Dialog Generation Using Multi-turn Reasoning Neural Networks

Xianchao Wu[1], **Ander Martínez**[2]*, **Momo Klyen**[1]

[1]A.I.&Research, Microsoft Development Co., Ltd.
[2] Graduate School of Information Science, Nara Institute of Science and Technology
{xiancwu, momokl}@microsoft.com, ander.martinez.zy4@is.naist.jp

Abstract

In this paper, we propose a generalizable dialog generation approach that adapts *multi-turn reasoning*, one recent advancement in the field of document comprehension, to generate responses ("answers") by taking current conversation session context as a "document" and current query as a "question". The major idea is to represent a conversation session into memories upon which attention-based memory reading mechanism can be performed multiple times, so that (1) user's query is properly extended by contextual clues and (2) optimal responses are step-by-step generated. Considering that the speakers of one conversation are not limited to be one, we separate the single memory used for document comprehension into different groups for speaker-specific topic and opinion embedding. Namely, we utilize the queries' memory, the responses' memory, and their unified memory, following the time sequence of the conversation session. Experiments on Japanese 10-sentence (5-round) conversation modeling show impressive results on how multi-turn reasoning can produce more diverse and acceptable responses than state-of-the-art single-turn and non-reasoning baselines.

对文本的分组

根据会话的时间序列使用查询"内存、响应"内存和它们的统一内存

多轮推理可以产生更多样的和可接受的反应，相较于最先进的单轮和非推理基线。

1 Introduction

Dialogue systems such as chatbots are a thriving topic that is attracting increasing attentions from researchers (Sordoni et al., 2015; Serban et al., 2016; Li et al., 2015; Wen et al., 2016). Recent achievements, such as deep neural networks for text generating, user profiling (Li et al., 2014), and natural language understanding, have accelerated the progresses of this field, which was historically approached by conventional rule-based and/or statistical response ranking strategies.

Response ranking models retrieve the most suitable response(s) from a fixed set of (question, answer) pairs given a dialogue context and current query from a user (Banchs and Li, 2012; Lowe et al., 2015). Learning-to-rank approaches were applied to compute the similarity scores of between (query, context) and indexed candidate (question, answer) pairs to return the optimal "answer" to the user. These ranking-based retrieval strategies have been well-applied as an important approach to dialogue systems, yet the set of scripted responses are limited and are short at generalization. On the other hand, statistical machine translation (SMT) systems have been applied to dialogue systems (Ritter et al., 2011), taking user's query as a source language sentence and the chatbot's response as a target language sentence. Labeled data for learning-to-ranking training will not be necessary anymore and all we need is the large-scale (question, answer) pairs.

这些基于排序的检索策略作为对话系统的一种重要方法得到了很好的应用，但是规范化的响应集有限且泛化能力较差。

统计机器翻译已应用于对话系统，将用户的查询作为源语言句，聊天机器人的响应作为目标语言句。

The sequence-to-sequence model proposed in (Sutskever et al., 2014) applied end-to-end training of neural networks to text generation. This model, further enhanced by an attention mechanism (Bahdanau et al., 2014), was generic and allowed its application to numerous sequence-to-sequence learning tasks such as neural machine translation (NMT) (Cho et al., 2014; Bahdanau et al., 2014), image captioning (Donahue et al., 2015; Mao et al., 2015), speech recognition (Chan et al., 2015) and constituency parsing (Vinyals et al., 2015). The simplicity of these models makes them attractive, since "translation" and "alignment" are learned jointly on the fly.

Specially, Vinyals and Le (2015) applied the sequence-to-sequence model to conversational modeling and achieved impressive results on various datasets. Their model was trained to predict a response given the previous sentence (s). Shang et al. (2015) combined local and global attentions

*Work done when Ander was an intern in Microsoft. Wu and Ander contributed equally to this paper.

Proceedings of NAACL-HLT 2018, pages 2049–2059
New Orleans, Louisiana, June 1 - 6, 2018.

and reported better results than retrieval based systems. Sordoni et al. (2015) explored three different end-to-end approaches for the problem of predicting the response given a query attached with a single message context.

多轮对话比机器翻译更难--因为对于给定的输入(上下文、查询)有更多可接受的响应，而这些响应通常依赖于外部知识和/或上下文推理。

Multi-turn conversation modeling is considered to be more difficult than machine translation, since there are many more acceptable responses for a given (context, query) input and these often rely on external knowledge and/or contextual reasoning. Dialogue systems trained with a maximum likelihood estimation (MLE) objective function, as most SMT utilizes, often learn to reply generic sentences as "I don't know" or "sounds good", which have a high incidence in the "answer" part of (question, answer) style training datasets. There have been various attempts at *diversifying* the responses (Li et al., 2016a; Yu et al., 2016; Li et al., 2017) but the lack of variations in the responses remains as an essential challenge. We wonder that if this stress can be relieved by modeling the prior context in a rather fine-grained way.

多样性回答的必要：使用最大似然估计(MLE)目标函数训练的对话系统(大多数SMT都使用这种方法)常常学会回答"我不知道"或"听起来不错"之类的泛型语句，这在(问答)样式训练数据集中的"回答"部分有很高的发生率。

In document comprehension fields, multi-turn reasoning (also called multi-hop reasoning) has delivered impressive results by assimilating various pieces of information to produce an unified answer (Hill et al., 2015; Dhingra et al., 2016). Through multi-turn reading the document's memory using attention models, current question can be extended with much richer knowledge. This makes it easier to figure out the correct answer from that document. Different documents need to be read different times to yield out the correct answer for the input question. Specially, Shen et al. (2016) use a dynamic number of turns by introducing a termination gate to control the number of iterations of reading and reasoning.

