the auxiliary decoding process.

#### 3.3.3 Fusion unit

To combine both the general decoding information and the auxiliary decoding information, we apply the fusion unit (Arevalo et al., 2017) integrating the hidden states of both standard GRU, i.e., $h_y$, and the cue word GRU, i.e., $h_w$, to compute the current hidden state $h_t$. The equations are as follows:

$$
\begin{aligned}
h'_y &= \tanh(W_1 h_y) \\
h'_w &= \tanh(W_2 h_w) \\
k &= \sigma(W_k[h'_y, h'_w]) \\
h_t &= k \circ h_y + (1 - k) \circ h_w \\
\theta &= \{W_1, W_2, W_k\}
\end{aligned}
\tag{10}
$$

with $\theta$ the parameters to be learned. From the equations above we can see that, the gate neuron $k$ controls the contribution of the information calculated from $h_y$ and $h_w$ to the overall output of the unit.

### 3.4 Model Training

When training on the aligned corpus, we randomly sample a noun in the reply as the cue word. The objective function was the cross entropy error between the generated word distribution $p_t$ and the actual word distribution $y_t$ in the training corpus.

## 4 Experiments

In this section, we compare our method with the-state-of-art response generation models based on a huge conversation resource. The objectives of our experiments are to 1) evaluate the effectiveness of our proposed HGFU model, and 2) explore how cue words affect the process of reply generation.

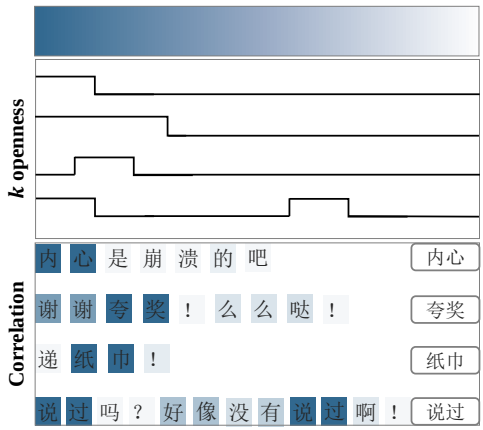


Figure 4: Heat map and the $k$ gate openness. Bottom: The correlation between the generated reply words and the cue word. Top: The openness of $k$ gate in fusion unit.

### 4.1 Experimental setup

We evaluated our model on a massive Chinese dataset of human conversation crawled from the Baidu Tieba[1] forum. There are 500,000 $\langle query - reply \rangle$ pairs for training, 2,000 for validation, and another unseen 27,871 samples for testing. In total, we kept about 63,000 distinct words.

In our experiments, the encoder, the standard decoder and the cue word decoder have 1,000 hidden units; the word embedding dimensionality is 610 which were initialized randomly and learned during training. We applied AdaDelta with a mini-batch size of 80 for optimization. These values were mostly chosen empirically. In order to prevent overfitting, early stopping was implemented using a held-out validation set.

### 4.2 Comparison Methods

In this paper, we conduct extensive experiments to compare our proposed method against several representative baselines. All the methods actually are implemented in two ways to utilize the cue word, which are local information initialization and global information inception.

**rGRU**: Through a specially designed Recall gate (Xu et al., 2016), domain knowledge was transformed into the extra global memory of a deep neural network.

**SCGRU**: In SCGRU (Wen et al., 2015b), an additional control cell was introduced to gate the dia-

[1] http://tieba.baidu.com

| Query (**Cue word**) | 班主任还拍了我超级丑的照片已被笑死.(**上镜**)<br>The teacher took a photo of me; it was really ugly and people laughed at me. (**Photogenic**) | Related Criterion | Labels |
|---|---|---|---|
| Reply1 | 谁的照片？ Whose photo? | Logic Consistency | Unsuitable |
| Reply2 | 什么时候拍的？ When did he took the photo? | Implicit Relevance | Neutral |
| Reply3 | 抱抱。Give you a hug. | Implicit Relevance | Neutral |
| Reply4 | 我拍照也都是巨丑的！ My photos are also ugly! | —— | Suitable |

Table 2: An example query, corresponding cue word in **bold** and its candidate replies with human annotation. The query states that people laughed at the author's photo, it is unsuitable to ask the ownership of this photo in Reply1. Generally, Reply2 and Reply3 apply to this scenario, but they do not reflect semantic relevance with the cue word. Reply4 talks about the respondent's situation and related to "Photogenic", thus it is a suitable response.

logue act (DA) features during the generation process.

