Joint Effects of Context and User History for Predicting Online Conversation Re-entries

Xingshan Zeng[1,2], **Jing Li**[3]*, **Lu Wang**[4], **Kam-Fai Wong**[1,2]
[1]The Chinese University of Hong Kong, Hong Kong
[2]MoE Key Laboratory of High Confidence Software Technologies, China
[3]Tencent AI Lab, Shenzhen, China
[4]Northeastern University, Boston, MA, United States
[1,2]{xszeng,kfwong}@se.cuhk.edu.hk
[3]ameliajli@tencent.com, [4]luwang@ccs.neu.edu

Abstract

As the online world continues its exponential growth, interpersonal communication has come to play an increasingly central role in opinion formation and change. In order to help users better engage with each other online, we study a challenging problem of re-entry prediction foreseeing whether a user will come back to a conversation they once participated in. We hypothesize that both the context of the ongoing conversations and the users' previous chatting history will affect their continued interests in future engagement. Specifically, we propose a neural framework with three main layers, each modeling context, user history, and interactions between them, to explore how the conversation context and user chatting history jointly result in their re-entry behavior. We experiment with two large-scale datasets collected from Twitter and Reddit. Results show that our proposed framework with bi-attention achieves an F1 score of 61.1 on Twitter conversations, outperforming the state-of-the-art methods from previous work.

1 Introduction

Interpersonal communication plays an important role in information exchange and idea sharing in our daily life. We are involved in a wide variety of dialogues every day, ranging from kitchen table conversations to online discussions, all help us make decisions, better understand important social issues, and form personal ideology. However, individuals have limited attentions to engage in the massive amounts of online conversations. There thus exists a pressing need to develop automatic conversation management tools to keep track of the discussions one would like to keep engaging in. To meet such demand, we study the problem of *predicting online conversation re-entries*, where we aim to forecast whether the users will return to a discussion they once entered.

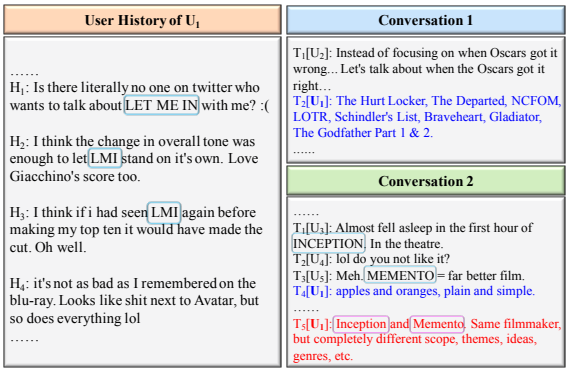


Figure 1: Sample tweets in the chatting history of user $\mathbf{U_1}$ and two Twitter conversation snippets $\mathbf{U_1}$ engaged in. $\mathbf{H_i}$: the i-th tweet in $\mathbf{U_1}$'s history. $\mathbf{T_i[U_j]}$: the i-th turn posted by $\mathbf{U_j}$. First entries by $\mathbf{U_1}$ are highlighted in blue in both conversations. $\mathbf{U_1}$ only returns to the second one.

What will draw a user back? To date, prior efforts for re-entry prediction mainly focus on modeling users engagement patterns in the ongoing conversations (Backstrom et al., 2013) or rely on the social network structure (Budak and Agrawal, 2013), largely ignoring the rich information in users' previous chatting history.

Here we argue that effective prediction of one's re-entry behavior requires the understanding of both the **conversation context**—what has been discussed in the dialogue under consideration, and **user chatting history** (henceforth user history)—what conversation topics the users are actively involved in. In Figure 1, we illustrate how the two factors together affect a user's re-entry behavior. Along with two conversations that user $\mathbf{U_1}$ participated in, also shown is their chatting history in previous discussions. $\mathbf{U_1}$ comes back to the second conversation since it involves topics on movies (e.g. mentioning *Memento* and *Inception*) and thus suits their interests according to the chatting his-

* Jing Li is the corresponding author.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2809–2818
Florence, Italy, July 28 - August 2, 2019.

tory, which also talked about movies.

In this work, we would like to focus on the joint effects of conversation context and user history, ignoring other information. It would be a more challenging yet general task, since information like social networks may be not available in some certain scenarios. To study how conversation context and user history jointly affect user re-entries, we propose a novel neural framework that incorporates and aligns the indicative representations from the two information source. To exploit the joint effects, four mechanisms are employed here: *simple concatenation* of the two types of representation, *attention* mechanism over turns in context, *memory networks* (Sukhbaatar et al., 2015) — able to learn context attentions in aware of user history, and *bi-attention* (Seo et al., 2016) — further capturing interactions from two directions (context to history and history to context). More importantly, our framework enables the re-entry prediction and corresponding representations to be learned in an end-to-end manner. On the contrary, previous methods for the same task rely on handcrafted features (Backstrom et al., 2013; Budak and Agrawal, 2013), which often require labor-intensive and time-consuming feature engineering processes. To the best of our knowledge, we are the first to explore the joint effect of conversation context and user history on predicting re-entry behavior in a neural network framework.