通过使用注意模型多轮读取文档的记忆，可以扩展当前问题的知识面。这使得从该文档中找出正确答案变得更容易。

Motivated by the reasoning network for document comprehension (Shen et al., 2016), we propose multi-turn reasoning neural networks that generate the proper response (or, "answer") by attention-based reasoning from current conversation session ("document") and current query (identical to "question" in document comprehension) from the user. In particular, our networks utilize conversation context and explicitly separate speakers' interventions into sentence-level and conversation-level memories. Our first model uses plain single-turn attention to integrate all the memories, and the second approach integrates multi-turn reasoning. The formulation of our proposed approach is designed in a generalized way, allowing for inclusion of additional information such as external knowledge bases (Yih and Ma, 2016; Ghazvininejad et al., 2017; Han et al., 2015) or emotional memories (Zhou et al., 2017). Moreover, our approach for two-speaker scenario can be easily extended to group chatting by a further speaker-specific memory splitting.

We evaluate the performances of our methods by comparing three configurations trained on a Japanese twitter conversation session dataset. Each conversation session contains 10 sentences which are 5-round between two real-world speakers. The results provide evidences that multi-turn reasoning neural networks can help improving the consistency and diversity of multi-turn conversation modeling.

数据集包含5论对话的10个句子

提高多轮对话的一致性和多样性

This paper is structured as follows: Section 2 gives a general description of multi-turn conversation modeling; Section 3 describes background neural language modeling, text generation, and attention mechanisms; Section 4.1 first introduces a model with multiple attention modules and then explains how the multi-turn reasoning mechanism can be further integrated into the previous models; Sections 5, 6 and 7 describe the experimental settings and results using automatic evaluation metrics, detailed human-evaluation based analysis, and conclusions, respectively.

2 Multi-turn Conversation Modeling

Consider a dataset $\mathcal{D}$ consisting of a list of conversations between two speakers. Each conversation $d \in \mathcal{D}$ is an ordered sequence of sentences s_i, where $i \in [1, T_d]$ and T_d is the number of sentences in d, produced by two speakers alternately. In this way, for $s_i, s_j \in d$, both sentences are from the same speaker if and only if $i \equiv j \pmod 2$. Note that, our definition includes the case that one speaker continuously expresses his/her message through several sentences. We simply concatenate these sentences into one to ensure that the conversation is modeled with alternate speakers.

我们的定义包括这样一种情况：一个说话者通过几句话不断地表达他/她的信息，作法是简单的拼接语句

A multi-turn conversation model is trained to search parameters that maximize the likelihood of every sentence $s_i \in d$ where $i \geq 2$, supposing that the beginning sentence s_1 is always given as a precondition:

$$\theta_M = \arg\max_{\theta} \{\mathcal{L}(\theta, \mathcal{D})\}, \quad (1)$$

where

$$\mathcal{L}(\theta, \mathcal{D}) = \sum_{d \in \mathcal{D}} \prod_{i=2}^{T_d} p(s_i | s_{<i}). \quad (2)$$

Here, $s_{<i}$ are sentences $s_j \in d$ and $j < i$.

与SMT和单轮对话区别：下一个句子的生成对会话不敏感，只由前一个句子决定，而多轮对话依靠前面的句子相当于依句子为单位的语言建模

The probability of each sentence, $p(s_i|s_{<i})$, is frequently estimated by a conditional language model. Note that, traditional single-turn conversation models or NMT models are a special case of this model by simply setting T_d to be 2. That is, the generation of the next sentence is session-insensitive and is only determined by the former single sentence. Another aspect of understanding this contextual conversation model is that, the number of reference contextual sentences $s_{<i}$ is not limited to be one. Suppose there are already 9 sentences known in one conversation session and we want to generate the 10-th sentence, then from $p(s_1)$ to $p(s_9)$ are all preconditions and we will only need to focus on estimating $p(s_{10}|s_{<10})$.

We adapt sequence-to-sequence neural models (Sutskever et al., 2014) for multi-turn conversation modeling. They are separated into an encoder part and a decoder part. The encoder part applies a RNN on the input sequence(s) $s_{<i}$ to yield prior information. The decoder part estimates the probability of the generated output sequence s_i by employing the last hidden state of the encoder as the initial hidden state of the decoder. Sutskever et al. (2014) applied this technique to NMT and impressive experimental results were reported thereafter.

但是，我们可以添加一些约束来确定生成谁的响应，是一个演讲者还是两个演讲者

Using Equation 2, we are modeling two chatbots talking with each other, since all the $s_{2,\ldots,T_d}$ are modeled step-by-step on the fly. However, we can add constraints to determine whose responses to be generated, either one speaker or both of them. That is, when i takes odd integers of 1, 3, 5 and so on, we are modeling the first speaker. Even integers of i indicates a generation of responses for the second speaker.

