Speaker or Listener? The Role of a Dialogue Agent

Yafei Liu[♢♡†], **Hongjin Qian**[△†], **Hengpeng Xu**[♡], **Jinmao Wei**[♡‡]
[♢] Jarvis Lab Tencent [△]Renmin University of China [♡]Nankai University
davenliu@tencent.com ian@ruc.edu.cn xuhengpeng@mail.nankai.edu.cn weijm@nankai.edu.cn

Abstract

For decades, chitchat bots are designed as a listener to passively answer what people ask. This passive and relatively simple dialogue mechanism gains less attention from humans and consumes the interests of human beings rapidly. Therefore some recent researches attempt to endow the bots with proactivity through external knowledge to transform the role from a listener to a speaker with a hypothesis that the speaker expresses more just like a knowledge disseminator. However, along with the proactive manner introduced into a dialogue agent, an issue arises that, with too many knowledge facts to express, the agent starts to talks endlessly, and even completely ignores what the other expresses in dialogue sometimes, which greatly harms the interest of the other chatter to continue the conversation. To the end, we propose a novel model named Initiative-Imitate to interact with adaptive initiative throughout a dialogue. It forces the agent to express in parallel with the appropriate role during the whole conversation. The corresponding experiments show the proposed Initiative-Imitate obtains competitive results both on the automatic and manual metrics. And the fluency and engagement of the chatbot have also been improved significantly. Besides, the case study indicates the Initiative-Imitate can constantly transfer to appropriate role timely and response more properly during the whole continuous conversation.

1 Introduction

Automatic human-machine conversation lies in the core of artificial intelligence (AI) and natural language processing (NLP). Many researchers have developed lots of dialogue systems, such as rule-based (Weizenbaum et al., 1966; Webb, 2000), retrieval-based (Wu et al., 2017) and generation-based like neural networks (Vinyals and Le, 2015; Vougiouklis et al., 2016; Yavuz et al., 2019).

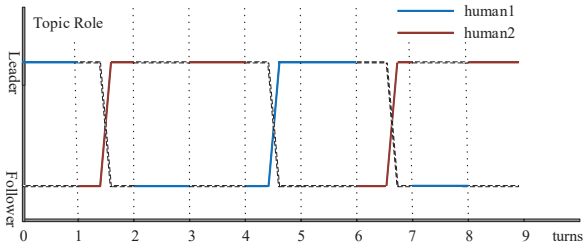

Figure 1: The role transformation of the two humans in a multi-turn human-human conversation started by human1. The dashed line means one's silent period, the solid line for the speaking period.

For decades, machine agent is generally presumed to be a passive role with the ability to answer what humans ask. However, this passive and relatively simple response mechanism consumes the interest of the other dialogue participant rapidly (Li et al., 2016b). Actually in a continuous human-human conversation, both participants need to be a speaker to lead current topics. This phenomenon is statistically summarized from the analysis on a real human-human dataset named DailyDialog (Li et al., 2017) as depicted in Figure 1. Two humans take the topic leading role like a speaker to introduce something new in turns. Thus in human-machine conversation, the dialogue agent side needs to act as a speaker timely and appropriately.

Furthermore, some researches (Yavuz et al., 2019; Ghazvininejad et al., 2018b; Wu et al., 2019; Zheng et al., 2019) try to make use of external knowledge to endow the machine agent with the ability to express proactively when generating responses. Models facilitated with external knowledge indeed generate more meaningful responses than peers that train only on the source-target dialogue dataset. However, these models tend to

[†]Equal contribution
[‡]Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4861–4869

fall into another situation where the machine agent talks too much ignoring what the other has said, let alone the inappropriate use of knowledge. Therefore, this response manner still can't gain more attention and satisfy the human being's practical expectations.

With these in mind, we start our work. The first and most important thing is a dataset with the role labeled. However, to obtain a role-annotated dialog dataset is both time-consuming and effort-consuming because in many real and practical dialogues one plays both the speaker and listener role alternately and switches from one role to the other irregularly. Luckily, last year Wu et al. (2019) proposed to guide dialogue with explicit goals and released their dataset Duconv [1]. The dataset collection setting is that one person plays a proactive role leading the whole dialogue as a content transmitter, while the other follows as an apprentice just like a listener. The dialogue roles on this dataset are pretty explicit and easily obtained.

