Multi-view Response Selection for Human-Computer Conversation

Xiangyang Zhou[1]*, **Daxiang Dong**[1]*, **Hua Wu**[1], **Shiqi Zhao**[1],
Dianhai Yu[1,2], **Hao Tian**[1,2], **Xuan Liu**[1] and **Rui Yan**[1]
[1]Baidu Inc., Beijing, China
[2]School of Information Science and Technology,
University of Science and Technology of China
{zhouxiangyang, dongdaxiang, wu_hua, zhaoshiqi,
yudianhai, tianhao, liuxuan, yanrui}@baidu.com

Abstract

In this paper, we study the task of response selection for multi-turn human-computer conversation. Previous approaches take word as a unit and view context and response as sequences of words. This kind of approaches do not explicitly take each utterance as a unit, therefore it is difficult to catch utterance-level discourse information and dependencies. In this paper, we propose a multi-view response selection model that integrates information from two different views, i.e., word sequence view and utterance sequence view. We jointly model the two views via deep neural networks. Experimental results on a public corpus for context-sensitive response selection demonstrate the effectiveness of the proposed multi-view model, which significantly outperforms other single-view baselines.

以前的方法将单词作为一个单元，将上下文和响应看作单词序列。从词级入手

这种方法没有明确地将每个话语作为一个单元，因此很难捕捉到话语层面的信息和依赖关系

提出一个多视图响应选择模型，该模型集成了来自两个不同视图的信息，即，单词序列视图和话语序列视图.通过神经网络来建模

1 Introduction

任务的挑战以及意义

从一组候选对象中选择一个潜在的响应是开放域人机对话，特别是基于检索的人机对话的一项重要而具有挑战性的任务。通常，检索索引的对话语料库中的一组候选响应，然后从候选响应中选择最佳响应作为系统的响应

Selecting a potential response from a set of candidates is an important and challenging task for open-domain human-computer conversation, especially for the retrieval-based human-computer conversation. In general, a set of candidate responses from the indexed conversation corpus are retrieved, and then the best one is selected from the candidates as the system's response (Ji et al., 2014).

前人的工作不足

Previous Deep Neural Network (DNN) based approaches to response selection represent context and response as two embeddings. The response is selected based on the similarity of these two embeddings (Lowe et al., 2015; Kadlec et al., 2015). In these work, context and response are taken as two separate word sequences without considering the relationship among utterances in the context and response. The response selection in these models is largely influenced by word-level information. We called this kind of models as *word sequence model* in this paper. Besides word-level dependencies, utterance-level semantic and discourse information are also very important to catch the conversation topics to ensure coherence (Grosz and Sidner, 1986). For example an utterance can be an affirmation, negation or deduction to the previous utterances, or starts a new topic for discussion. This kind of utterance-level information is generally ignored in word sequence model, which may be helpful for selecting the next response. Therefore, it is necessary to take each utterance as a unit and model the context and response from the view of utterance sequence.

这篇论文的直接原因：上下文和回复作为两个独立的单词序列，没有考虑上下文和回复中话语之间的关系

例如，一个话语可以是对之前话语的肯定、否定或演绎，或者开始一个新的话题进行讨论。

论中需要加一些现象级实例来阐述

This paper proposes a multi-view response selection model, which integrates information from both **word sequence view** and **utterance sequence view**. Our assumption is that each view can represent relationships between context and response from a particular aspect, and features extracted from the word sequence and the utterance sequence provide complementary information for response selection. An effective integration of these two views is expected to improve the model performance. To the best of our knowledge, this is the first work to improve the response selection for multi-turn human-computer conversation in a multi-view manner.

我们的假设是，每个视图都可以从一个特定的方面表示上下文和响应之间的关系，从单词序列和话语序列中提取的特征为响应选择提供了补充信息。

经典套句

We evaluate the performance of the multi-view response selection model on a public corpus containing about one million context-response-label triples.

*These two authors contributed equally

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 372–381,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

简单描述一下数据集

This corpus was extracted from an online chatting room for Ubuntu troubleshooting, which is called the Ubuntu Corpus in this paper (Lowe et al., 2015). Experimental results show that the proposed multi-view response selection model significantly outperforms the current best single-view models for multi-turn human-computer conversation.

实验结果

The rest of this paper is organized as follows. In Section 2, we briefly introduce related works. Then we move on to a detailed description of our model in Section 3. Experimental results are described in Section 4. Analysis of our models is shown in Section 5. We conclude the paper in Section 6.