3 Language Modeling and Text Generation

Language models (LM) are trained to compute the probability of a sequence of tokens (words or characters or other linguistic units) being a linguistical sentence. Frequently, the probability of a sentence s with T_s tokens is computed by the production of the probabilities of each token $y_j \in s$ given its contexts $y_{<j}$ and $y_{>j}$:

$$p(s) = \prod_{j=1}^{T_s} p(y_j | y_{<j}, y_{>j}). \quad (3)$$

When generating a sequence based on a LM, we can generate one word at a time based on the previously predicted words. In this situation, only the previously predicted words are known and the probability of current sequence is approximated by dropping the posterior context. That is,

$$p(y_j | y_{<j}, y_{>j}) \approx p(y_j | y_{<j}). \quad (4)$$

We construct a sequence generation LM using sequence-to-sequence neural network f. The neural network intercalates linear combinations and non-linear activate functions to estimate the probability of mass function. Then, in the encoder part of f, the contextual information is represented by a fixed-size hidden vector h_j:

$$p(y_j | y_{<j}, y_{>j}) \approx f(y_{j-1}, h_j, \theta_f), \quad (5)$$

where θ_f represents f's trainable parameters.

To embed the previous word sequence into a fixed-size vector, recurrent neural networks (RNN) such as long short term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014) are widely used. These networks repeat a recurrent operation on each input word:

$$h_j = g(h_{j-1}, y_{j-1}, \theta_g), \quad (6)$$

where θ_g represents the trainable parameters of a RNN function g, and h_j is the hidden state of the RNN at time j.

3.1 Conditional Language Modeling

The hidden state h of a RNN can accumulate information from previous words ($y_j \in s$ and $j < T_s$) or previous sentences ($s_i \in d$ and $i < T_d$) which ensures the encoding and decoding processes in sequence-to-sequence models. Since the contextual sentences are known already, the encoder can represent them in both forward ($\overrightarrow{h}_j$) and backward ($\overleftarrow{h}_j$) directions. The results from both recursions can be combined by a concatenated operation. This is referred to as bidirectional RNN shorted as BiRNN (Schuster and Paliwal, 1997).

$$h_j = [\overrightarrow{h}_j^\top; \overleftarrow{h}_j^\top]^\top. \quad (7)$$

For each sentence $s_i \in d$ ($i < T_d$), we annotate the combination of the final states of each RNN direction as a memory vector $m_i = [(\overrightarrow{h}^{(i)}_{T_{s_i}})^{\mathrm{T}}; (\overleftarrow{h}^{(i)}_1)^{\mathrm{T}}]^{\mathrm{T}}$. A projection of annotation m_i can be used as the decoder's initial state t_0 such as $t_0 = \tanh(W_s m_{T_{d'}})$ and $T_{d'} < T_d$. W_s here is a weight matrix that projects $m_{T_{d'}}$ into a vector that shares a same dimension with t_0. In (Bahdanau et al., 2014) for NMT, $\overleftarrow{h}_1$, backward encoding of a single source sentence, was used to initialize $t_0 = \tanh(W_s \overleftarrow{h}_1)$.

3.2 Attention Mechanism

使用注意力机制的原因：随着上下文信息的增长，将所有的上下文信息汇总成一个固定长度的向量将变得越来越弱，从而无法指导目标句子的生成

Summarizing all contextual information into one single fixed-length vector becomes weaker to guide the generation of the target sentence as the contextual information grows longer. To tackle this problem, an attention mechanism (Bahdanau et al., 2014) was applied to NMT for learning to align and translate jointly. In this attention-based model, the conditional probability in Equation 4 is defined as:

$$p(y_j|y_{<j}) = f(y_{j-1}, t_j, c_j, \theta_f), \tag{8}$$

where

$$t_j = g(t_{j-1}, y_{j-1}, c_j, \theta_g) \tag{9}$$

is a RNN hidden state in the decoder part for time j and $c_j = \sum_{i=1}^{T_s} \alpha_i^{(j)} h_i$ is a *context vector*, a weighted combination of the annotation set memory $(h_1, ..., h_{T_s})$ produced by encoding a source sentence s with length T_s. The weight $\alpha_i^{(j)}$ of each (source) annotation h_i is computed by

$$\alpha_i^{(j)} = \frac{\exp(e_i^{(j)})}{\sum_{l=1}^{T_s} \exp(e_l^{(j)})}. \tag{10}$$

where $e_i^{(j)}$ is an *alignment model* and is implemented by a feed-forward network a:

$$e_i^{(j)} = a(h_i, t_{j-1}). \tag{11}$$

This method was applied to single-turn conversation modeling (Vinyals and Le, 2015). We use this model, with attention over each immediately previous sentence $s_{i-1} \in d$ for generating s_i, as a baseline for our experiments. We annotate this model as SIMPLE subsequently in this paper.

4 Multi-turn Reasoning Network with Multiple Type Memories

The attention mechanism described in Equations 8 and 9 is performed in a single-turn feed-forward fashion. However, for complex context and complex queries, human readers often revisit the given context in order to perform deeper inference after one turn reading. This real-world reading phenomenon motivated the multi-turn reasoning networks for document comprehension (Shen et al., 2016). Considering dialog generation scenario with given rich context, we intuitively think if the attentions can be performed multi-turns so that the conversation session is better understood and the simple query, which frequently omits unknown number of context-sensitive words, can be extended for a better generation of the response. The domain adaptation from document comprehension to dialog generation is feasible by taking the rich context of the speakers as a "document", current user's query as a "question" and the chatbot's response as an "answer".