The second is proper knowledge fusion. Previously, Ghazvininejad et al. (2018b) fused knowledge facts representation by adding it to the encoder context vector. Zhou et al. (2018) fused knowledge facts representation by concatenating it with the encoder context vector. More researches (Vougiouklis et al., 2016; Yavuz et al., 2019) fed facts representation into the decoder state to predict the response. These models provide complete knowledge representation without controlling the proportion of knowledge to fuse. Our knowledge fusion module is inspired by the Child-Sum method (Tai et al., 2015; Zoph and Knight, 2016), Tai fused different tree-structured long short-term memory networks (LSTM) outputs into one output. Zoph combined multi-source translations into unified representation space to perform the translation. It is a similar circumstance in knowledge dialogues because external knowledge facts can be also viewed as one source of input and incorporated with other source inputs organically.

Finally it's time to resolve the problem when to be a speaker. We know a speaker holds more abilities to express knowledge. So in a two-side dialogue when one becomes proactive to convey some useful and meaningful knowledge facts, the other attempts to be a listener. As a result, the dialogue proceeds further with both sides deeply involved in. This also means the proactive side needs more knowledge to prepare the response. While it's opposite for a listener. A listener tends to know less about knowledge and responds based on what the other has said rather than the knowledge. This behavior is in parallel with the characteristic of the forget gate in the Child-Sum method. The forget gate controls how much knowledge to be discarded during the multi-source fusion phase. When one leads the dialog, the other takes a backseat and forgets more. Therefore we predict the role with this forget gate and generate a response not only on the default decoder state but also on the predicted role simultaneously. And we will detail the concrete procedures later. Furthermore, we introduce the metric Engagement in website analysis into the dialogue quality measurement to measure the degree that a chatbot is involved when to converse with a human.

In this paper, our contributions are summed up to three points:

- We first point out the role-to-play problem existing in automatic human-machine conversation and design a solution for it.
- A new knowledge fusion model is proposed to fuse knowledge selectively and appropriately along with the role, **Initiative-Imitate**, which is applied to imitate the volatile initiative of the chatter in conversations.
- Different from the previous common manual evaluation metric Fluency and Coherence, Engagement is firstly introduced to measure how deeply a chatbot is engaged when to chat with a human.

2 Related Work

In the last several years, neural conversational models become prevalent due to a new round of artificial intelligence. And usually, they provide better responses than early rule-based or template-based dialogue systems (Webb, 2000; Varges et al., 2009). Before this paper, there are two dominated trends in human-machine conversation research: the passive and the proactive. Actually the passive starts much earlier than the other.

The passive models always attempt to answer what a human asks. Among these models, Shang et al. (2015) trained end-to-end neural conversation models on massive data. Li et al. (2016a) proposed a Maximum Mutual Information objective function

[1] https://ai.baidu.com/broad/subordinate?dataset=duconv

to promote generating diverse responses. However, these dialogue systems often generate generic, safe, and inconsistent responses (Ram et al., 2018). This leads to the arising of the second research line of neural conversation models.

Utilizing knowledge in dialogue can generate diverse, engaging, meaningful and personalized responses in a way. Ghazvininejad et al. (2018b) fused knowledge encoder representation with encoder final state to initialize the initial decoder state by element-wise sum. Vougiouklis et al. (2016) fed knowledge representation at each decoder hidden state computation. Zhou et al. (2018) encoded commonsense knowledge during searching and attended over retrieved sub-graph when generating words. Zhang et al. (2018) introduced the PERSONA-CHAT dataset with dialogue agents personalized and showed various baseline performance including some generation based models like Seq2Seq and Profile Memory. DeepCopy (Yavuz et al., 2019) applied the copy mechanism both on source words and knowledge facts words during the word generation period. This enhances the system's expressive ability especially in terms of the out-of-vocabulary words. Wu et al. (2019) applied KnowlegePost (Lian et al., 2019) to minimize the divergence between knowledge prior and knowledge posterior distribution to learn a better knowledge representation and concatenated the knowledge context vector with decoder feed.