We experiment with two large-scale datasets, one from Twitter (Zeng et al., 2018), the other from Reddit which is newly collected[1]. Our framework with bi-attention significantly outperforms all the comparing methods including the previous state of the art (Backstrom et al., 2013). For instance, our model achieves an F1 score of 61.1 on Twitter conversations, compared to an F1 score of 57.0 produced by Backstrom et al. (2013), which is based on a rich set of handcrafted features. Further experiments also show that the model with bi-attention can consistently outperform comparisons given varying lengths of conversation context. It shows that bi-attention mechanism can well align users' personal interests and conversation context in varying scenarios.

After probing into the proposed neural framework with bi-attention, we find that meaningful representations are learned via exploring the joint effect of conversation context and user history, which explains the effectiveness of our framework in predicting re-entry behavior. Finally, we carry out a human study, where we ask two humans to perform on the same task of first re-entry prediction. The model with bi-attention outperforms both humans, suggesting the difficulty of the task as well as the effectiveness of our proposed framework.

[1]The datasets and codes are released at: https://github.com/zxshamson/re-entry-prediction

2 Related Work

Response Prediction. Previous work on response prediction mainly focuses on predicting whether users will respond to a given social media post or thread. Efforts have been made to measure the popularity of a social media post via modeling the response patterns in replies or retweets (Artzi et al., 2012; Zhang et al., 2015). Some studies investigate post recommendation by predicting whether a response will be made by a given user (Chen et al., 2012; Yan et al., 2012; Hong et al., 2013; Alawad et al., 2016).

In addition to post-level prediction, other studies focus on response prediction at the conversation-level. Zeng et al. (2018) investigate microblog conversation recommendation by exploiting latent factors of topics and discourse with a Bayesian model, which often requires domain expertise for customized learning algorithms. Our neural framework can automatically acquire the interactions among important components that contribute to the re-entry prediction problem, and can be easily adapted to new domains. For the prediction of re-entry behavior in online conversations, previous methods rely on the extraction of manually-crafted features from both the conversation context and the user's social network (Backstrom et al., 2013; Budak and Agrawal, 2013). Here we tackle a more challenging task, where the re-entries are predicted without using any information from social network structure, which ensures the generalizability of our framework to scenarios where such information is unavailable.

Online Conversation Behavior Understanding. Our work is also in line with conversational behavior understanding, including how users interact in online discourse (Ritter et al., 2010) and how such behavior signals the future trajectory, including their continued engagement (Backstrom et al., 2013; Jiao et al., 2018) and the appearance of impolite behavior (Zhang et al., 2018). To

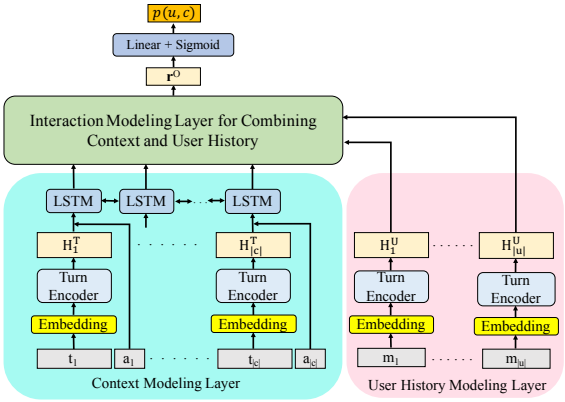


Figure 2: The generic framework for re-entry prediction. We implement it with three encoders (Average Embedding, CNN, and BiLSTM) for turn modeling and four mechanisms (Simple Concatenation, Attention, Memory Networks, and Bi-attention) for modeling interactions between context and user history.

better understand the structure of conversations, Recurrent Neural Network (RNN)-based methods have been exploited to capture temporal dynamics (Cheng et al., 2017; Zayats and Ostendorf, 2018; Jiao et al., 2018). Different from the above work, our model not only utilizes the conversations themselves, but also leverages users' prior posts in other discussions.

3 Neural Re-entry Prediction Combining Context and User History

This section describes our neural network-based conversation re-entry prediction framework exploring the joint effects of context and user history. Figure 2 shows the overall architecture of our framework, consisting of three main layers: context modeling layer, user history modeling layer, and interaction modeling layer to learn how information captured by the previous two layers interact with each other and make decisions conditioned on their joint effects. Here we adopt four mechanisms for interaction modeling: simple concatenation, attention, memory networks, and bi-attention, which will be described later.