走向多跳的原因：对于复杂的上下文和复杂的查询，人类读者通常会重新访问给定的上下文，以便在一次阅读之后执行更深入的推理。

直觉的想法：如果注意事项可以多轮执行，以便更好地理解会话，并且可以扩展简单查询，以便更好地生成响应，而简单查询常常会省略未知数量的上下文敏感单词。

However, there are still several major challenges for this domain adaptation. First, a document is frequently written by a single author with one (hidden) personality, one writing style, and one distribution of the engagement rates of the topics appearing in that document. These are not the case for conversation scenario in which at least two speakers are involved with different (hidden) personalities, personalized speaking styles, and diverse engagement rate distributions of the topics in that conversation session. Second, for document comprehension, the output is frequently a single named entity (Shen et al., 2016) and thus a single softmax function can satisfy this one-shot ranking problem. However, we will need a RNN decoder utilizing context vectors for generating the target response sentence being a sequence of tokens (words or characters) instead of one single named entity.

文本理解中使用的多轮对话不适合用于对话：(1)文本是一个作者一种写作风格，者不适用于至少有两个说话者具有不同(隐藏的)个性、个性化的说话风格以及会话中主题的不同参与率分布的会话场景。(2)对于文档理解，输出通常是单个命名实体，因此一个softmax函数可以满足这个一次性排序问题。然而，我们将需要一个RNN解码器，它利用上下文向量生成目标响应语句，使其成为标记序列(单词或字符)，而不是单个命名实体。

We tackle the first challenge by separating the context into multiple type memories upon which attention models are performed. For the second difference, we replace the simple softmax output layer by a GRU decoder employing reasoning-attention context vectors.

针对于上述两个问题的解决方法：对于(1)将上下文分隔为多个类型的记忆，在这些类型的记忆上执行注意力模型。对于(2)我们用GRU解码器替换了简单的softmax输出层，该解码器使用了推理注意上下文向量

4.1 Separation of contextual information

The SIMPLE model can use multiple turns of context to infer the response by concatenating them

during decoding, using a separator symbol such as EOS for end-of-sentence. Sordoni et al. (2015) separated the query message and the previous two context messages when conditioning the response. The previous context messages were concatenated and treated as a single message.

In our proposed models, we use more than three turns for the context. We separate the last message (the query) from the previous turns to produce a set of annotations h, one per character[1] in the sentence. While encoding the contextual information, we separate the m_i from each speaker into two sets. The motivation is to capture individual characteristics such as personalized topical information and speaking style (Li et al., 2016b). We refer to the set of annotations from the same speaker as one memory. That is, the sentences for which the probabilities are being predicted as M_r (response memory, specially corresponds to the chatbot's side) and the question set as M_q (query memory, specially corresponds to the user's side). We further apply a RNN on top of m_i to produce one more set of vectors M_c (context memory):

$$M_c = \bigcup_{i=0}^{T_c} \{m_i^{(c)}\}, \tag{12}$$

in which,

$$m_i^{(c)} = \text{RNN}(m_i, m_{i-1}^{(c)}), \tag{13}$$

where T_c is the number of turns (sentences) in the conversation. The initial state $m_0^{(c)}$ is a trainable parameter.

We apply an attention mechanism on each of the memories M_q, M_r, M_c and M_h (of current query) separately. Refer to Figure 1 for an intuitive illustration of these memories. Following (Shen et al., 2016), we choose projected cosine similarity function as the attention module. The attention score $a_{j,i}^q$ on memory $m_i^q \in M_q$ for a RNN hidden state t_j in the decoder part is computed as follows:

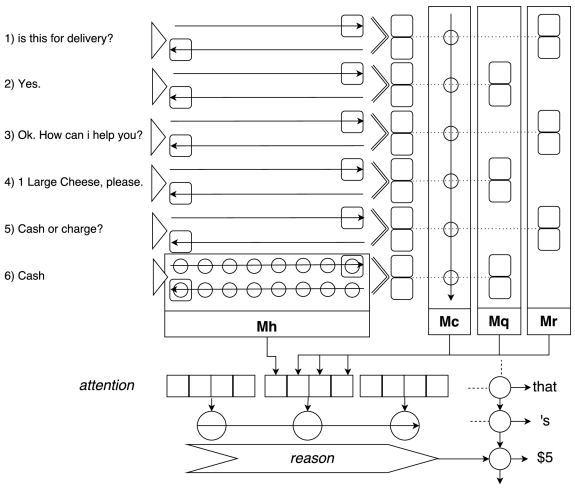
$$a_{j,i}^q = \text{softmax}_{i=1,\ldots,|M_q|} \cos(W_1^q m_i^q, W_2^q t_j), \tag{14}$$

where W_1^q and W_2^q are weight vectors associated with m_i^q and t_j, respectively. Consequently, the attention vector on the query sequences is given by:

$$c_j^q = \sum_{i=1}^{|M_q|} a_{j,i}^q m_i^q. \tag{15}$$

[1]Most Japanese characters, such as Kanji, have independent semantic meanings other than English letters

在对上下文信息进行编码时，我们将每个说话者的mi分成两组。动机是捕捉个人特征，如个性化的话题信息和说话风格



m_h的产生是输入一句话每一步的隐状态

m_c的产生是BiRNN最后的隐状态

m_q是用户的查询的隐状态

m_r是机器的回复隐状态

注意力分别作用于这四个产生四个上下文向量 然后四个拼接得到最终的上下文向量

然后解码器的隐状态和上下文的隐状态以及先前的预测值词向量一起解码出当前生成词

Figure 1: Illustration of the architecture of our REASON model. In the example conversation session, six context sentences are encoded independently by a biRNN. The character-specific annotations from the sixth sentence (i.e., current query) compose the only *sentence-level* M_h memory. $M_{r,q,c}$ are *conversation-level* memories. The last m_i from each sentence are distributed alternately in M_q and M_r memories. Also, $m_{1,\ldots,6}$ are iterated sequentially by a single-direction RNN to produce six context annotations in M_c using Equation 13. In the bottom of the diagram, a reasoning module (Figure 2) is used instead of the plain attention used in MULTI.