Zoph and Knight (2016) fused multi-source languages aligned to the target translation with Child-Sum Tree-LSTMs (Tai et al., 2015), and we draw enlightenment from the Child-Sum Tree-LSTMs and apply it into knowledge fusion as a completely novel knowledge fusion method. However, this doesn't resolve the current role-to-play problem during the response phase, that is whether the agent needs to be a listener or speaker at present. Looking into Child-Sum Tree-LSTMs, we have found that the forget gate controls the proportion of one source to forget. It's just like the behavior of the role. The speaker needs more knowledge to transmit relative knowledge to the other, so he needs to forget less knowledge. While the listener requires more forgetting to decrease the influence of the knowledge input. As a result, it seems the forget gate here is just a hidden variable which represents the role. So we keep an absolute role label (1: proactive, 0: passive) to supervise the forget gate and make the forget gate to supervise the knowledge fusion proportion.

Different from the knowledge models above, our Initiative-Imitate fuses knowledge with source context in a novel manner. And the utilized knowledge adaptively changes according to the predicted role at each turn in a multi-turn dialog. To the best of our knowledge, our Initiative-Imitate is the first chatbot which models the role initiative of the participants in a continuous dialogue.

3 Model

In this section, we first set up the problem, then demonstrate the individual modules in the proposed model framework in Figure 2.

3.1 Problem Definition

In general dialogue systems, we make $x = \{x_1, x_2, ..., x_{n_x}\}$ as the dialogue source input. x usually represents the dialogue history. $y = \{y_1, y_2, ..., y_{n_y}\}$ is the response. The traditional conversation models take the $\{x, y\}$ pair as data set, and then feed it into a neural model to learn the conditional probability of y given x: $P(y|x)$.

In knowledge dialogues, there are also some sentence-level knowledge sequences, denoted as $k = \{k_1, k_2, ..., k_{n_k}\}$. For each k_i, $k_i = \{k_{i_1}, k_{i_2}, ..., k_{i_{n_{k_i}}}\}$. x_i and k_{i_j} are both word-level tokens. n_* means the count of elements of knowledge sequence $*$. Both history context x and knowledge items k are inputs. So the conditional probability of y given x and k shall be $P(y|x, k)$. Note that y is open-ended. That is, y is generated token by token rather than selected from a candidate set.

What's more in contrast to the previous knowledge models is the proper role to play. We formulate it as $role$ here. And the conditional probability turns to be $P(y|role; x, k)$. $role$ is predicted by the forget gate in the knowledge fusion module. During training, the prediction of $role$ is supervised with the labeled role of the current turn. And then y is predicted with both predicted $role$ and decoder state.

3.2 Encoder

Encoders encode variable-length input sequence into fixed-length vector representation through recurrent neural network (RNN) typed models, i.e.

$$h_t = f(x_t, h_{t-1}); \quad c = \phi(\{h_1, h_{n_x}\}), \qquad (1)$$

where h_t is the RNN state, c is the so-called context vector, f is the dynamics function, for example,

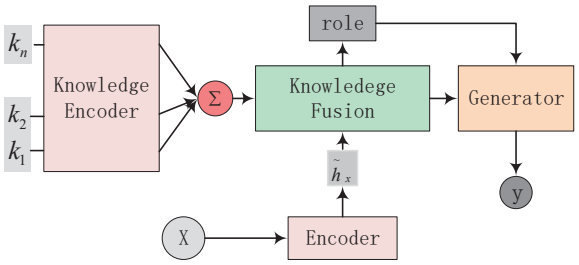

Figure 2: The framework of Initiative-Imitate with attention excluded.

LSTM (Hochreiter and Schmidhuber, 1997). And ϕ summarizes the hidden states, e.g. summing the hidden states.

Source x and knowledge k are both encoded according to equation 1 without parameters shared. The source encoder encodes dialogue history x into hidden states $h_t^{(x)}$. While each fact sequence k_i in the knowledge is encoded as $h_t^{k_i}$, $t \in \{1, 2, ..., n_{k_i}\}$.