3.1 Input and Output

We start with formulating model input and output. At input layer, our model is fed with two types of information, the chatting history of the target user u and the observed context of the target conversation c. The goal of our model is to output a Bernoulli distribution $p(u, c)$ indicating the estimated likelihood of whether u will re-engage in the conversation c. Below gives more details.

Formally, we formulate the context of c as a sequence of chronologically ordered turns $\langle t_1, t_2, \cdots, t_{|c|} \rangle$, where the last turn $t_{|c|}$ is posted by u (we then predict u's re-entries afterwards). Each turn t is represented by a sequence of words $\mathbf{w}_t$, and an auxiliary triple, $\mathbf{a}_t = \langle i_t, r_t, u_t \rangle$, where i_t, r_t, and u_t are three indexes indicating the position of turn t, which turn t replies to, and the author of t, respectively. Here $\mathbf{a}_t$ is used to record the replying structures as well as the user's involvement pattern.

For the user history, we formulate it as a collection of u's chatting messages $\{m_1, m_2, \cdots, m_{|u|}\}$, all posted before the time $t_{|c|}$ occurs. Each message m is denoted as its word sequence, $\mathbf{w}_m$.

In the following, we explain how the aforementioned representations are processed by our model to make predictions. The three main layers in Figure 2 are described in Sections 3.2, 3.3, and 3.4, respectively. The learning objective is presented in Section 3.5.

3.2 Context Modeling Layer

The context modeling layer captures representations from the observed context for the target conversation c. To this end, we jointly model the content in each turn (henceforth **turn modeling**) and the turn interactions in conversation structure (henceforth **structure modeling**).

Turn Modeling. The turn representations are modeled via turn-level word sequence with a turn encoder. We exploit three encoders here: **Average Embedding** (Averaging each word's embedding representation), **CNN** (Convolutional Neural Networks), and **BiLSTM** (Bidirectional Long Short-Term Memory). BiLSTM's empirical performance turns out to be slightly better (will be reported in Table 2).

Concretely, given the conversation turn t, each word w_i of t is represented as a vector mapped by an embedding layer $I(\cdot)$, which is initialized by pre-trained embeddings and updated during training. The embedded vector $I(w_i)$ is then fed into the turn encoder, yielding the *turn representation* for t, denoted by H_t^T.[2]

[2]For all the BiLSTM encoders in this work, without otherwise specified, we take the concatenation of all hidden states from both the directions as its learned representations.

Structure Modeling. To learn the conversational structure representations for c, our model applies BiLSTM, namely structure encoder, to capture the interactions between adjacent turns in its context. Each state of this structure encoder sequentially takes t's turn representation, H_t^T, concatenated with the auxiliary triple, $\mathbf{a}_t$, as input to produce the structure representation H^C. Our intuition is that H^C should capture both the content of the conversation and interaction patterns among its participants. Then H^C, considered as the *context representation* for c, is sent to interaction modeling layer as part of its input.

3.3 User History Modeling Layer

To encode the user history for target user u, in this layer, we first apply the same encoder in turn modeling to encode each chatting message m by u, as they both explore the post-level representations. The turn encoder is sequentially fed with the embedded word in m, and produce the message-level representation H_m^M. All messages in u's user history are further concatenated into a matrix H^U, serving as u's *user history representation* and the input of the next layer.

3.4 Interaction Modeling Layer

To capture whether the discussion points in c match the interests of u, H^C (from context modeling) and H^U (from user history modeling) are merged through an interaction modeling mechanism over the two sources of information. We hypothesize that users will be likely to come back to a conversation if its topic fits their own interests. Here, we explore four different mechanisms for interaction modeling. Their learned interaction representation, denoted as $\mathbf{r}^O$, is fed into a sigmoid-activated neural perceptron (Glorot et al., 2011), for predicting final output $p(u, c)$. It indicates how likely the target user u will re-engage in the target conversation c. We then describe the four mechanisms to learn $\mathbf{r}^O$ in turn below.

Simple Concatenation. Here we simply put context representation (last state) and user representations (with average pooling) side by side, yielding $\mathbf{r}^O = [H^C_{|c|}; \sum_j^{|u|} H^U_j/|u|]$ as the interaction representation for re-entry prediction.