每句话的最后一个mi交替分布在Mq和Mr内存中。

Similarly, the attention scores and attention vectors on the other three memories can be derived by replacing q with r, c, and h in Equations 14, 15.

We then concatenate these resulting attention vectors into a final context vector c_j^M, which is consequently applied to Equations 8 and 9. Since the dimension of the updated context vector c_j^M is four times larger, its weight matrix C will need to be enlarged with a same column dimension with the dimension of c_j^M so that Cc_j^M still aligns with the dimension of the hidden layer vector t_j. More details of the GRU function style definition of t_j using c_j can be found in (Bahdanau et al., 2015). We refer to this model that integrates multiple types of memories through separated attention mechanisms as MULTI.

Note that, by separately embedding conversation context into multiple type memories following the number of speakers, we can easily extend this two speaker scenario into group chatting in which tens or hundreds of speakers can be engaged in. The only extension is to further separate M_q by speakers. Consequently, the context vector can be concatenated using the attention vectors by read-

模型的训练流程：第一句为机器人的轮次，先经过encoder得到h_1，存在M_h_1和M_r_1中，然后用h_1作为输入经过又一RNN得到M_c_1，然后decoder的初始状态t_0分别和M_h_1、M_r_1、M_c_1以及M_q_1(初始为0)做attention得到四个context vector拼接，作为REASON中RNN的输入，输出来判断是否停止迭代，若为停止迭代输出拼接好的context vector，解码器继续往下进行。

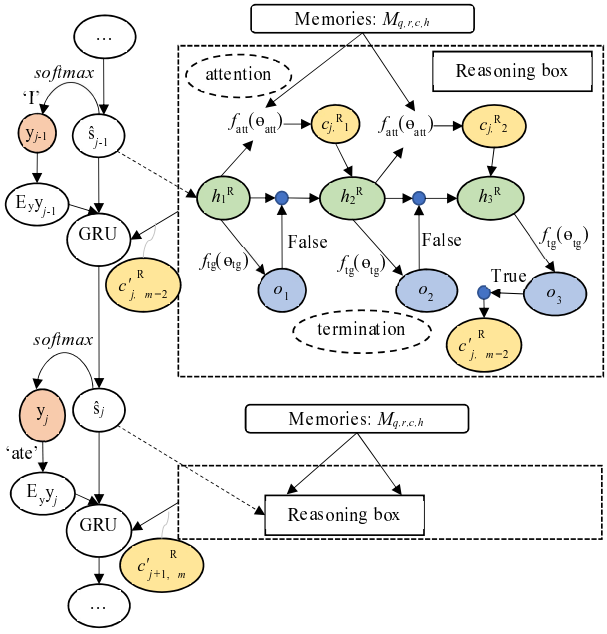


Figure 2: Illustration of the *reason* part of the REASON model. This figure is drawn by partially referring Figure 1 in (Shen et al., 2016).

理论上的好处是，聊天机器人可以流畅地跟踪每个说话者的话题，然后决定如何回应说话者。

ing all the memories. The theoretical benefit is that the chatbot can softly keep track of each individual speaker's topics and then make a decision of how to response to that speaker. Another extension will be using a reinforcement learning policy to determine when to let the chatbot to give a response to which speaker in current group chatting.

另一个扩展将使用强化学习策略来确定何时让聊天机器人对当前群组中的说话者做出响应

Generally, the number and type of memories can be enlarged in a reasonable way, such as by introducing external knowledge (Yih and Ma, 2016; Ghazvininejad et al., 2017; Han et al., 2015) or performing sentiment analysis to the "fact memories" to yield emotional memories (Zhou et al., 2017). A detailed description and experimental testifying is out of the scope of this paper.

记忆的数量和类型可以通过合理的方式扩大，如引入外部知识或对"事实记忆"进行情绪分析，产生情感记忆

4.2 Reasoning Neural Dialogue System

As illustrated in Figure 2, we apply a multi-turn reasoning mechanism, following Shen et al. (2016), to the multiple-type annotation memories. This reasoning mechanism replaces the single-turn attention mechanism. We adapt the idea of using a termination state during the inference to dynamically determine how many turns to reason. The termination module can decide whether to continue to infer the next turn (of re-reading the four types of memories) after digesting intermediate topical and speaker-specific information, or to terminate the whole inference process when it concludes that existing information is sufficient to generate the next word in a response. Generally, the idea is to construct a reasoning attention vector that works as a context vector during generating the next word. This idea is included in the "Reasoning box" in Figure 2. Specially, y_{j-1} stands for a former word generated by the hidden state $\hat{s}_{j-1}$ in the GRU decoder. E_y is the embedding matrix. We use a full-connection layer to map from $\hat{s}_{j-1}$ to the initial reasoning hidden state h_1^{R}, since h_m^{R} should be with the same length alike each memory vector in $M_{q,r,c,h}$ and $\hat{s}_{j-1}$'s dimension is smaller than that. Thus, (1) outside the "reasoning box", we use a GRU decoder to yield $\hat{s}_j$ so that a next word y_j can be generated, and (2) inside the "reasoning box", we read the memories to yield the "optimal" contextual vector. The "reasoning box" takes the memories $M_{q,r,c,h}$ and $\hat{s}_{j-1}$ as inputs and finally outputs $c_{j,m}^{\mathrm{R}}$.