3.3 Knowledge Fusion

After encoding source x and knowledge k, we need to fuse them to obtain the joint representation h. First, we need the context vector $\tilde{h_x}$ and $\tilde{h_k}$ corresponding to the source hidden states $h_t^{(x)}$ and knowledge hidden states $h_t^{k_i}$. We formulate them as follows:

$$\tilde{h_x} = h_{n_x}; \quad \tilde{h_k} = \sum_{j=1}^{n_f} h_t^{k_i}. \tag{2}$$

$h_j^{k_i}$ is the last hidden state of $h_t^{k_i}$, h_{n_x} the last hidden state of x.

As depicted in Figure 3, we fuse $\tilde{h_x}$ and $\tilde{h_k}$ into h. The precise formulations are as follows:

$$i = \sigma(W_1^i \tilde{h_x} + W_2^i \tilde{h_k} + b^i), \tag{3}$$

$$\begin{aligned} f_x &= \sigma(W_1^{f_x} \tilde{h_x} + b^{f_x}); \\ f_k &= \sigma(W_2^{f_k} \tilde{h_k} + b^{f_k}), \end{aligned} \tag{4}$$

$$o = \sigma(W_1^o \tilde{h_x} + W_2^o \tilde{h_k} + b^o), \tag{5}$$

$$u = tanh(W_1^u h_x + W_2^u h_k + b^u), \tag{6}$$

$$c = i \odot u + f_x \odot \tilde{c_x} + f_k \odot \tilde{c_k}, \tag{7}$$

$$h = o \odot tanh(c). \tag{8}$$

All W and b are trainable parameters. The symbol $\odot$ means an element-wise multiplication. In equation 7, it obtains new current cell state in LSTM with $\tilde{c_x}$ and $\tilde{c_k}$, so $\tilde{c_x}$ and $\tilde{c_k}$ are not the context vector as in Equation 1 but cell state. Note that we have f_k in Equation 4. It stands for the proportion to forget of the cell state $\tilde{c_k}$. We will show how to apply it to predict the role in the section below.

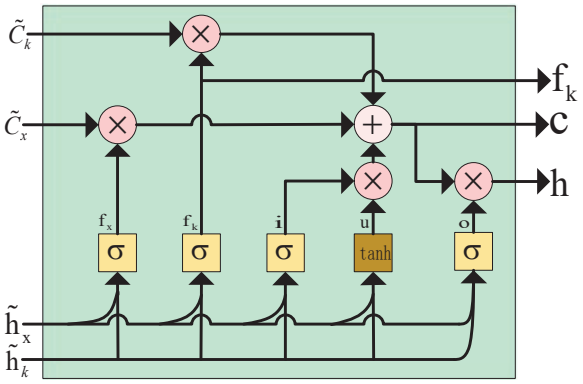

Figure 3: Knowledge Fusion Module

3.4 Generator

Before generating a word, the role $role_p$ needs to be predicted as follow:

$$role_p = \sigma(W f_k + b). \tag{9}$$

And the predicted role $role_p$ will be concatenated with c_t in Equation 11 in (Yavuz et al., 2019) to obtain a new c_t.

Our model generator generates words in two modes: generate mode and copy mode. The former generates words on a fixed vocabulary token by token. While the latter copies tokens from the input sources (x and k). Following (2019), we first compute the two distributions on the two modes, and then integrate the two distributions into the final. Luong attention (Luong et al., 2015) is applied in the decoder phase. Please refer to (Yavuz et al., 2019; Luong et al., 2015) to get the thorough generation process.