Attention. To capture the context information useful for re-entry prediction, we exploit an attention mechanism (Luong et al., 2015) over H^C. Attentions are employed to "soft-address" important context turns according to their similarity with user representation (with average pooling). Here we adopt dot attention weights and define the attended interaction representation as:

$$\mathbf{r}^O = \sum_i^{|c|} \alpha_i \cdot H_i^C, \alpha_i = softmax(H_i^C \cdot \sum_j^{|u|} H_j^U/|u|) \quad (1)$$

Memory Networks. To further recognize indicative chatting messages in user history, we also apply end-to-end memory networks (MemN2N) (Sukhbaatar et al., 2015) for interaction modeling. It can be seen as a recurrent attention mechanism over chatting messages (stored in memory). Hence fed with context representation, memory networks will yield a memory-aware vector as interaction representation:

$$\mathbf{r}^O = \sum_j^{|u|} \alpha_j \cdot f_{turn}(H_j^U), \alpha_j = softmax(H^C_{|c|} \cdot H_j^U) \quad (2)$$

where $f_{turn}(\cdot)$ denotes the unit function used for turn modeling.

Here we adopt multi-hop memory mechanism to allow deep user interests to be learned from chatting history. For more details, we refer the readers to Sukhbaatar et al. (2015).

Bi-attention. Inspired by Seo et al. (2016), we also apply bi-attention mechanism to explore the joint effects of context and user history. Intuitively, the bi-attention mechanism looks for evidence, if any, indicating the topics of the current conversation that align with the user's interests from two directions (i.e. context to history and history to context), such as the names of two movies *Inception* and *Let Me In* shown in Figure 1. Concretely, bi-attention mechanism captures context-aware attention over user history messages:

$$\alpha_{ij}^U = \frac{\exp(f_{score}(H_i^C, H_j^U))}{\sum_{j'=1}^{|u|} \exp(f_{score}(H_i^C, H_{j'}^U))} \quad (3)$$

where the alignment score function takes a form of $f_{score}(H_i^C, H_j^U) = W_{bi-att}[H_i^C; H_j^U; H_i^C \circ H_j^U]$. It captures the similarity of the i-th context turn and the j-th user history message. The weight vector W_{bi-att} is learnable in training.

Likewise, we compute user-aware attention over context turns. Afterwards, the bi-directional attended representations are concatenated and passed into a ReLU-activated multilayer perceptron (MLP), yielding representation $\mathbf{r}$. $\mathbf{r}$, as turn-level representation, is then sequentially fed into a two-layer BiLSTM, to produce the interaction representation $\mathbf{r}^O$.

3.5 Learning Objective

For parameter learning in our model, we design the objective function based on cross-entropy loss as following:

$$\mathcal{L} = -\sum_{i} \left[\lambda y_i \log(\hat{y}_i) + \mu(1-y_i)\log(1-\hat{y}_i)\right] \quad (4)$$

where the two terms reflect the prediction on positive and negative instances, respectively. Moreover, to take the potential data imbalance into account, we adopt two trade-off weights λ and μ. The parameter values are set based on the proportion of positive and negative instances in the training set (see Section 4). $\hat{y}_i$ denotes the re-entry probability estimated from $p(u, c)$ for the i-th instance, and y_i is the corresponding binary ground-truth label (1 for re-entry and 0 for the opposite).

4 Experimental Setup

Data Collection and Statistic Analysis. To study re-entry behavior in online conversations, we collected two datasets: one is released by Zeng et al. (2018) containing Twitter conversations formed by tweets from the TREC 2011 microblog track data[3] (henceforth **Twitter**), and the other is *newly collected* from Reddit (henceforth **Reddit**), a popular online forum. In our datasets, the conversations from Twitter concern diverse topics, while those from Reddit focus on the political issues. Both datasets are in English.

To build the Reddit dataset, we first downloaded a large corpus publicly available on Reddit platform.[4] Then, we selected posts and comments in subreddit "politics" posted from Jan to Dec 2008. Next, we formed Reddit posts and comments into conversations with replying relations revealed by the "*parent_id*" of each comment. Last, we removed conversations with only one turn.

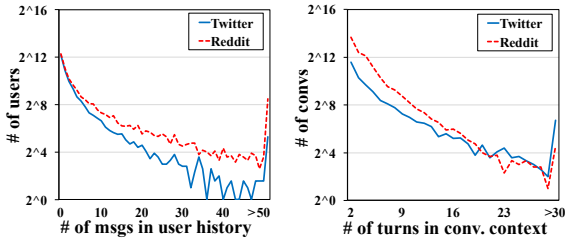
In our main experiment, we focus on *first re-entry* prediction, i.e. we predict whether a user u will come back to a conversation c, given current turns until u's first entry in c as context and u's past chatting messages (posted before u engaging in c). For model training and evaluation, we randomly select 80%, 10%, and 10% conversations to form training, development, and test sets.