The number of reasoning turns for yielding the "reasoning attention vectors" (c_j^{R} which is further indexed by reasoning steps of 1, 2 in Figure 2) during the decoding inference is dynamically parameterized by both the contextual memories and current query, and is generally related to the complexities of the conversation context and current query.

The training steps are performed as per the general framework as described in Equations 8 and 9. For each reasoning hidden state h_m^{R}, the termination probability o_m is estimated by $f_{tg}(h_m^{\mathrm{R}}; \theta_{tg})$, which is

$$o_m = (1 - o_{m-1}) * \sigma(w_t^\top h_m^{\mathrm{R}} + b_t), \qquad (16)$$

where $\theta_{tg} = \{w_t, b_t\}$, w_t is a weight vector, b_t is a bias, and σ is the sigmoid logistic function. Then, different hidden states h_m^{R} are first weighted by their termination probabilities o_m and then summed to produce a reasoning-attention context vector c_j^{R} (using the equations as described previously in Section 3.2), which is consequently used to construct the next reasoning step's h_2^{R} = RNN(h_1^{R}, $c_{j,1}^{\mathrm{R}}$). The final $c_{j,m}^{\mathrm{R}}$ ($m \geq 1$ is the final reasoning step) will be used in Equations 8 and 9 in a way alike former attention vectors. During our experiments, we instead used a sum of from $o_2 \times c_{j,1}^{\mathrm{R}}$ to $o_{m+1} \times c_{j,m}^{\mathrm{R}}$ as the final $c_{j,m}^{\prime\mathrm{R}}$ for next word generation.

采用在推理过程中使用终止状态的思想来动态地确定有多少次转向推理

终止模块可以在消化了中间话题和特定说话者的信息后决定是否继续推断下一步(重读这，或者当推断出现有信息足以生成响应中的下一个单词时，终止整个推理过程。四种类型的记忆)

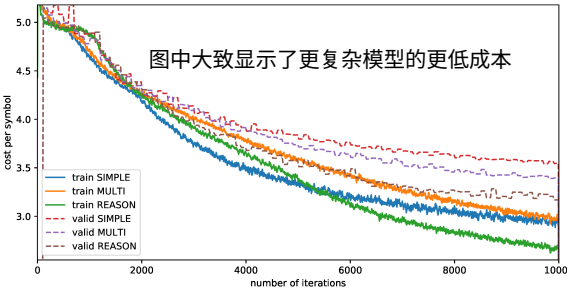
During generating each word in the response, our network performs a response action r_m at the m-th step, which implies that the termination gate variables $o_{1:m} = (o_1 = 0, o_2 =$

```
A: i'm bored
B: boooring
A: isn't it? (^_^)
B: everybody is asleep
A: really?
   (^_^) my friends are still awake! :P
B: lucky!
   xD feels lonely with everyone asleep
A: almost everyone is awake (^v^)
B: whyyy?!
   my friends go early to bed.
   Or is it me that's late? xD
A: it's us who are late! xD
B: true, very true xD
```

Figure 3: A 5-round conversation of speakers A and B.

	Conversation sessions	Sentences	Characters (Unique)
Train	253K	2.5M	24M (6,214)
Validation	100	1K	10K (836)
Test	100	1K	9.3K (780)

Table 1: Statistics of the filtered datasets.



结论：
(1)在迭代10000时，SIMPLE和MULTI的最终训练损失非常接近；
(2)最后的训练损失之间有很大的差额REASON与SIMPLE或REASON与MULTI；
(3)验证损失完全遵循以下顺序SIMPLE>MULTI>RESON

Figure 4: Cost curves of NLL during training.

$0, ..., o_{m-1} = 0, o_m = 1$). A stochastic policy $\pi((o_m, r_m)|h^{\mathrm{R}}_m, t_j; \theta)$ with parameters θ to get a distribution of termination actions, to continue reading the conversation context (i.e., $M_{q,r,c,h}$) or to stop, and of response actions r_m for predicting the next word if the model decides to stop at current step. In our experiments, we set a maximum step parameter T_{max} to be 5 for heuristically avoiding too many reasoning times. We follow (Shen et al., 2016) to compute the expected reward and its gradient for one instance. We refer to this model with multi-turn reasoning attentions as REASON.

5 Experiments

日本twitter对话。每段对话包含10个句子，来自两个真实世界的交替说话者。

In our experiments, we used a dataset consisting of Japanese twitter conversations. Each conversation contains 10 sentences from two real-world alternating speakers. Given the origin of the dataset, it is quite noisy, containing misspelled words, slang and *kaomoji* (multi-character sequences of facial emoticons) among meaningful words and characters. Preliminary experiments by using a word-based approach resulted in the vocabulary size being too big and with too many word breaking errors, we instead used a character-based approach. Figure 3 shows a sample 10-sentence conversation in which original Japanese sentences were translated into English and similar spelling patterns were kept in a sense (such as boooring for boring and whyyy for why).

为何要进行字符方法：通过使用基于单词的方法进行初步实验，我们发现词汇量过大，出现了太多的断词错误，因此我们采用了基于字符的方法

每一句不超过20个字符

We kept the conversations in which all sentences were no more than 20 characters. This filtering strategy resulted in a dataset of 254K conversations from which 100 (1K sentences) where taken out for testing and another 100 for validating and hyper-parameter tuning. The training set contains 6,214 unique characters, which are used as our vocabulary with the addition of two special symbols, an UNK (out-of-vocabulary unknown word) and an EOS (end-of-sentence). Table 1 shows major statistics of the dataset.