3.5 Loss

Traditional objective loss function in dialogue systems will include negative log-likelihood (NLL) loss function as follow:

$$L_{NLL} = -\frac{1}{|y|} \sum_{t=1}^{|y|} log(p_t(y_t | y_{<t}, x, k)). \tag{10}$$

NLL measures the difference between the gold response and the generated response as the main metric in most training objective function. There are also some other auxiliary losses applied in dialogue model training. One is the bag-of-word (BOW) loss proposed by Zhao et al. (2017). It measures the relevance between the encoded knowledge k_c and the golden response, and helps learn a better representation of the knowledge context vector. The followings are the formulation of BOW loss:

$$L_{BOW} = -\frac{1}{|y|}\sum_{t=1}^{|y|} log(p(y_t|\tilde{h_k})), \quad (11)$$

$$p(y_t|\tilde{h_k}) = softmax(MLP(\tilde{h_k})). \quad (12)$$

MLP function transforms the input vector $\tilde{h_k}$ into a vector with a size of target vocabulary. Then the function $softmax$ gets a distribution over the target vocab.

To this moment, the role prediction problem is still kept in hand. As mentioned in section 1, it's necessary to observe the role state of the other participant throughout the whole conversation in order to react and respond in a proper and gentle manner. After studying the dialogue manners between human beings, we have found that the characteristic of the role astonishingly parallels with the behavior of the knowledge forget gate f_k in Equation 4. Let 1 stand for the absolute speaker role, 0 for the absolute listener role. Generally, when a human expresses a lot of contents or knowledge as a speaker, the bot activates the listener mode, so it forgets more knowledge reducing the influence of too much knowledge. At the same time, the values of f_k become very small. For the contrary case, f_k shall be bigger. So we use f_k to predict the role to play. The explicit loss function formulates as follow:

$$L_{Role} = BCE(role_p, role). \quad (13)$$

Here BCE is the binary cross-entropy function. The $role$ is the ground-truth role label.

Summing up all token-level losses defined above yields the final loss function. Hence the final loss is:

$$Loss = L_{NLL} + L_{BOW} + L_{Role}. \quad (14)$$

Finally, the overall architecture of our model comes to the surface with the accomplishment of objective function definition including the role-control mechanism.

4 Experiments

The experiment part consists of several procedures: dataset preparation to adapt our task, a brief baseline introduction, training setup, and the result comparison with analysis.

4.1 Dataset Preparation

As far as we know, all current dialogue datasets don't include a label indicating a speaker role or listener role in a session of conversation. To annotate dataset with role state is both time-consuming and effort-consuming. Fortunately, Wu et al. (2019) released a dataset Duconv, whose dialogue collection setting is that one man plays the role of leader, the other as a follower. So the role states here are pretty explicit. Furthermore, not like PERSONA-CHAT (Zhang et al., 2018), the form of knowledge in Duconv is the triplet. However, we found that the unnatural concatenated triplet may harm the performance of models slightly. So we transform the triplet into natural language with some stopwords or conjunctions. For example, the triplet (”战狼/ Wolf Warriors”, ”主演/ protagonist”, ”吴京/ Jason Wu”) will be transformed into ”战狼的主演是吴京/ The protagonist of Wolf Warriors is Jason Wu”. Besides, we follow Wu et al. (2019) to perform normalization on the topic, which is proved to be effective in better response generation.

4.2 Baselines

Only generative models are considered to be the baselines here. What follows are several baselines of generation-based models [2]:

- Seq2Seq: Standard sequence to sequence without attention.
- Seq2Seq$_{attn}$ (Luong et al., 2015): Standard sequence to sequence with global attention.
- CopyNet (Gu et al., 2016): Standard sequence to sequence with copy mechanism which may copy words from the source end when to decode.
- Generation-base [3] : This generation-based model is released by Wu et al. (2019) along with their dataset.

[2]Memory-based models are not included because MemNet (Ghazvininejad et al., 2018a) doesn't perform better than Seq2Seq on Duconv (Wu et al., 2019).

[3]https://github.com/PaddlePaddle/models/tree/develop/PaddleNLP/Research/ACL2019-DuConv

Models	F1	BLEU-1	BLEU-2	Distinct-1	Distinct-2
Seq2Seq	37.60	0.265	0.172	0.083	0.189
Seq2Seq$_{attn}$	38.26	0.264	0.173	0.085	0.192
CopyNet	39.01	0.229	0.154	**0.132**	0.307
Generative-base	35.98	**0.341**	0.189	0.062	0.178
DeepCopy	43.31	0.308	0.213	0.129	0.311
Initiative-Imitate	**44.11**	0.335	**0.231**	0.127	**0.319**

Table 1: Evaluation results for different models with automatic metric. Best scores on each metric are in bold.