**SLGD**: We implemented the Stochastic Language Generation in Dialogue (SLGD) method (Wen et al., 2015a), which added additional features in each gate of the neural cell.

**FGRU**: To explore more fusion strategies, intuitively, we fused the cue word and hidden states by vector concatenation during the decoding process.

Note that rGRU and SCGRU incorporate additional information by gating mechanisms, while SLGD and FGRU fuse the information into each gate of the neural cell directly.

### 4.3 Experiment Evaluation

**Objective metrics.** To evaluate the performance of different methods for the conversation generation task, we leverage BLEU (Papineni et al., 2002) as the automatic evaluation metric, which is originally designed for machine translation and evaluates the output by using n-gram matching between the output and the reference. Here, we use BLEU-1, BLEU-2 and BLEU-3 in our experiments.

**Subjective metrics.** Since automatic metrics may not consistently agree with human perception (Stent et al., 2005), human testing is essential to assess subjective quality. Hence, we randomly sampled 150 queries in the test set, then we invited five annotators to offer a judgment. For fairness, all of our human evaluation was conducted in a random, blind fashion, i.e., replies obtained from the five evaluated models are pooled and randomly permuted for each annotator. Three levels are assigned to a reply with scores from 0 to 2: 0 = Unsuitable reply, 2 = Suitable reply, and 1 = Neutral reply.

| Method | | BLEU-1 | BLEU-2 | BLEU-3 | Human score |
|---|---|---|---|---|---|
| Local | rGRU | 1.087 | 0.419 | 0.249 | |
| | SCGRU | 2.135 | 0.622 | 0.255 | |
| | SLGD | 1.678 | 0.508 | 0.209 | |
| | FGRU | 2.262 | 0.598 | 0.208 | |
| | HGFU | 1.861 | 0.545 | 0.209 | |
| Global | rGRU | 1.793 | 0.676 | 0.277 | 0.542 |
| | SCGRU | 3.637 | 0.981 | 0.369 | 0.73 |
| | SLGD | 4.146 | 1.059 | 0.367 | 0.71 |
| | FGRU | 4.197 | 1.013 | 0.282 | 0.677 |
| | HGFU | **4.893** | **1.225** | **0.393** | **0.942** |

Table 4: Performance of evaluated methods.

To make the annotation task operable, the suitability of the generated reply is judged not only based on *Grammar and Fluency, Logic Consistency and Semantic Relevance* following (Shang et al., 2015), but also *Implicit Relevance*, i.e., the generated reply should be semantically relevant to the predicted cue word, no matter the cue word explicitly appears in the reply or not. If any of the first three criteria is contradicted, the reply should be labeled as "Unsuitable". Only the replies conforming to all requirements are labeled as "Suitable". Table 2 shows an example of the annotation results of a query and its replies. The first reply is labeled as "Unsuitable" because of the logic consistency. Reply2 and Reply3 are not semantically related to the cue word, and is therefore annotated as "Neutral".