这里可以理下贡献或者章节安排

2 Related Work

2.1 Conversation System

Establishing a machine that can interact with human beings via natural language is one of the most challenging problems in Artificial Intelligent (AI). Early studies of conversation models are generally designed for specific domain, like booking restaurant, and require numerous domain knowledge as well as human efforts in model design and feature engineering (Walker et al., 2001). Hence it is too costly to adapt those models to other domains. Recently leveraging "big dialogs" for open domain conversation draws increasing research attentions. One critical issue for open domain conversation is to produce a reasonable response. Responding to this challenge, two promising solutions have been proposed: 1) retrieval-based model which selects a response from a large corpus (Ji et al., 2014; Yan et al., 2016; Yan et al.,). 2) generation-based model which directly generates the next utterance (Wen et al., 2015a; Wen et al., 2015b).

2.2 Response Selection

Research on response selection for human-computer conversation can be classified into two branches, i.e., single-turn and multi-turn response selection. Single-turn models only leverage the last utterance in the context for selecting resposne and most of them take the word sequence view. Lu and Li (2013) proposed a DNN-based matching model for response selection. Hu et al., (2014) improved the performance using Convolutional Neural Networks (CNN) (LeCun et al., 1989). In 2015, a further study conducted by Wang et al. (2015a) achieved better results using tree structures as the input of a DNN model. Nevertheless, those models built for single-turn response selection ignore the whole context information, which makes it difficult to be implemented in the multi-turn response selection tasks.

On the other hand, research on multi-turn response selection usually takes the whole context into consideration and views the context and response as word sequences. Lowe et al., (2015) proposed a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) based response selection model for multi-turn conversation, where words from context and response are modeled with LSTM. The selection of a response is based on the similarity of embeddings between the context and response. Similar to the work of Lowe et al., Kadlec et al., (2015) replaced LSTM with Temporal Convolutional Neural Networks (TCNN) (Kim, 2014) and Bidirect-LSTM. Their experimental results show that models with LSTM perform better than other neural networks. However, the utterance-level discourse information and dependencies have been left out in these studies since they view the context and response as word sequences.

2.3 Response Generation

套话

Another line of related research focuses on generating responses for human-computer conversation. Ritter et al., (2011) trained a phrase-based statistical machine translation model on a corpus of utterance pairs extracted from Twitter human-human conversation and used it as a response generator for single-turn conversation. Vinyals and Le (2015) regarded single-turn conversation as a sequence-to-sequence problem and proposed an encoder-decoder based response generation model, where the post response is first encoded using LSTM and its embedding used as the initialization state of another LSTM to generate the response. Shang et al., (2015) improved the encoder-decoder based model using attention signals. Sordoni et al., (2015) proposed a context-sensitive response generation model, where the context is represented by bag-of-words and fed into a recurrent language model to generate the next response.

介绍相关模型的时候要简练 准确

In this paper, we focused on the task of response selection.

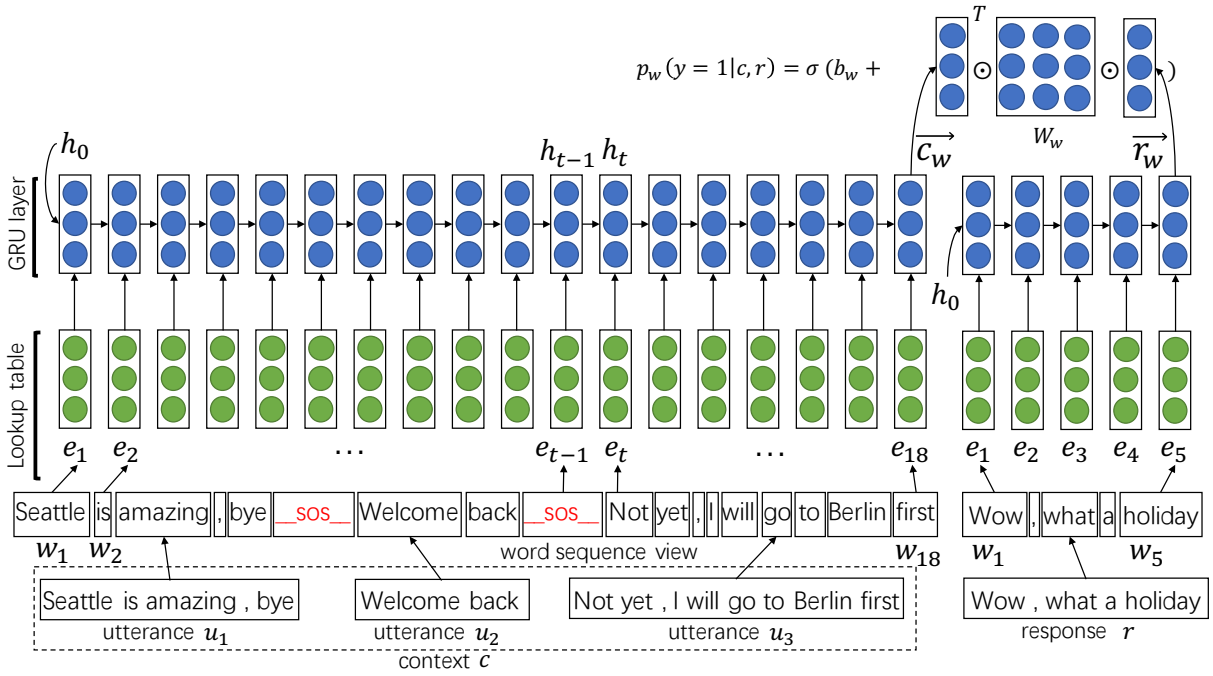


Figure 1: Word sequence model for response selection

3 Response Selection Model

In the task of response selection, a conventional DNN-based architecture represents the context and response as low dimensional embeddings with deep learning models. The response is selected based on the similarity of these two embeddings. We formulate it as