- DeepCopy (Yavuz et al., 2019): A knowledge model decoding with a hierarchical pointer network.

4.3 Setup

All models except Generation-base are implemented with OpenNMT [4] (Klein et al., 2017). We use a fixed vocabulary including 30000 most frequent tokens and a dynamic dict with source input and relative knowledge tokens in terms of the copy mechanism. Pre-trained word vectors [5] are from Li et al. (2018). Encoders and Decoders in every model are both 2-layer LSTMs with the same hidden size 500. The model parameters are optimized with Adam with a batch size of 64, a fixed learning rate of 0.001, decay after 10000 steps with weight decay rate 0.5. And during the test, we apply the beam search strategy with size 5.

4.4 Results

In this section, we present the experimental results in terms of both automatic measures and human evaluation.

4.4.1 Automatic Evaluation

Table 1 demonstrates the automatic evaluation results of different models on several metrics. F1 and BLEU-1 measure the similarity between predictions and golden responses on the uni-gram level, BLEU-2 on the bi-gram level. Distinct metric measures the token-level diversity of response on corpus level. The Initiative-Imitate obtains the best results on F1, BLEU-2, Distinct-2. The Generation-base model ranks first on BLEU-1, but worse on other metrics. We suppose it's associated with the mandatory utilizing of knowledge because we have found more repetitions of knowledge in the responses by this model. And this also leads to bad diversity which is reflected from the Distinct scores. This also indicates that excessive utilizing of knowledge won't guarantee a better response when to chat. The proper proportion of knowledge utilizing helps more. This corresponds to the initial idea of an appropriate role to play during a chat. Being a speaker means the bot needs to express more knowledge, but not that case for a listener. With copy mechanism, CopyNet, DeepCopy, and Initiative-Imitate all gain better and comparable results in terms of the Distinct. It can be inferred that the copy mechanism indeed increases the diversity of response to some extent.

Finally, our model Initiative-Imitate is inherited from DeepCopy, and surpasses the DeepCopy model by a moderate margin, which proves the positive effectiveness of our knowledge fusion module and role-control setting in a way.

4.4.2 Manual Evaluation

For manual evaluation, we first select metrics Fluency and Coherence clearly defined in (Wu et al., 2019) to measure whether the dialog agent can express fluently and logically. However, there doesn't exist a metric which tells how deeply the agent is engaged in the dialog. Inspired by the metric *Engagement* in web analysis, a new dialog met-

Models	Flu.	Coh.	Eng.
Seq2Seq	1.40	1.05	0.25
Seq2Seq$_{attn}$	1.45	1.10	0.50
CopyNet	1.75	1.30	0.15
Generative-base	0.95	0.80	0.40
DeepCopy	**1.80**	1.68	0.30
Initiative-Imitate	**1.80**	**1.75**	**0.70**

Table 2: Manual evaluation on three metrics. The range of Fluency (Flu.) and Coherence (Coh.) are both from 0 to 2. While the range of Engagement (Eng.) is between 0 and 1. The higher value is better.