The training minimizes negative log-likelihood (NLL) per character on the nine sentences $s_{2,...,10}$ of each conversation. One configuration in MULTI and REASON is that, we respectively use the reference contexts (instead of former automatically generated sentences) to generate current sentence. That is, when generating s_i, we use the golden contextual sentences of from s_1 to s_{i-1}. These three systems were respectively trained 3 epochs (10,000 iterations) on an AdaDelta (Zeiler, 2012) optimizer. Character embedding matrix was shared by both the encoder and the decoder parts. All the hidden layers, in the encoding/decoding parts and the attention models, were of size 200 and the character embeddings were of size 100. The recurrent units that we used were GRU. The gradients were clipped at the maximum gradient norm of 1. The reasoning module's maximum steps $T_{\max}$ was set to be 5. The data was iterated on mini-batches of less than 1,500 symbols each.

训练细节，训练次数，训练步长，优化器

梯度截断

We initialized the recurrent weight matrices in GRUs as random orthogonal matrices. Unless specially mentioned, all the elements of the 0-indexed vectors and all bias vectors were initialized to be zero. Any other weight matrices were initialized by sampling from the Gaussian distribution of mean 0 and variance 0.01.

参数初始化细节

Figure 4 shows the progression of the NLLs per

		SIMPLE	MULTI	REASON
BLEU-4	Train	1.98	1.97	**2.30**
	Validation	1.80	2.12	**2.62**
	Test	2.20	2.13	**2.89**
BLEU-2	Train	6.77	6.78	**7.03**
	Validation	6.67	6.89	**8.14**
	Test	7.19	7.24	**7.97**

Table 2: Character-level BLEU-2/4 (%) scores.

character during training. The validation costs begun converging in the third epoch for the three models. The plot roughly shows lower cost for more complex models.

当使用BLEU-2而不是BLEU-4时，与人类评价的相关性更好。

Galley et al. (2015) obtained better correlation with human evaluation when using BLEU-2 rather than BLEU-4. We thus report both of these scores for automatic evaluation and comparison. The character-level BLEU-4 and BLEU-2 scores for the trained models are reported in Table 2. The REASON model achieved consistently better BLEU-2 and BLEU-4 scores in the three datasets. MULTI performed slightly better than SIMPLE on the validation set yet that performance is less stable than REASON.

在图4中

Figure 4 also reflects that, (1) the final training costs of SIMPLE and MULTI are quite close with each other at iteration 10,000; (2) there is a big margin of between the final training cost of REASON and that of SIMPLE or MULTI; and (3) the validation costs exactly follows an order of SIMPLE > MULTI > REASON.

6 Analysis

图5展示了应用中的一个实例，MULTI效果不好

Figure 5 illustrates an English translation of a conversation and the responses suggested by each of the described models. This conversation is extracted from the test set. The three responses are different from the reference response, but the one from REASON looks the most consistent with the given context. The response from MULTI is contradicting the context of speaker B as he/she said *Not at all* in a former sentence.

BLEU与人类的判断不太相关

As it has been shown in (Liu et al., 2016) that BLEU doesn't correlate well with human judgments, we asked three human evaluators to respectively examine 900 responses from each of the models given their reference contexts. The evaluators were asked to judge (1) whether one response is acceptable and (2) whether one response is better than the other two responses. A summary of this evaluation is displayed in Table 3. The *acceptable* column refers to the percentage of responses

用三个人来评估，标准是：(1)一种反应是否可以接受；(2)一种反应是否优于另外两种反应。

A: I feel nostalgic. This was so cute. B: It's gross. What's that picture? A: It's cute! It used to be on TV. B: Isn't that from a children's show? A: Yes B: I know that one!! A: Isn't it cute? B: Not at all A: Uh! I can't believe it...
SIMPLE (B:) Uh! it isn't. / MULTI (B:) It's cute! / REASON (B:) Do you think it's cute?
Reference B: Are you asking me whether the picture is cute?

Figure 5: Sample responses generated by the three models and the recorded reference response, translated to English. Both MULTI and REASON include *cute*, yet MULTI contradicts a previous response *Not at all*.

最后两列进行一对一的比较

	acceptable	best-of-three †	> SIMPLE‡	> MULTI‡
SIMPLE	42%	25%	-	45%
MULTI	52%	29%	55%	-
REASON	**65%**	**46%**	**59%**	**58%**

Table 3: Human evaluation of the responses generated by the three models. † Percentage over the conversations that had at least one response accepted. ‡ From the cases where any of both compared models was acceptable. > = "better than".

that were considered acceptable by at least two of the human evaluators while the *best-of-three* columns refers to the percentage of times that each model's response was considered by at least two evaluators to be better than the other two's, from the contexts that had at least one acceptable response. The last two columns make one-to-one comparisons. In 18% of the contexts, none of the models produced an acceptable response.

REASON的多跳attention的有效性

This human evaluation shows that complexer models are more likely to produce acceptable responses. The MULTI and REASON models are only different in the attention mechanism of multi-turn reasoning. The reasoning module performed better than single-turn attention 58% of the times.

Table 4 contains the character-level *distinct-n* (Li et al., 2016a) metrics for n-grams where $1 \leq n \leq 5$. This metric measures the number of distinct n-grams divided by the total number of n-grams in the generated responses. The displayed results are computed on the concatenation of all the responses to the test-set contexts. The *Reference* column was computed on the reference responses and represents the optimal human-like ratio.

