2021 IEEE Spoken Language Technology Workshop (SLT) | 978-1-7281-7066-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/SLT48900.2021.9383544

GETTING YOUR CONVERSATION ON TRACK: ESTIMATION OF RESIDUAL LIFE FOR CONVERSATIONS

Zexin Lu[1]*, Jing Li*[1]*, Yingyi Zhang*[2]*, Haisong Zhang*[3]

[1] Department of Computing, The Hong Kong Polytechnic University
[2] Nanjing University of Science and Technology [3] Tencent AI Lab

ABSTRACT

This paper presents a predictive study on the progress of conversations. Specifically, we **estimate the residual life for conversations**, which is defined as *the count of new turns to occur* in a conversation thread. While most previous work focus on coarse-grained estimation that classifies the number of coming turns into two categories, we study fine-grained categorization for varying lengths of residual life. To this end, we propose a hierarchical neural model that jointly explores indicative representations from the content in turns and the structure of conversations in an end-to-end manner. Extensive experiments on both human-human and human-machine conversations demonstrate the superiority of our proposed model and its potential helpfulness in chatbot response selection.

Index Terms— Conversation Understanding, Dialogue System, Social Computing, User Behavior Analysis, Natural Language Processing

1. INTRODUCTION

Conversations play an important role in opinion exchange and idea sharing in our daily life. We are involved in a wide variety of conversations every day, ranging from meetings for project collaboration to chitchats for forming our personal ideology. Being in these conversations, it sometimes occurs to us that the conversation is out of control. One example is the raise of red herrings that distracts the focus of a meeting and result in lengthy and meaningless arguments. Another example is the appearance of a conversation killer in an interesting and active chat that turns all other participants away and ruins their experience of being engaged.

In light of these concerns, there exits a pressing need to track the conversation progress [1, 2, 3] and advancing the user interaction experience [4, 5, 6]. It is hence interesting to investigate whether the progress of a conversation can be algorithmically predicted, given the first few turns. To that end, we approach this problem via estimating the conversations' **residual life**, which is defined as *how many new turns a conversation thread will receive* [4]. Specifically, following previous work [7] roughly dividing conversations into four stages, we define conversations' residual life in each stage to be very long, long, short, and very short. Foreseeing a conversation's progress will help one in doing the right things at the right time. For instance, when curing conversations, it is inappropriate to recommend ending discussions for users to be involved. Another promising application is on response selection [8, 9], where participants might want to forecast the risks in their responses that will inadvertently kill an active conversation. Particularly in human-computer interactions, our study can help chatbots in identifying responses that actively move a conversation forward. Without adopting such strategy, it is likely that a chatbot yield generic and boring responses, such as " I don't know" and "Me too" [10, 11, 12, 13], and thus turn human participants away.

[T_1]: I was a registered libertarian for 10 years. I left after 2008 financial meltdown, which proved conclusively we MUST have regulations.
[T_2]: interesting, how will having more rules stop people from committing crimes? Death penalty doesn't stop murder
[T_3]: great topic that has absolutely nothing to do with financial regulation. Any other non sequiturs?
[T_4]: no. Alot of good it did stopping the Wells Fargo fiasco tho
[T_5]: it did stop it. Don't you get that? Laws existed so they couldn't willfully continue. Which is my point. Lib would make that legal
...

Fig. 1: A Twitter conversation snippet. [T_i]: The i-th turn in the conversation snippet. There are nine new turns to occur.

To date, most progress made in related fields has been limited to the coarse-grained categorization for human-human conversations, such as the detection of "active" discussions [14, 4] and ended chats [5]; while we look at a wider range of conversation genres in both human-human and human-machine conversations, where a conversation's progress is estimated via fine-grained residual life prediction in four ordered categories. Such study, to the best of our knowledge, has never been explored before. Another line in previous research predicts user responses for individual social media messages, such as the number of replies or retweets [15], message diffusion patterns [16, 17, 18], etc. Different from them, we focus on response prediction at conversation level, where the entire context of a conversation is examined for estimating its future trajectory.

Jing Li and Zexin Lu are supported by PolyU Internal Fund: 1-BE2W.

To illustrate how the history contexts can affect residual life of conversations, Figure 1 displays a snippet of Twitter conversation about "financial regulation". From the snippet, it is observed that the conflicting opinions voiced via making statements (in T_1), showing doubts (in T_2), expressing disagreement and asking questions (in T_3), etc., result in the back-and-forth debate fashion, which in fact carries the discussion on for another nine turns. Thus we argue that *effective estimation of residual life requires an understanding on both turn content and conversation structure*. To this end, we propose a hierarchical neural model that jointly exploits the content representations of turns and the structure representations from turn interactions in an end-to-end manner. In contrast to most existing methods that rely on manually-crafted features, such as topology structure of conversations [17, 18], simple lexical statistics [15, 19], and social networks of users [14, 16, 4], our model does not require features from either manual design or external resource. Such capability ensures our generality in the scenarios where some certain information is unavailable. Moreover, our model explores two tasks simultaneously, one is to distinguish ongoing and ended conversations, and the other is to tackle fine-grained categorization for the residual life, where the latter one serves as our focus and is in a more challenging scenario.