[4] https://github.com/OpenNMT/OpenNMT-py

[5] https://drive.google.com/open?id=1kSAl4_AOg3_6ayU7KRM0Nk66uGdSZdnk

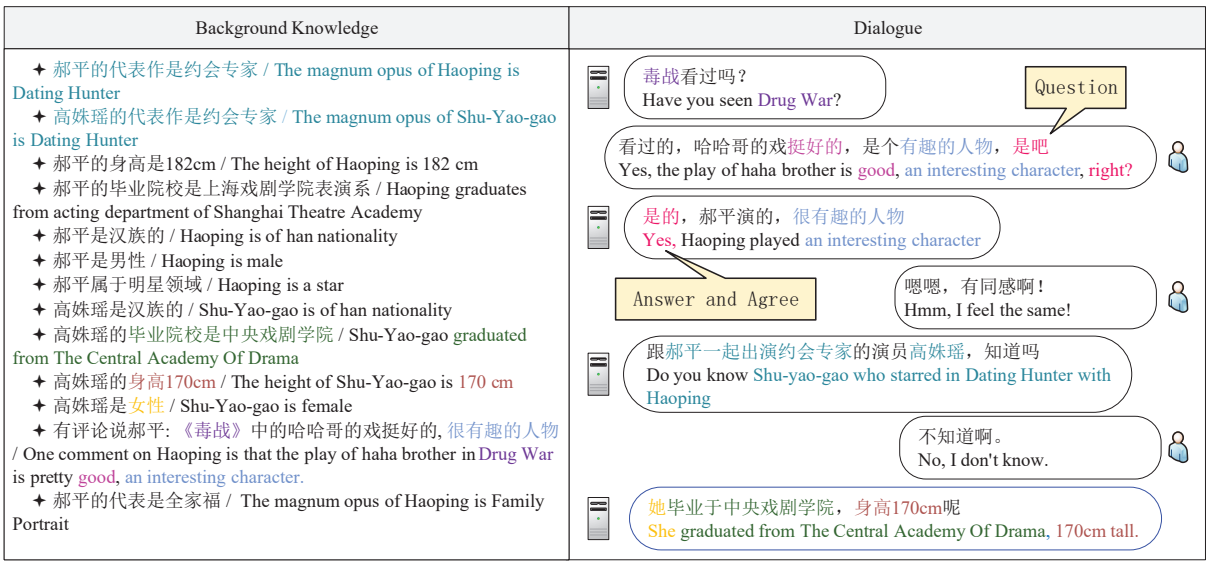

Figure 4: One case by Initiative-Imitate

ric *Engagement* (Eng.) is developed as follow:

$$Eng. = \frac{card(\{u_a : u_a\, answers\, any\, u_q\})}{card(\{u_q : u_q\, is\, a\, question\})} \quad (15)$$

u_q means a question utterance that expects a clear and direct reply. While u_a is an utterance which follows a u_q and then answers this u_q to the point. The $card$ function will count the element count of an utterance set. Thereby, *Engagement* in the dialog field is the proportion of direct replies to all questions asked by the other participant, which measures the engagement degree of the chatbot in dialogues.

With all evaluation metrics ready, we collect 20 sessions of dialogue for each of the six models above. We ask two university students to individually score each response uttered by machine in all collected dialogues according to the unified standard, and then ask them to negotiate controversial scores with each other to reach an agreement. The final manual evaluation overall score is the average of all the utter scores by one model. As we can see, our model ranks first among all the three metrics especially for the Engagement, which corresponds to our introduction of role-play control mechanism. What's more, with copy mechanism, CopyNet, DeepCopy, and Initiative-Imitate perform better in terms of fluency and coherence because of the utilizing of proper knowledge. However, comparing CopyNet and DeepCopy with Seq2Seq$_{attn}$, the Engagement becomes worse because too much knowledge harms the ability to react to the proposed question very likely. This emphasizes the importance of appropriate knowledge utilizing again. In the meantime, Generation-base model doesn't score high in terms of fluency and coherence. It is probably relevant to the mechanism of knowledge forcing utilizing in the model, which enlightens us the appropriate knowledge utilizing in the Initiative-Imitate. What's more, with attention, Seq2Seq$_{attn}$ performs much better than the Seq2Seq concerning the Engagement. We think attending over history states on encoder provides more precise information during decoding than the final states of the encoder.

Finally, our Initiative-Imitate is inherited from DeepCopy except for knowledge fusion along with the role-to-play control mechanism. It can be seen that the engagement degree is greatly improved. Meanwhile, the Coherence is also enhanced to some degree. All these before-mentioned reveal the positive effectiveness of proper and appropriate knowledge utilizing together with the role-to-play setting.

4.5 Case Study

Previously we analyzed the strengths and drawbacks of different models at the granularity of gram and session with different common metrics and our newly introduced metric Engagement. Here we will provide a concrete session of dialogue example by Initiative-Imitate shown in Figure 4.