以前做法求出文本和回复之间的相似性

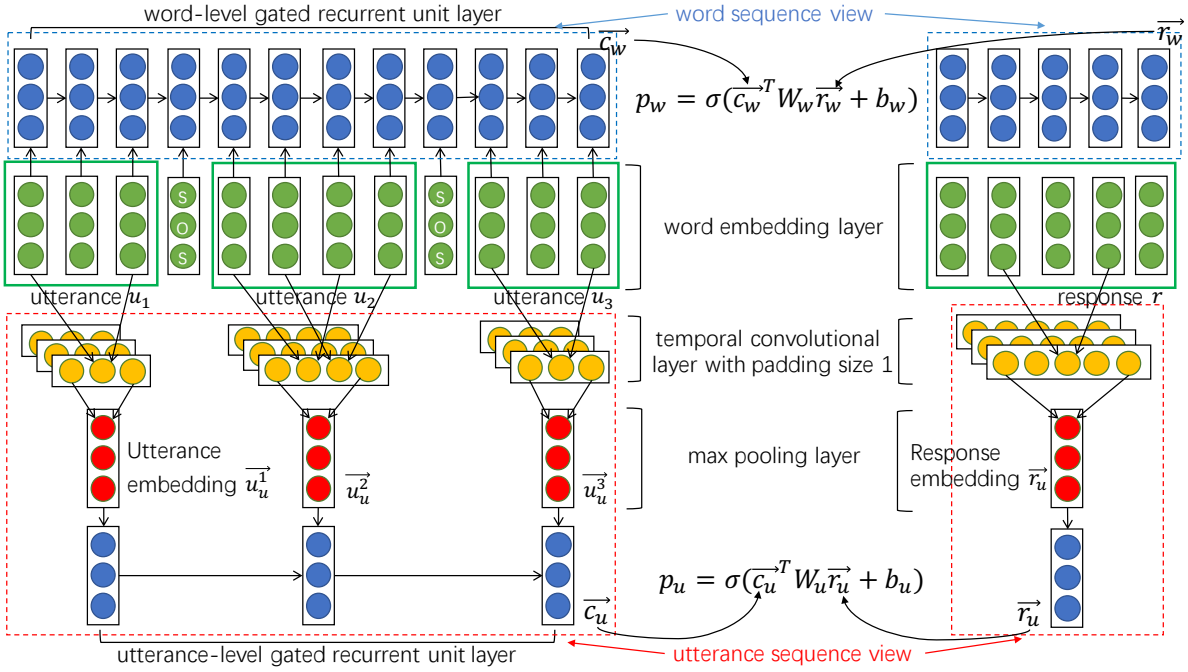
$$p(y = 1|c, r) = \sigma(\overrightarrow{c}^T W \overrightarrow{r} + b) \quad (1)$$

where c and r denote the context and response, $\overrightarrow{c}$ and $\overrightarrow{r}$ are their embeddings constructed with DNNs. $\sigma(x)$ is a sigmoid function defined as $\sigma(x) = \frac{1}{1+e^{-x}}$. $p(y = 1|c, r)$ is the confidence of selecting response r for context c. The matrix W and the scalar b are metric parameters to be learned to measure the similarity between the context and response.

We extend this architecture in a multi-view manner, which jointly models the context and response in two views. In this section, we first briefly describe the word sequence model. Then we introduce the utterance sequence model and multi-view response selection model in details.

3.1 Word Sequence Model

The word sequence model in this paper is similar to the LSTM-based model proposed in Lowe et al. (2015). As shown in Figure 1, three utterances of context c, written as u_1, u_2 and u_3, are connected as a sequence of words. A special word *__sos__* is inserted between every two adjacent utterances, denoting the boundary between utterances. Given the word sequences of context and response, words are mapped into word embeddings through a shared lookup table. A Gated Recurrent Unit neural network (GRU) (Chung et al., 2014) is employed to construct the context embedding and response embedding. It operates recurrently on the two word embedding sequences as Equation 2 to Equation 5, where h_{t-1} is the hidden state of GRU when it reads a word embedding e_{t-1} of word w_{t-1}, h_0 is a zero vector as the initiation state, z_t is an *update gate* and r_t is a *reset gate*. The new hidden state h_t for embedding e_t is a combination of the previous hidden state h_{t-1} and the input embedding e_t, controlled by the update gate z_t and reset gate r_t. U, U_z, U_r, W, W_z and W_r are model parameters of GRU to be learned. $\otimes$ denotes element-wise multiplication.