To evaluate our proposed model, we experiment on both human-human and human-machine conversation datasets in our experiments. The results show that our model outperforms baselines based on handcrafted features. For example, our model achieves 48.0% accuracy on human-machine conversations, compared with 35.3% given by a prior model based on hand-coded features [4]. To better understand our superiority, a case study on Twitter conversations is provided and the results demonstrate that our model is able to capture indicative representations in the conversation history. More interestingly, we present a preliminary discussion on the correlation between the predicted residual life and manual annotation of the response quality and point out our potential to benefit response selection for chatbots.

2. PRELIMINARIES

2.1. Basic Notions for Conversations

We follow the definitions for common concepts of conversations from previous studies [20, 21]. The unit of a conversation is a **turn**, defined as an utterance given by one participant. Specifically, for most social media conversations [22], e.g., Twitter and online forums, a message being part of a discussion is considered as a turn. For human-machine conversations, a human-written prompt or a machine-generated response refers to a turn.

A sequence of turns forms a **conversation thread** where normally, for each two adjacent turns, the latter one *replies to* the previous one.[1] We then clarify this definition for two different cases. For multi-party conversations, e.g., most social media discussions, an entire conversation (with an original post and all its direct and indirect replies) is organized in tree structure [23], because a message may spark multiple replies. Under this circumstance, we consider a root-to-leaf path of such trees as a conversation thread. For conversations held between two participants, e.g., most human-machine conversations, the turns in a conversation thread can be modeled in the chronological order. For our task, a conversation thread serves as a data instance, and for human-machine conversations, their residual life only takes human turns into account as machines will always answer a human prompt.

Dataset	# of convs	Avg turns per conv	Avg length per turn	|Vocab|
Twitter	49,290	8.67	16.58	194,629
Movie	100,648	4.77	10.41	67,247
Wiki	40,890	4.05	38.87	118,111
ChatbotCN	34,270	7.09	5.74	35,393

Table 1: Statistics of datasets. # of convs: conversation count. Avg turns per conv: average turns per conversation. Avg length per turn: average word number per turn.

To track the progress of a conversation, we follow previous study [7] to assume that a conversation, from the greetings at the very beginning to the farewells at the closing, can be roughly segmented into four stages, each interpreted as **childhood**, **adolescence**, **adulthood**, and **old age** in its life cycle. Conversations in each stage in order further have their residual life to fall into one of the following categories: very short, short, long, and very long.

2.2. A Study on Conversation Data

To study residual life of real-life conversations, we conduct a pilot data analysis. Here we collect and investigate four conversation datasets, three of which are human-human conversations and the rest human-machine conversations. Statistics of the four datasets are shown in Table 1: 80% for training, 10% validation, and 10% test.

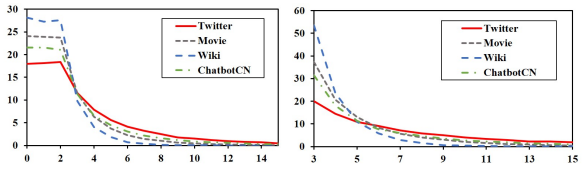
Data Collection. We collect three human-human conversation datasets, one from Twitter (henceforth **Twitter**), one movie scripts (henceforth **Movie**), and one Wikipedia talk-pages[2] (henceforth **Wiki**).

For Twitter, we first collected seed tweets initializing conversations using Twitter Streaming API[3] from January to December, 2016. Then, we used the names of authors and the IDs of seed tweets to locate the corresponding discussion pages and obtained the conversations via HTML page crawling and parsing. Finally, we recovered the missing messages

[1] In this paper, unless otherwise stated, a conversation is used as the short form for a conversation thread.

[2] https://en.wikipedia.org/wiki/Help:Talk_pages

[3] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview



(a) Residual Life (b) All Turns

Fig. 2: The turn distributions for conversations corresponding to residual life (on the left) and all turns (on the right). The historical axis shows the number of turns and the vertical axis indicates the proportion of conversations (%).

using Twitter search API[4] recursively with "in-reply-to" relations (the HTML pages only display partial conversations). The Movie dataset is released by [24], which contains fictional conversations held between two characters from movie scripts. It is close to off-line conversations held in our daily life [5]. The Wiki dataset is released by [4] consisting of editor discussions on Wikipedia projects and exhibiting working discussion styles. [5] In particular, as Twitter and Wiki conversations are multi-party conversations in tree structure, we randomly select a root-to-leaf path from each tree as a conversation thread as [5].