SIMPLE performed the best at uni-gram diver-

	SIMPLE	MULTI	REASON	Reference
distinct-1	**.039**	.028	.032	.088
distinct-2	.112	.095	**.121**	.407
distinct-3	.199	.180	**.238**	.588
distinct-4	.248	.241	**.310**	.587
distinct-5	.255	.265	**.328**	.530

Table 4: N-gram diversity metrics of between (1) the responses generated to the test set and (2) their reference responses.

sity. For n-grams $n \geq 2$, REASON produced the most diverse outputs. While the results for REASON were consistently better than the other two models, the results for MULTI were not always better than SIMPLE. This indicates MULTI does not always benefit from the augmented context without the multi-turn reasoning attentions.

这表明，如果不注意多轮推理，MULTI并不总是能从增强上下文中获益

7 Conclusions

We have presented a novel approach to multi-turn conversation modeling. Our approach uses multiple explicitly separated memories to represent rich conversational contexts. We also presented multi-turn reasoning attentions to integrate various annotation memories. We run experiments on three different models with and without the introduced approaches and measured their performances using automatic metrics and human evaluation.

Experimental results verified that the increased contexts are able to help producing more acceptable and diverse responses. Driven by the depth of the reasoning attention, the diversities of the responses are significantly improved. We argue that the reasoning attention mechanism helps integrating the multiple pieces of information as it can combine them in a more complex way than a simple weighted sum. We further observed that as the accuracy of the conversation model improves, the diversity of the generated responses increases.

The proposed approach of multi-turn reasoning over multiple memory attention networks is presented in a general framework that allows the inclusion of memories of multiple resources and types. Applying to group chatting with more than two speakers and reasoning over emotion embeddings or knowledge vectors included from an external knowledge base/graph are taken as our future directions.

Acknowledgments

The authors thank the anonymous reviewers for their impressive comments and suggestions for improving the earlier version.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]* ArXiv: 1409.0473. http://arxiv.org/abs/1409.0473.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2015. End-to-End Attention-based Large Vocabulary Speech Recognition. *arXiv:1508.04395 [cs]* ArXiv: 1508.04395. http://arxiv.org/abs/1508.04395.

Rafael E. Banchs and Haizhou Li. 2012. IRIS: a Chat-oriented Dialogue System based on the Vector Space Model. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, Jeju Island, Korea, pages 37–42. http://www.aclweb.org/anthology/P12-3007.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. Listen, Attend and Spell. *arXiv:1508.01211 [cs, stat]* ArXiv: 1508.01211. http://arxiv.org/abs/1508.01211.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]* ArXiv: 1406.1078. http://arxiv.org/abs/1406.1078.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2016. Gated-Attention Readers for Text Comprehension. *arXiv:1606.01549 [cs]* ArXiv: 1606.01549. http://arxiv.org/abs/1606.01549.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 445–450. http://www.aclweb.org/anthology/P15-2073.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *CoRR* abs/1702.01932. http://arxiv.org/abs/1702.01932.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, pages 129–133. http://aclweb.org/anthology/W15-4616.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]* ArXiv: 1511.02301. http://arxiv.org/abs/1511.02301.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A Diversity-Promoting Objective Function for Neural Conversation Models. *arXiv:1510.03055 [cs]* ArXiv: 1510.03055. http://arxiv.org/abs/1510.03055.

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*. Association for Computational Linguistics, San Diego, California, pages 110–119. http://www.aclweb.org/anthology/N16-1014.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 994–1003. http://www.aclweb.org/anthology/P16-1094.

Jiwei Li, Will Monroe, Tianlin Shi, Sbastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. *arXiv:1701.06547 [cs]* ArXiv: 1701.06547. http://arxiv.org/abs/1701.06547.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 165–174. http://www.aclweb.org/anthology/P14-1016.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2122–2132. https://aclweb.org/anthology/D16-1230.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *arXiv:1506.08909 [cs]* ArXiv: 1506.08909. http://arxiv.org/abs/1506.08909.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *ICLR* .

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 583–593. http://www.aclweb.org/anthology/D11-1054.

M. Schuster and K.K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.* 45(11):2673–2681. https://doi.org/10.1109/78.650093.

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*. AAAI Press, Phoenix, Arizona, AAAI'16, pages 3776–3783. http://dl.acm.org/citation.cfm?id=3016387.3016435.

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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1577–1586. http://www.aclweb.org/anthology/P15-1152.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2016. ReasoNet: Learning to Stop Reading in Machine Comprehension. *arXiv:1609.05284 [cs]* ArXiv: 1609.05284. https://doi.org/10.1145/3097983.3098177.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015.

A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 196–205. http://www.aclweb.org/anthology/N15-1020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 3104–3112.

Oriol Vinyals, ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a Foreign Language. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 2773–2781. http://papers.nips.cc/paper/5635-grammar-as-a-foreign-language.pdf.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv:1506.05869 [cs]* ArXiv: 1506.05869. http://arxiv.org/abs/1506.05869.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain Neural Network Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 120–129. http://www.aclweb.org/anthology/N16-1015.

Wen-tau Yih and Hao Ma. 2016. Question Answering with Knowledge Base, Web and Beyond. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, San Diego, California, pages 8–10. http://www.aclweb.org/anthology/N16-4003.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2016. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *arXiv:1609.05473 [cs]* ArXiv: 1609.05473. http://arxiv.org/abs/1609.05473.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv:1212.5701 [cs]* ArXiv: 1212.5701. http://arxiv.org/abs/1212.5701.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *CoRR* abs/1704.01074. http://arxiv.org/abs/1704.01074.