缺点：
(1)尽管将utterance作为单元处理，但处理框架比较简单，不讲究。
(2)尽管上下文和回复进行交互，但是交互的层次较高，utterance提取信息阶段回复没有参与，提取信息目的性不明确。简单的交互，没有充分利用上下文和回复的交互

为什么使用CNN
另一种提取文本的信息方式，提取文本的信息的多样性。
建模句子之间的交互的需要。将句子聚合为RNN式的输入，CNN后跟MaxPooling降维。

不用RNN的最后一个隐状态，有可能为了防止最后几个单词的权重较大，对整个句子交互产生偏移。

CNN先去提炼出utterance中的核心信息
RNN将这个核心信息编码为语境信息

Figure 2: Multi-view response selection model

$$h_t = (\mathbf{1} - z_t) \otimes h_{t-1} + z_t \otimes \hat{h}_t \quad (2)$$

$$z_t = \sigma(W_z e_t + U_z h_{t-1}) \quad (3)$$

$$\hat{h}_t = tanh(W e_t + U(r_t \otimes h_{t-1})) \quad (4)$$

$$r_t = \sigma(W_r e_t + U_r h_{t-1}) \quad (5)$$

After reading the whole word embedding sequence, word-level semantic and dependencies in the whole sequence are encoded in the hidden state of GRU, which represents the meaning of the whole sequence (Karpathy et al., 2015). Therefore we use the last hidden state of GRU as the context embedding and response embedding in word sequence model, named $\overrightarrow{c_w}$ and $\overrightarrow{r_w}$ respectively[1]. The confidence of selecting response in word sequence model is then calculated as in Equation 6:

$$p_w(y = 1|c, r) = \sigma(\overrightarrow{c_w}^T W_w \overrightarrow{r_w} + b_w) \quad (6)$$

where W_w and b_w are metric parameters to be trained in word sequence model. $\overrightarrow{c_w}$ and $\overrightarrow{r_w}$ are constructed by a same GRU in word sequence model.

[1]We use two subscripts, i.e., w and u, to distinguish notation in the two views.

3.2 Utterance Sequence Model

Utterance sequence model regards the context as a hierarchical structure, where the response and each utterance are first represented based on word embeddings, then the context embedding is constructed for the confidence calculation of response selection. As the lower part of Figure 2 illustrates, the construction of the utterance embedding and response embedding is in a convolutional manner, which contains the following layers:

话语序列模型将语境看作一个层次结构，首先基于词的嵌入来表示响应和每个话语，然后构造上下文嵌入来进行响应选择的置信度计算。

对话嵌入和响应嵌入的构造采用卷积方式

Padding Layer: Given a word embedding sequence belonging to a certain utterance (response), namely $[e_1, ..., e_m]$, the padding layer makes its outer border with $\lfloor n/2 \rfloor$ zero vectors, the padded sequence is $[0_1, .., 0_{\lfloor n/2 \rfloor}, e_1, ..., e_m, 0_1, .., 0_{\lfloor n/2 \rfloor}]$, where n is the size of convolution window used in temporal convolutional layer.

时序卷积层

Temporal Convolutional Layer: Temporal convolutional layer reads the padded word embedding sequence through a sliding convolution window with size n. For every step that the sliding window moves, a *region vector* is produced by concatenating the word embeddings within the sliding window, denoted as $[e_i \oplus ... \oplus$

这里的时序卷积 目前看来就是一个非因果的1d卷积
但是独立于每一句的时间段

$e_{i+n-1}] \in \mathbb{R}^{n|e|}$, where $\oplus$ denotes the concatenation of embeddings, $|e|$ is the size of word embedding. The temporal convolutional layer consists of k kernels, each of which implies a certain dimension and maps the *region vector* to a value in its dimension by convolution operation. The convolution result of each kernel, termed $conv_i$, is further activated with the *RELU* non-linear activation function (Xu et al., 2015), which is formulated as:

$$f_{relu}(conv_i) = max(conv_i, 0) \quad (7)$$

Pooling Layer: Because utterance and response are naturally variable-sized, we put a *max-over-time pooling layer* on the top of temporal convolutional layer (Kim, 2014), which extracts the max value for each kernel, and gets a fix-sized representation of length k for utterance and response.