The statistics of the two datasets are shown in Table 1. As can be seen, users participate twice on average in Twitter conversations, and the number is only 1.3 on Reddit. This results in the severe imbalance over instances of re-entry and non re-entry (negative samples where users do not come back) on both datasets. Therefore, strategies should be adopted for alleviating the data imbalance issue, as done in Eq. (4). It indicates the sparse user activity in conversations, where most users engage in a conversation only once or twice. Thus predicting user re-entries only with context will not perform well, and the complementary information underlying user history should be leveraged.

We further study the distributions of message number in user history and turn number in conversation context on both datasets. As shown in Figure 3, there exists severe sparsity in either user history or conversation context. Thus combining them both might help alleviate the sparsity in one information source. We also notice that Twitter and Reddit users exhibit different conversation behaviors. Reddit users tend to engage in more conversations, resulting in more messages in user history (as shown in Figure 3(a)). Twitter users are more likely to stay within each conversation, leading to lengthy discussions and larger re-entry frequencies on average, as shown in Figure 3(b) and Table 1.

	Twitter	Reddit
# of users	10,122	13,134
# of conversations	7,500	29,477
# of re-entry instances	5,875	12,780
# of non re-entry instances	8,677	39,988
Avg. # of convs per user	1.7	5.9
Avg. # of msgs in user history	3.9	8.4
Avg. # of entries per user per conv	2.0	1.3
Avg. # of turns per conv	5.2	3.7
Avg. # of users per conv	2.3	2.6

Table 1: Statistics of two datasets.



(a) User history (b) Conversation context

Figure 3: Distributions of message number in user history and turn number in conversation context on the two datasets.

[3] https://trec.nist.gov/data/tweets/

[4] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

Data Preprocessing and Model Setting. For preprocessing Twitter data, we applied Glove tweet preprocessing toolkit (Pennington et al., 2014).[5] For the Reddit dataset, we first applied the open source natural language toolkit (NLTK) (Loper and Bird, 2002) for word tokenization. Then, we replaced links with the generic tag "URL" and removed all the non-alphabetic tokens. For both datasets, a vocabulary was built and maintained in experiments with all the tokens (including emoticons and punctuation) from training data.

For model setups, we initialize the embedding layer with 200-dimensional Glove embedding (Pennington et al., 2014), where Twitter version is used for our Twitter dataset and the Common Crawl version applied on Reddit dataset.[6] All the hyper-parameters are tuned on the development set by grid search. The batch size is set to 32. Adam optimizer (Kingma and Ba, 2014) is adopted for parameter learning with initial learning rate selected among $\{10^{-3}, 10^{-4}, 10^{-5}\}$. For the BiLSTM encoders, we set the size of their hidden states to 200 (100 for each direction). For the CNN encoders, we use filter windows of 2, 3, and 4, each with 50 feature maps. In MemN2N interaction mechanism, we set hop numbers to 3. In the learning loss, we set $\mu = 1$ and $\lambda = 2$, the weights to tackle data imbalance. For re-entry prediction, a user is considered to come back if the estimated probability for re-entry is larger than 0.5.

Baselines and Comparisons. For comparisons, we consider three baselines. RANDOM baseline: randomly pick up a "yes-or-no" answer. HISTORY baseline: predict based on users' history re-entry rate before current conversation, which will answer "yes" if the rate exceeds a pre-defined threshold (set on development data), and "no" otherwise. (For users who lack such information before current conversation, it predicts "yes or no" randomly.) ALL-YES baseline: always answers "yes" in re-entry prediction. Its assumption is that users tend to be drawn back to the conversations they once participated by the platform's auto messages inviting them to return.

For supervised models, we compare with CCCT, the state-of-the-art method proposed by Backstrom et al. (2013), where the bagged decision tree with manually-crafted features (including arrival patterns, timing effects, most related terms, etc.) are employed for re-entry prediction. We do not compare with Budak and Agrawal (2013), since most of its features are related to social networks or Twitter group information, which is unavailable in our data.

In our proposed neural framework, we further compare varying encoders for turn modeling and mechanisms to model the interactions between user history and conversation context. We first compare three turn encoders — AVG-EMBED (average embedding), CNN, and BILSTM, to examine their performance in turn representation learning. Their results are compared on our variant only with context modeling layer and the best encoder (turned out to be BILSTM) is applied on the full model. For the interaction modeling layer, we also study the effectiveness of four mechanisms to combine user history and conversation context — simple concatenation (CON), attention (ATT), memory networks (MEM), and bi-attention (BIA).

5 Results and Analysis

This section first discusses prediction results of first re-entry in Section 5.1. We then present the results of the second and third re-entry prediction in Section 5.2, as well as an analysis on user history effects. Section 5.3 then provides explanations on what we learn from the joint effects from context and user history, indicative of user re-entries. Finally, we conduct a human study to compare human performance on the same task with our best model (Section 5.4).