### 4.4 Overall Performance

The overall results against all baseline methods are listed in Table 4. Our proposed HGFU model in global schema obviously shows better performance than the baseline methods; it obtains the

| | Chinese Sentence | English Tranlation |
|---|---|---|
| Query | 写的真心棒！(**夸奖**) | What a nice written! (**Appreciation**) |
| Reply | 谢谢夸奖！么么哒！ | Thanks for your appreciation! Love you! |
| Query | 还是无法淡定。(**内心**) | Still cannot calm down. (**Heart**) |
| Reply | 内心是崩溃的吧。 | Your heart must be broken. |
| Query | 我先去哭一会。(**纸巾**) | I am going to cry for a while. (**Tissue**) |
| Reply | 递纸巾！ | Offer you a tissue! |
| Query | 当初你们不是说过他是诺维斯基吗？(**说过**) | Didn't you say that he was $Nowitzki^{\dagger}$? (**Say**) |
| Reply | 说过吗？好像没有说过啊！？ | Did I say it? I don't seem to say it!? |

Table 3: The explicit introducing-content cases of our HGFU model. The predicted cue word in **bold** explicitly occurs in the generated reply. $Nowitzki^{\dagger}$ is a NBA basketball player.

highest BLEU scores as well as the highest human score.

In terms of automatic evaluations, the global-based methods perform much better than a set of local-based methods, which demonstrates the effectiveness of global information inception. As mentioned above, the global schema provides the model a broader and more flexible space for learning, which is benefit for information fusion. When it comes to human scores (For the sake of convenience, we only conducted human evaluation in global schema), there are similar conclusions to BLEU results.

From Table 4, we can see that the performance of rGRU is not as good as the other systems, while SCGRU outperforms the others in the local pattern and shows comparative performance in the global schema. These two methods both augment the standard neural network with specially designed gate to control the cue word, but the results vary greatly. It is the limitation of gating mechanisms that is lacking in adaptiveness. Besides, SLGD adding cue word term in each gate of the neural cell has the similar result as FGRU method, which concatenates cue word with hidden state. Basically, our proposed HGFU has a significant improvement against the baseline systems. The most probable credits come from the cue word GRU: we apply the extra GRU unit to control the auxiliary information instead of fusion in the standard GRU, which is more flexible.

Till now, we have elaborated the overall performance of all methods. Next we will come to a closer look at some representative cases of our HGFU model for further analysis and discussions.

### 4.5 Analysis and Case Studies

Given a query and the cue word, our HGFU model generates a meaningful and informative response. In Table 3, the predicted cue word occurs in the generated response and we treat this kind of generation as the explicit introducing-content. However, we do not strictly restrict tothis. As shown in Table 5,our HGFUmodel also generates the replies without containing the cue word, while the responsesare still somehow related to the cue word and the query. This reflects our expectation: the information of the cue word will be fused into the generation process but not necessarily occur explicitly. It provesthe characteristics of our proposed new "soft" schema, whichare more flexible, extensible, and controllable.

We further analyze these explicit cases using a heat map as shown in Figure 4. We use various shades of blue to present the extent of correlation between the cue word and the generated reply. The darker the blue is, the higher correlation they have. For the added information in the reply (Here is exactly the cue word in darkblue), its position and occurrence times are not fixed, which are autonomously controlled by our model.

Besides, the rectangular pulse is also a significant presentation of this correlation, which indicates how the $k$ gate in fusion unit balance the influence of $h_y$ and $h_w$. When in the high level of the rectangular pulse, $k$ "opens" the switch of $h_w$ to generate the current word; when in the low level, the fusion unit mainly takes $h_y$ for generation. We observe that the switch corresponds with the heat map: the generated word is more correlated with the cue word when the switch is open.

| | Chinese Sentence | English Tranlation |
|---|---|---|
| Query | 在微博看到这样一个评论真的是完全无法反驳。(**观点**) | I saw such a comment in the microblog which cannot be refused completely. (**View**) |
| Reply | 路人表示不服啊！ | As an outsider, I am not convinced! |
| Query | 怎么突然就下雨了？(**委屈**) | Why is it raining suddenly? (**Grievance**) |
| Reply | 好伤心啊。 | So sad. |
| Query | 泰民这张也是做了很久桌面。(**屏保**) | This photo of $Taemin^{\dagger}$ was also taken as a desktop for a long while. (**Screenshot**) |
| Reply | 锁屏吗? | As the lockscreen? |
| Query | 混脸熟求勾搭！(**小新**) | Make acquaintance and seek chances for further relations! (**Freshman**) |
| Reply | 同新人！求认识。 | I am also the new! Nice to meet you. |

Table 5: The implicit introducing-content cases of our HGFU model. The cue word in **bold** is not contained in the reply, while the response is still related to the cue word. $Taemin^{\dagger}$ is a Korean singer.