尤其是CNN构建的最大汇聚表征，体现了话语和回复的核心意义

In particular, representations constructed by CNN with max-pooling reflect the core meanings of utterance and response. The embeddings of utterance u_i and response r in utterance sequence view are referred to as $\overrightarrow{u_u^i}$ and $\overrightarrow{r_u}$. Utterance embeddings are connected in the sequence and fed into a GRU, which captures utterance-level semantic and discourse information in the whole context and encodes those information as context embedding, written as $\overrightarrow{c_u}$. The confidence of selecting response r for context c in utterance sequence model, named $p_u(y = 1|c, r)$, is calculated using Equation 8:

对话的经过卷积池化后的向量拼接之后送入GRU中，在整个语境中获取话语层面的语义和话语信息，并将这些信息编码为语境嵌入信息

$$p_u(y = 1|c, r) = \sigma(\overrightarrow{c_u}^T W_u \overrightarrow{r_u} + b_u) \quad (8)$$

值得注意的是，这里使用的TCNN在构建话语嵌入和响应嵌入时是共享的。

It is worth noticing that the TCNN used here is shared in constructing the utterance embedding and response embedding. The word embeddings are also shared for both the context and response. The $\_\_sos\_\_$ tag in word sequence view is not used in the utterance sequence model.

3.3 Multi-view Model

Organic integration of different views has been proven to be very effective in the field of recommendation, representation learning and other research areas (Elkahky et al., 2015; Wang et al., 2015b). Most existing multi-view models integrate different views via a linear/nonlinear combination. Researchers have demonstrated that jointly minimizing two factors, i.e., 1) the *training error* of each view and 2) the *disagreement* between complementary views can significantly improve the performance of the combination of multi-views (Xu et al., 2013).

研究人员已经证明，共同最小化两个因素，即(1)各视图的训练误差 (2)互补视图之间的不一致可以显著提高多视图组合的性能

Our multi-view response selection model is designed as shown in Figure 2. As we can see, the context and response are jointly represented as semantic embeddings in these two views. The underlying word embeddings are shared across the context and response in these two views. The complementary information of these two views is exchanged via the shared word embeddings. The utterance embeddings are modeled through a TCNN in the utterance sequence view. Two independent Gated Recurrent Units are used to model the word embeddings and utterance embeddings separately on word sequence view and utterance sequence view, the former of which captures dependencies in word level and the latter captures utterance-level semantic and discourse information. Confidences for selecting the response in these two views are calculated separately. We optimize the multi-view model by minimizing the following loss:

这两个视图中的上下文和响应共享底层单词embeddings。这两个视图的互补信息通过共享的word embeddings交换。

$$\mathcal{L} = \mathcal{L}_{\mathcal{D}} + \mathcal{L}_{\mathcal{L}} + \frac{\lambda}{2}\|\theta\| \quad (9)$$

$$\mathcal{L}_{\mathcal{D}} = \sum_i (p_w(l_i)\bar{p_u}(l_i) + p_u(l_i)\bar{p_w}(l_i)) \quad (10)$$

$$\mathcal{L}_{\mathcal{L}} = \sum_i (1 - p_w(l_i)) + \sum_i (1 - p_u(l_i)) \quad (11)$$

where the object function of the multi-view model $\mathcal{L}$ is comprised of the *disagreement loss* $\mathcal{L}_{\mathcal{D}}$, the *likelihood loss* $\mathcal{L}_{\mathcal{L}}$ and the regular term $\frac{\lambda}{2}\|\theta\|$. $p_w(l_i) = p_w(y = l_i|c, r)$ and $p_u(l_i) = p_u(y = l_i|c, r)$ denote the likelihood of the i-th instance with label l_i from training set in these two views. Only two labels, $\{0, 1\}$, denote the correctness of the response during training. $\bar{p}_w(l_i)$ and $\bar{p}_u(l_i)$ denote the probability $p_w(y \neq l_i)$ and $p_u(y \neq l_i)$ respectively. The multi-view model is trained to jointly minimize the *disagreement loss* and the *likelihood loss*. θ denotes all the parameters of the multi-view model.

The unweighted summation of confidences from these two views is used during prediction, defined as

Model/Metrics	1 in 10 R@1	1 in 10 R@2	1in 10 R@5	1 in 2 R@1
Random-guess	10%	20%	50%	50%
TF-IDF	41.0%	54.5%	70.8%	65.9%
Word-seq-LSTM (Lowe et al., 2015)	60.40%	74.50%	92.60%	87.80%
Word-seq-GRU	60.85%	75.71%	93.13%	88.55%
Utter-seq-GRU	62.19%	76.56%	93.42%	88.83%
Multi-view	**66.15%**	**80.12%**	**95.09%**	**90.80%**

Table 1: Performance comparison between our models and baseline models. In the table, **Word-seq-LSTM** is the experiment result of the LSTM-based word sequence model reported by Lowe et at (2015). **Word-seq GRU** is the word sequence model that we implement with GRU. **Utter-seq-GRU** is the proposed utterance-sequence model. The **Multi-view** is our multi-view response selection model. In addition, we list the performance of **Random-guess** and **TF-IDF**

in Equation 12:

$$s_{mtv}(y = 1|c, r) = p_w(y = 1|c, r) + p_u(y = 1|c, r) \quad (12)$$

The response with larger $s_{mtv}(y = 1|c, r)$ is more likely to be selected. We will investigate other combination models in our future work.