5.1 First Re-entry Prediction Results

In main experiment, we adopt the automatic evaluation metrics — AUC, F1 score, precision, and recall, and focus on the prediction of the major re-entry type — *first re-entry*, where conversation context up to user's first participation is given. As shown in Table 1, most users, if re-entry, only return once to a conversation. Also, in conversation management, the prediction of first re-entry is a challenging yet practical problem. We will discuss second and third re-entry prediction later in Section 5.2. The comparison results are reported in Table 2. On both datasets, we observe:

• *First re-entry prediction is challenging*. All models produce AUC and F1 scores below 70.

[5] https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

[6] https://nlp.stanford.edu/projects/glove/

Models	Twitter				Reddit			
	AUC	F1 Score	Precision	Recall	AUC	F1 Score	Precision	Recall
Baselines								
RANDOM	51.0	45.0	40.3	50.9	49.4	32.6	24.5	48.7
HISTORY	50.1	46.4	42.2	51.4	50.7	35.2	26.9	50.9
ALL-YES	50.0	54.9	37.9	**100.0**	50.0	38.5	23.8	**100.0**
S.O.T.A								
CCCT	57.7	57.0	45.5	76.4	59.9	39.8	**44.7**	36.0
W/O History								
AVG-EMBED	60.4	59.0	43.5	91.8	63.7	42.4	31.0	67.2
CNN	58.8	59.1	43.2	93.5	64.0	42.8	31.1	68.5
BILSTM	60.4	59.4	45.8	85.0	64.1	43.1	31.4	69.5
With History								
BILSTM+CON	51.0	58.0	40.9	**100.0**	50.1	38.6	24.0	98.3
BILSTM+ATT	58.4	59.0	44.6	87.3	60.3	41.3	27.8	82.4
BILSTM+MEM	61.3	59.9	45.7	87.5	65.5	43.7	31.8	69.9
BILSTM+BIA	**62.7**	**61.1**	**47.0**	87.7	**67.1**	**45.4**	33.9	68.9

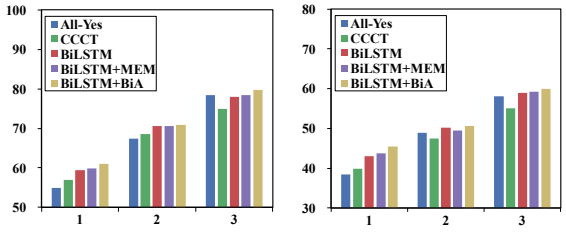
Table 2: Results on first re-entry prediction. The best results in each column are in **bold**. Model BILSTM+BIA yields significantly better AUC and F1 scores than all other comparisons ($p < 0.05$, paired t-test).

In particular, models built on rules and features with shallow content and network features perform poorly, suggesting the need of better understanding of conversations or more information like user's chatting history. We also observe that HISTORY yields only slightly better results than RANDOM. It suggests that users' re-entries depend on not only their past re-entry patterns, but also the conversation context.

• *Well-encoded user chatting history is effective.* Among neural models, our BILSTM+MEM and BILSTM+BIA models outperform other comparisons by successfully modeling users' previous messages and their alignment with the topics of ongoing conversations. However, the opposite observation is drawn for BILSTM+CON and BILSTM+ATT. It is because the interactions between context and user history are effective yet complex, requiring well-designed merging mechanisms to exploit their joint effects.

• *Bi-attention mechanism better aligns the users' interests and the conversation topics.* BILSTM+BIA achieves the best AUC and F1 scores, significantly outperforming all other comparison models on both datasets. In particular, it beats BILSTM+MEM, which also able to learn the interaction between user history and conversation content, indicating the effectiveness of bi-attention over memory networks in this task.

Interestingly, comparing the results on the two datasets, we notice all models yield better recall and F1 on Twitter than Reddit. This is due to the fact that Reddit users are more likely to abandon conversations, reflected as the fewer number of entries in Table 1. Twitter users, on the other hand, tend to stay longer in the conversations, which encourages all models to predict the return of users.



(a) Twitter Dataset (b) Reddit Dataset

Figure 4: F1 scores for prediction on the first, second, and third re-entries (given the conversation context until the last entry). X-axis: # of turns in the given conversation context. Both figures, from left to right, show the F1 scores by ALL-YES, CCCT, BILSTM, BILSTM+MEM, and BILSTM+BIA.

5.2 Predicting Re-entries with Varying Context and User History

Here we study the effects of varying conversation context and user history over re-entry prediction.