## 5 Related work

### 5.1 Conversation Systems

Automatic human-computer conversation has attractedincreasing attention over the past few years. At the very beginning, people start the research using hand-crafted rules and templates (Walker et al., 2001; Misu and Kawahara, 2007; Williams et al., 2013). These approaches require no data or little data for trainingbuthuge manual effort to build the model, which is very time-consuming. For now, buildinga conversation systemmainly falls into two categories: retrieval-based and generation-based. As information retrieval techniques are developing fast, Leuski et al. (2009) build systems to select the most suitable response from the query-reply pairs using a statistical language model in cross-lingual information retrieval. Yan et al. (2016) propose a retrieval-based conversation system with the deep learning-to-respond schema through a deep neural network framework driven by web data. Recently, generation-based conversation systems have shownimpressive potential. Shang et al. (2015) generate replies for short-text conversation by Seq2Seq-basedneural networks with local and global attentions.

### 5.2 Content Introducing

In vertical domains, Wen et al. (2015b) apply an additional control cell to gate the dialogue act (DA) features during the generation process to ensure the generated repliesexpressthe intended meaning. Also, the Stochastic Language Generation in Dialogue method (Wen et al., 2015a) adds additional features in each gate of the neural cell. Xu et al. (2016) introduce a new trainable gate to recall the global domain memory to enhance the ability of modeling the sequence semantics. Different from the above work, our paper addresses the problem of content introducing in the open-domain generative conversation systems.

In open domains, Xing et al. (2016) incorporate topic information into Seq2Seq framework to generate informative and interesting responses. To provide informative clues for content introducing, Li et al. (2016b) detect entities from previous utterances and search for more related entities in a large knowledge graph. A very recent study similar to ours is Mou et al. (2016), where the predicted word explicitly occurs in the generated utterance. Unlike the existing work, we explore an implicit content-introducing method for neural conversation systems, which utilizes the additional cue word in a "soft" manner to generate a more meaningful response given a user-issued query.

## 6 Conclusion

In this paper, we explore an implicit content-introducing method for generative short-text conversation system. Given a user-issued query, our proposed HGFU incorporates an additional cue word in a "soft" manner to generate a more meaningful response. The HGFU model consists of three components: the standard GRU, the cue word GRU and the fusion unit. The standard GRU operates a general decoding process, and the cue word GRU imitates this process but treats the predicted cue word as the current input. As for the fusion unit, it combines both the hidden states of the standard GRU and the cue word GRU to generate

the current output word. The experimental results demonstrate the effectiveness of our approach.

## Acknowledgments

This work was supported by the National Hi-Tech R&D Program of China No. 2015AA015403; the National Science Foundation of China No. 61672058.

## References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Charles Lee Isbell, Michael Kearns, Dave Kormann, Satinder Singh, and Peter Stone. 2000. Cobot in lambdamoo: A social statistics agent. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 36–41. AAAI Press.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2009. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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*, pages 110–119. Association for Computational Linguistics.

Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016b. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2845–2851.

Teruhisa Misu and Tatsuya Kawahara. 2007. Speech-based interactive information guidance system using question-answering technique. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–145. IEEE.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3349–3358.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3776–3783. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *Deep Learning Workshop of the 32nd International Conference on Machine Learning*.

Marilyn A Walker, Rebecca Passonneau, and Julie E Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 515–522.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2016. Topic augmented neural response generation with a joint attention mechanism. *arXiv preprint arXiv:1606.08340*.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2016. Incorporating loose-structured knowledge into lstm with recall gate for conversation modeling. *CoRR*, abs/1605.05110.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64. ACM.