4 Experiment

4.1 Dataset

介绍数据集

Our model is evaluated on the public Ubuntu Corpus (Lowe et al., 2015), designed for response selection study of multi-turn human-computer conversation (Serban et al., 2015). The dataset contains 0.93 million human-human dialogues crawled from an Internet chatting room for Ubuntu trouble shooting. Around 1 million context-response-labeled triples, namely $< c, r, l >$, are generated for training after preprocessing[2], where the original context and the corresponding response are taken as the positive instances while the random utterances in the data set taken as the negative instances, and the number of positive instance and negative instance in training set is balanced. The validation set and testing set are constructed in a similar way to the training set, with one notable difference that for each context and the corresponding positive response, 9 negative responses are randomly selected for further evaluation.

4.2 Experiment Setup

介绍实验setup

套话 Following the work of Lowe et al., (2015), the evaluation metric is 1 in m Recall@k (denoted 1 in m R@k), where a response selection model is designed to select k most likely responses among m candidates, and it gets the score "1" if the correct response is in the k selected ones. This metric can be seen as an adaptation of the precision and recall metrics previously applied to dialogue datasets (Schatzmann et al., 2005). It is worth noticing that 1 in 2 R@1 equals to precision and recall in binary classification.

设计了一个响应选择模型，在m个候选项中选择k个最可能的响应，如果正确的响应在k个候选项中，则得分为"1"

4.3 Model Training and Hyper-parameters

介绍超参数

We initialize word embeddings with a pre-trained embedding matrix through GloVe (Pennington et al., 2014) [3]. We use Stochastic Gradient Descent (SGD) for optimizing. Hidden size for a gated recurrent unit is set to 200 in both word sequence model and utterance sequence model. The number of convolutional kernels is set to 200. Our initial learning rate is 0.01 with mini-batch size of 32. Other hyper-parameters are set exactly the same as the baseline. We train our models with a single machine using 12 threads and each model will converge after 4-5 epochs of training data. The best model is selected with a holdout validation dataset.