Results with Varying Context. We first discuss model performance given different amounts of conversation context by varying the number of user entries. Figure 4 shows the F1 scores for predicting the first, second, and third re-entries. For predicting second or third re-entries, turns of current context until given user's second or third entry

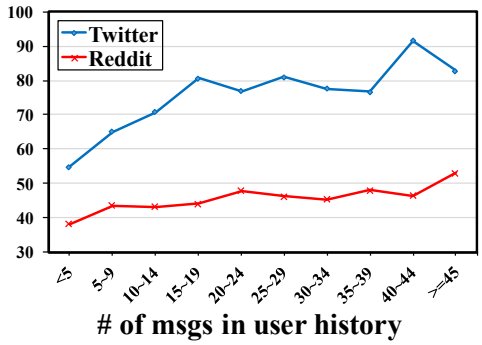


Figure 5: F1 scores of model BILSTM+BIA on first re-entry prediction, with varying numbers of chatting messages given in user history.

will be given. As can be seen, all models' performance monotonically increases when more context is observed. Our BILSTM+BIA uniformly outperforms other methods in all setups. Interestingly, baseline ALL-YES achieves the most performance gain when additional context is given. This implies that the more a user contributes to a conversation, the more likely they will come back.

Results with Varying User History. We further analyze how model performance differs when different amounts of messages are given in the user history. From Figure 5, we can see that it generally yields better F1 scores when more messages are available for the user history, suggesting the usefulness of chatting history to signal user re-entries. The performance on Reddit does not increase as fast as observed on Twitter, which may mainly because the context from Reddit conversations is often limited.

5.3 Further Discussion

We further discuss our models with an ablation study and a case study to understand and interpret their prediction results.

Ablation Study. To examine the contribution of each component in our framework, we present an ablation study on first re-entry prediction task. Table 3 shows the results of our best full model (BILSTM+BIA) together with its variant without using turn-level auxiliary meta $\mathbf{a}_t$ (defined in Section 3.1 to record user activity and replying relations in context), and that without structure modeling layer (to capture conversation discourse in context described in Section 3.2); also compared are variants without using user chatting history (described in Section 3.3).

Our full model yields the best F1 scores, showing the joint effects of context and user history can usefully indicate user re-entries. We also see that auxiliary triples, though conveying simple meta data for context turns, are helpful in our task. In addition, interestingly, conversation structure looks more effective in models leveraging user history, because they can learn deeper semantic relations between context turns and user chatting messages.

Models	Twitter			Reddit		
	F1	Pre	Rec	F1	Pre	Rec
W/O History						
W/O SML	58.8	42.6	**95.1**	39.6	25.2	**92.9**
With SML	59.4	45.9	85.0	43.1	31.4	69.5
With History						
W/O SML	57.5	43.2	86.7	43.8	31.3	74.4
W/O Meta	60.4	46.6	86.1	44.3	31.3	75.8
Full model	**61.1**	**47.0**	87.7	**45.4**	**33.9**	68.9

Table 3: Results of our variants. SML: structure modeling layer. Meta: auxiliary triples $\mathbf{a}_t$. Our full model BILSTM+BIA obtains the best F1.

Models	Conv. 1 ($\mathbf{C_1}$)	Conv. 2 ($\mathbf{C_2}$)
CCCT	1.0	1.0
BILSTM	0.386	0.480
BILSTM+MEM	0.583	0.712
BILSTM+BIA	0.460	0.581

Table 4: Predicted probabilities by different models for user $\mathbf{U_1}$'s re-entry to conversations $\mathbf{C_1}$ and $\mathbf{C_2}$ in Figure 1. CCCT can only yield binary outputs. For other neural models, predicting threshold is 0.5.

Case Study. We further utilize a case study based on the sample conversations shown in Figure 1 to demonstrate what our model learns. Table 4 displays the outputs from different models on estimating how likely $\mathbf{U_1}$ will re-engage in conversation 1 ($\mathbf{C_1}$) and conversation 2 ($\mathbf{C_2}$), where $\mathbf{U_1}$ returns to the latter. All neural models successfully forecast that $\mathbf{U_1}$ is more likely to re-engage in $\mathbf{C_2}$, while only BILSTM+BIA yields correct results (given threshold 0.5).

We further visualize the attention weights output by BILSTM+BIA's bi-attention mechanism with a heatmap in Figure 6. As can be seen, it assigns higher attention values to turns $\mathbf{T_2}$ and $\mathbf{T_3}$ in conversation $\mathbf{C_2}$, due to their topical similarity with user $\mathbf{U_1}$'s interests, i.e. movies, as inferred from their previous messages about *Let Me In*. The attention weights then guide the final prediction for higher chance of re-entry to $\mathbf{C_2}$ rather than $\mathbf{C_1}$.