As we can see, the colored knowledge texts are selected for response generation. And each utterance utilizes at least one piece of knowledge. These

knowledge fused in generated responses are appropriate and fit well. Furthermore, when human asks a question about knowledge (texts in red), the Initiative-Imitate seems to understand the question and responds timely and precisely. The Initiative-Imitate recognizes the role of the human and controls the knowledge utilizing in the next sentence generation. Thus it is more engaged in the whole dialog. In the meantime, proper knowledge utilizing also gets rid of repetition to some degree and makes the dialog more coherent, which has been proved during the previous evaluation phase.

It has to be mentioned that our Initiative-Imitate still can not answer all questions completely no matter when and where. As the Engagement score 0.7 shown in Table 2, we humans can easily obtain a score very close to 1 because we care more about what the other has said and will respond after being asked timely and appropriately. While it's not that easy case for a dialog agent. Therefore, the engagement improvement of a dialog agent is an important direction towards real general artificial intelligence in the long term. And we will stick to our initiative improving purpose constantly.

5 Conclusion

In this paper, we first point out the knowledge overusing problem existing in current proactive knowledge dialogue models. Meanwhile, if the bot always plays a single role (proactive or passive) in a dialogue, the other dialog partner loses his/her interest rapidly.

To deal with all these arising issues, we propose to distinguish the dialogue role in the whole dialogue. And for this purpose, we have designed Initiative-Imitate to deal with the knowledge utilizing and role-to-play problem. With the forget gate in the knowledge fusion module, we predict the role to play. Correspondingly, the role-to-play label supervises the forget gate on how much knowledge left at this turn of response generation. In this way, we deal with the two issues simultaneously. As for the dataset, we prepare our dataset with Duconv because of the implicit role setting in the dialog. The automatic evaluation of the Initiative-Imitate on the prepared dataset shows the enhancement of the introduction of the adaptive role setting in human-machine dialog on the gram level. For the human evaluation, we introduce a new metric Engagement to measure the engagement degree of chatbot in dialogues. This metric reflects how much the dialogue agent cares about what the other has said, which can be a vital measurement of the quality of the human-machine conversation. After that, models are evaluated with three metrics including Fluency, Coherence, and Engagement. And the final results prove the positive effectiveness of the role-control to the response generation.

In summary, we take an initial and meaningful step on the role-to-play setting and proper knowledge utilizing. As for future work, we will still insist on working on better role-to-play modeling. What's more, currently many metrics are applied in the result evaluation because they are complementary and can only reflect one aspect of the characteristics of dialogue respectively. So one better and unified metric is in urgent need in the field of automatic conversation evaluation just like the main metric BLEU in machine translation.

References

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Scott Wen-tau Yih, and Michel Galley. 2018a. A knowledge-grounded neural conversation model. In *AAAI*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018b. A knowledge-grounded neural conversation model. In *the 32th AAAI Conference on Artificial Intelligence*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640.

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

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. *2016 Conference of the North American Chapter of the Association for Computational Linguistics*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143.

Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016b. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *the 25th International Joint Conference on Artificial Intelligence*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *the 28th International Joint Conference on Artificial Intelligence*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1577–1586.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Sebastian Varges, Silvia Quarteroni, Giuseppe Riccardi, Alexei Ivanov, and Pierluigi Roberti. 2009. Leveraging pomdps trained with user simulations and rule-based dialogue management in a spoken dialogue system. In *Proceedings of the SIGDIAL 2009 Conference*, pages 156–159.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In *the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3370–3380.

Nick Webb. 2000. Rule-based dialogue management systems. In *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.2854*, pages 164–169.

Joseph Weizenbaum et al. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

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 *the 57th Annual Meeting of the Association for Computational Linguistics*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. 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)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. Deepcopy: Grounded response generation with hierarchical pointer networks. In *the 31th Neural Information Processing Systems*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Junyuan Zheng, Surya Kasturi, Xin Chen Mason Lin, Onkar Salvi, and Harry Jiannan Wang. 2019. The oneconn-memnn system for knowledge-grounded conversation modeling. In *the 33th AAAI Conference on Artificial Intelligence*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *the 27th International Joint Conference on Artificial Intelligence*, pages 4623–4629.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 30–34.