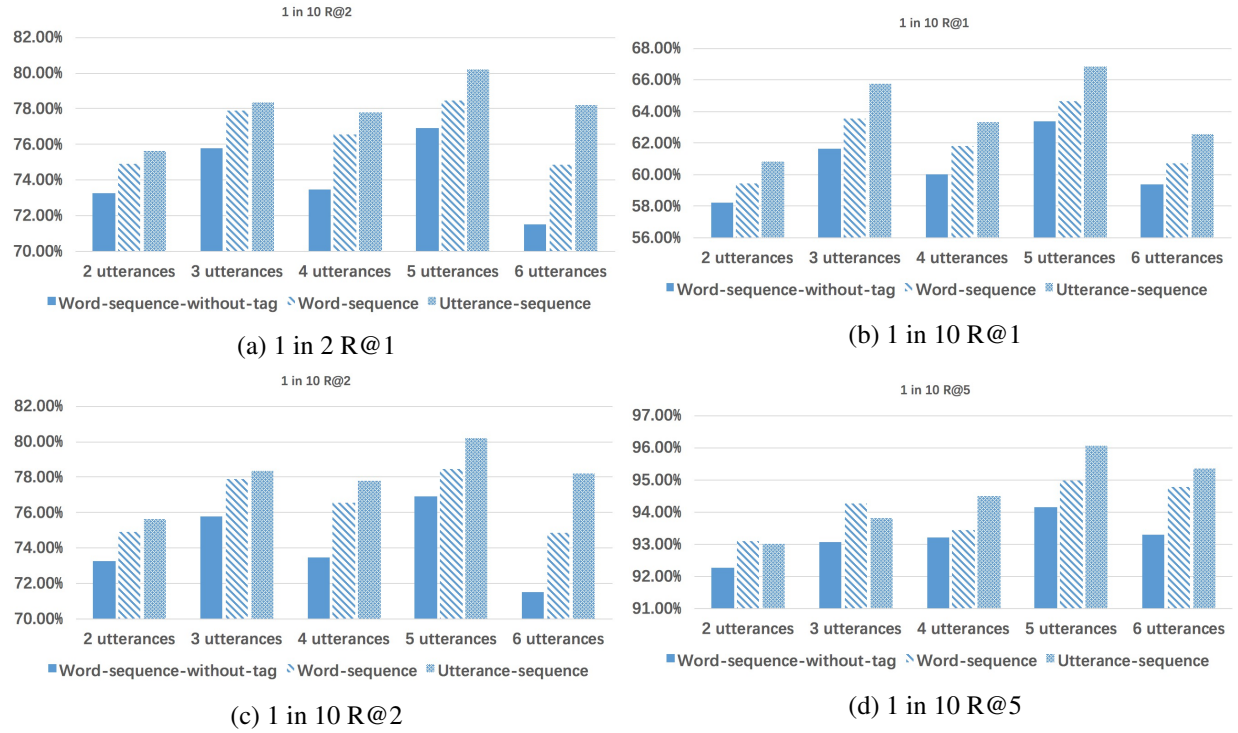
4.4 Comparison Approaches

介绍baseline

We consider the word sequence model implemented by Lowe et at., (2015) with LSTM as our baseline, the best model in context-sensitive response selection so far. Moreover, we also implement the word sequence model and the utterance sequence model with GRU for further analysis. Two simple approaches are also implemented, i.e., the Random-

[2]Preprocessing includes tokenization, recognition of named entity, urls and numbers.

[3]Initialization of word embedding can be obtained on https://github.com/npow/ubottu



(a) 1 in 2 R@1

(b) 1 in 10 R@1

(c) 1 in 10 R@2

(d) 1 in 10 R@5

Figure 3: Performance comparison between word sequence model (with/without $\_\_sos\_\_$ tags) and utterance sequence model. We choose the number of utterances in range of [2,6], since most samples in testset fall in this interval

guess and the TF-IDF, as the bottom line for performance comparison. The performance of Random-guess is calculated by mathematics with an assumption that each response in candidates has the equal probability to be selected. The TF-IDF is implemented in the same way in Lowe et al., (2015). TF for a word is calculated as the count of times it appears in a certain context or response. IDF for each word w is $log(\frac{N}{|d \in D : w \in d|})$, where D denotes the whole training set, N is the size of D, d is a conversation in D. The context and the response in testset are represented as a bag-of-words according to TF-IDF. The selection confidence is estimated as the cosine score between context and response.

4.5 Experimental Result

We summarize the experiment result in Table 1. As shown in Table 1, all DNN-based models achieve significant improvements compared to Random-guess and TF-IDF, which implies the effectiveness of DNN models in the task of response selection. The word sequence models implemented with GRU and LSTM achieve similar performance. The utterance sequence model significantly outperforms word sequence models for 1 in 10 R@1. Multi-view model significantly outperforms all the other models, especially for 1 in 10 R@1, which is more difficult and closer to the real world scenario than other metrics. The experimental result demonstrates the effectiveness of multi-view model and proves that word sequence view and utterance sequence view can bring complementary information for each other.

5 Analysis

We examine the complementarity between word sequence model and utterance sequence model in two folds, i.e., via statistic analysis and case study.

5.1 Statistical Analysis

We compare the performance of word sequence model[4] and utterance sequence model for different number of utterances in the contexts. In addition, we also examine what the contribution $\_\_sos\_\_$ tag makes in word sequence view. The performance

[4]The GRU-based word sequence model that we implemented is used for comparison.