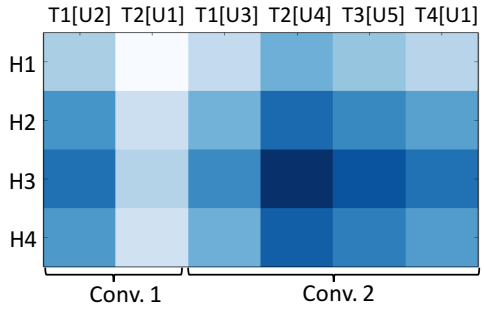


Figure 6: Attention output of model BILSTM+BIA for the two sample conversations in Figure 1.

Predictor	Twitter	Reddit
Human 1	26 (**29**)	30 (**30**)
Human 2	25 (**28**)	28 (**29**)
BILSTM+BIA	35	33

Table 5: Numbers of correct predictions made by humans, reading conversation context only and further seeing users' chatting history (boldfaced numbers), compared to the results of our best model in same setting. A random guess gives 25 (out of 50 pairs).

5.4 Comparing with Humans

We are also interested in how human performs for the first re-entry prediction task, in order to find out how challenging such a task is. To achieve this, we design a human evaluation. Concretely, from each dataset, we randomly sample 50 users who have been involved in at least 4 conversations, with both re-entry and non re-entry behaviors exhibited. Then for each user u, we construct paired samples based on randomly selected conversations c_1 and c_2, where u re-engage in one but not the other. The rest of the conversations that u participated in are collected as their user history. Then, we invite two humans who are fluent speakers of English, to predict which conversation user u will re-engage, after reading the context up to user's first participation in the paired conversations c_1 and c_2. They are requested to make a second prediction after reading user's chatting history.

Humans' prediction performance is shown in Table 5 along with BILSTM+BIA model's output on the same data. As can be seen, humans can only give marginally better predictions than a random guess, i.e., 25 out of 50 pairs. Their performance improves after reading the user's previous posts, however, still falls behind our model's predictions. This indicates the ability of our model to learn from large-scaled data and align users' interests with conversation content. In addition, we notice that humans yield better performance on Reddit conversations than Twitter. It might be due to the fact that Reddit conversations are more focused, and it is easier for humans to identify the discussion points. While for Twitter discussions, the informal language usage further hinders humans' judgment.

6 Conclusion

We study the joint effects of conversation context and user chatting history for re-entry prediction. A novel neural framework is proposed for learning the interactions between two source of information. Experimental results on two large-scale datasets from Twitter and Reddit show that our model with bi-attention yields better performance than the previous state of the art. Further discussions show that the model learns meaningful representations from conversation context and user history and hence exhibits consistent better performance given varying lengths of context or history. We also conduct a human study on the first re-entry prediction task. Our proposed model is observed to outperform humans, benefiting from its effective learning from large-scaled data.

Acknowledgements

This work is partly supported by HK RGC GRF (14232816, 14209416, 14204118), NSFC (61877020). Lu Wang is supported in part by National Science Foundation through Grants IIS-1566382 and IIS-1813341. We thank the three anonymous reviewers for the insightful suggestions on various aspects of this work.

References

Noor Aldeen Alawad, Aris Anagnostopoulos, Stefano Leonardi, Ida Mele, and Fabrizio Silvestri. 2016. Network-aware recommendations of novel tweets. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 913–916. ACM.

Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–606. Association for Computational Linguistics.

Lars Backstrom, Jon M. Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2013. Characterizing and curating conversation threads: expan-

sion, focus, volume, re-entry. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 13–22.

Ceren Budak and Rakesh Agrawal. 2013. On participation in group chats on Twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 165–176. ACM.

Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR Conference on Research and development in information retrieval*, pages 661–670. ACM.

Hao Cheng, Hao Fang, and Mari Ostendorf. 2017. A factored neural network model for characterizing online discussions in vector space. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2296–2306.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Liangjie Hong, Aziz S Doumith, and Brian D Davison. 2013. Co-factorization machines: modeling user interests and predicting individual decisions in Twitter. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, pages 557–566. ACM.

Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei. 2018. Find the conversation killers: A predictive study of thread-ending posts. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1145–1154. International World Wide Web Conferences Steering Committee.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 172–180.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Rui Yan, Mirella Lapata, and Xiaoming Li. 2012. Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 516–525. Association for Computational Linguistics.

Victoria Zayats and Mari Ostendorf. 2018. Conversation modeling on reddit using a graph-structured LSTM. *TACL*, 6:121–132.

Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. 2018. Microblog conversation recommendation via joint modeling of topics and discourse. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 375–385.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

Qi Zhang, Yeyun Gong, Ya Guo, and Xuanjing Huang. 2015. Retweet behavior prediction using hierarchical Dirichlet process. In *AAAI*, pages 403–409.