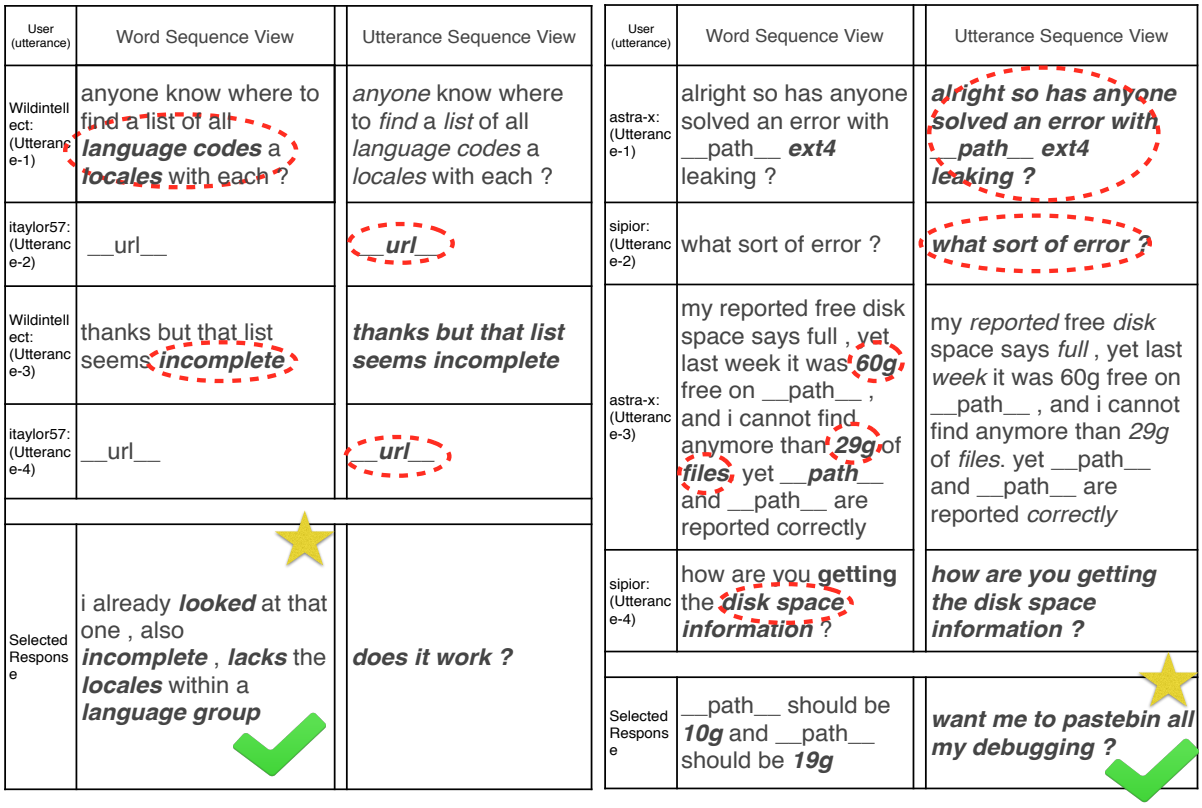


Figure 4: Case studies for analysis of word sequence model and utterance sequence model. The context and the selected responses are collected from testset. Response with a green checkmark means it is a correct one, otherwise it is incorrect. Words (Utterances) in **bold** are the important elements recognized by our importance analysis approach. The yellow start denotes the selection of multi-view model.

is shown in Figure 3. We can see that as the number of turns increases, the utterance sequence model outperforms word sequence model more significantly, which implies that utterance sequence model can provide complementary information to word sequence model for a long context. Furthermore, word sequence model without *__sos__* tag has an obvious fall in performance compared with word sequence model with *__sos__*, which implies its crucial role in distinguishing utterances for modeling context.

5.2 Case Study

We analyze samples from testset to examine the complementarity between these two views. The key words for word sequence model and core utterances for utterance sequence model are extracted for analysis. These important elements are recognized based on the work of Li et al. (2015), where the gradients of their embeddings are used for importance analysis. After studying the testset, we find that the word sequence model selects responses according to the **matching of key words** while the utterance sequence model selects responses based on the **matching of core utterances**. We list two cases in Figure 4 as examples.

我们发现，单词序列模型根据关键词的匹配来选择响应，而话语序列模型根据核心话语的匹配来选择响应。

As it shows, the word sequence model prefers to select the response that shares similar key words to the context, such as the words *"incomplete"* and *"locales"* in example 1 or *"60g"* and *"19g"* in example 2. Although key word matching is a useful feature in selecting response for cases such as example 1, it fails in cases like example 2, where incorrect response happens to share similar words with the context. Utterance sequence model, on the other side, leverages core utterances for selecting response. As shown in example 2, utterance-1 and utterance-2 are recognized as the core utterances, the main topic of the two utterance is *"solved"* and *"error"*, which is close to the topic of the correct re-

正确回答的核心意义是与不同话语中的不同词汇相结合的，而话语序列模型并没有很好地表现出来

sponse. However, for cases like example 1, where the core meaning of correct response is jointly combined with different words in different utterances, the utterance sequence model does not perform well.

The multi-view model can successfully select the correct responses in both two examples, which implies its ability to jointly leverage information from these two views.

6 Conclusion

In this paper, we propose a multi-view response selection model for multi-turn human-computer conversation. We integrate the existing word sequence view and a new view, i.e., utterance sequence view, into a unified multi-view model. In the view of utterance sequence, discourse information can be learnt through utterance-level recurrent neural network, different from word sequence view. The representations learnt from the two views provide complementary information for each other in the task of response selection. Experiments show that our multi-view model significantly outperforms the state-of-the-art word sequence view models. We will extend our framework to response generation approaches in our future work. We believe it will help construct a better representation of context in the encoding phrase of DNN-based generation model and thus improve the performance.

Acknowledgement

This paper is supported by National Basic Research Program of China (973 program No. 2014CB340505). We gratefully thank the anonymous reviewers for their insightful comments.

References

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. International World Wide Web Conferences Steering Committee.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

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

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, pages 2042–2050.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

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 preprint arXiv:1506.08909*.

Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. In *Advances in Neural Information Processing Systems*, pages 1367–1375.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In Proc. EMNLP*, pages 1532–1543.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *In Proc. EMNLP*, pages 583–593. Association for Computational Linguistics.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

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. *arXiv preprint arXiv:1506.06714*.

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

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. Association for Computational Linguistics.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015a. Syntax-based deep matching of short texts. *arXiv preprint arXiv:1503.02427*.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015b. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1083–1092.

Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, 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 (SIGDIAL)*. Association for Computational Linguistics, September.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, 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 (EMNLP)*. Association for Computational Linguistics, September.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. Docchat: An information retrieval approach for chatbot engines using unstructured documents.

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.