Besides talks among human participants, we also study human-machine conversations and collect a dataset from the chatting logs between anonymous users and a Chinese online chatbot (henceforth **ChatbotCN**), where users may chitchat on a wide range of topics. For each user, we segment the corresponding logs into varying conversations using timings via assuming that a new conversation is initialized if the user comes back after a long time.[6]

Residual Life Analysis. Here we further analyze on the data distributions of residual life on these real-life cases. Each thread is randomly cut into two parts: the *history* part, as observable context, and the *future* part, whose turn number is considered as the residual life. For human-machine conversations, we let the last turn in history to come from the machine and predicts how many human turns will be received. Afterwards, for each dataset, we study the residual life distributions on training data and the results are displayed in Figure 2a. We can observe a severe imbalance for varying numbers of future turns. To understand the cause of such imbalance, in Figure 2b, we show the distributions of the total turn numbers, including both the history and future turns in conversations.[7] We observe that only a small proportion of conversations can grow into lengthy discussions, which is consistent with the discoveries from previous studies [5, 26]. Based on the data

[4]https://developer.twitter.com/en/docs/tweets/search/overview/standard

[5]http://www.mpi-sws.org/ cristian/Echoes_of_power.html

[6]Assuming that users' response time for inter-conversation and intra-conversation turns satisfy two distinct Gaussian distributions, we assign the time spans between a machine turn and the next human turn into two clusters via Gaussian mixture model [25], one with smaller mean for inter-conversation spans, and the other intra-conversation spans.

[7]Conversations ended in 1-2 turns are not considered for better display.

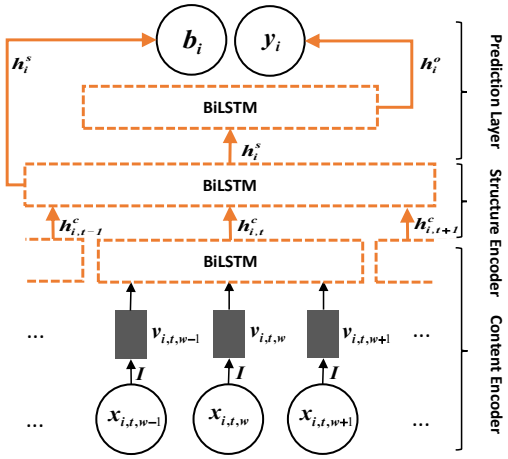


Fig. 3: The hierarchical BiLSTM model for estimating residual life categories of conversations. Here h refers to final state of their corresponding encoder cell.

distributions, we further determine our four residual life categories in the similar manner of [4] (used to separate "*active*" and "*inactive*" discussions). Specifically, we order instances by their residual life and divide them into four equal segments. For all the four residual life categories, i.e., very short, short, long, and very long, we determine their boundaries at 25%, 50%, and 75% of the instances in increasing order according to future turn numbers. In doing so, we can adapt residual life definitions to new data in varying distributions. Thus our framework can better fit diverse conversation genres, whose residual life distributions might be very different (as indicated by Figure 2a). For boundary cases, we assign them to one side of categories if more instances are found in the corresponding quantile.[8]

3. OUR MODEL FOR RESIDUAL LIFE ESTIMATION

To examine the conversation history for residual life estimation, our model employs a hierarchical Bidirectional Long Short-Term Memory (BiLSTM) network [27] and jointly explores the content of turns and the structure of conversations. Our overall architecture is illustrated in Figure 3.

Inputs and Outputs. Our model takes the input of a conversation $\mathbf{x}_i$ formulated as the sequence of its history turns: $\langle \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,|\mathbf{x}_i|} \rangle$, where $|\mathbf{x}_i|$ denotes the number of history turns in $\mathbf{x}_i$. Each turn $\mathbf{x}_{i,t}$ in $\mathbf{x}_i$ is formulated as a word sequence $\langle x_{i,t,1}, x_{i,t,2}, ..., x_{i,t,|\mathbf{x}_{i,t}|} \rangle$, where $|\mathbf{x}_{i,t}|$ is the number of words in turn $\mathbf{x}_{i,t}$ and $x_{i,t,w}$ denotes the w-th word in turn $\mathbf{x}_{i,t}$. Our final output y_i indicates the residual life category of conversation $\mathbf{x}_i$, where $y_i \in$ {very short, short, long, very long}.

[8]Boundary cases refer to instances shared in two adjacent quantiles. For example, if instances with zero and one future turn each holds 18% of the data. The instances with one future turn (at 25%) are the boundary cases for the first two quantiles. We assign them to the second category, i.e., short residual life, where $\frac{11}{18}$ (over 50%) of the instances are found.

Model Description. To jointly capture turn content and conversation structure, here we present our two BiLSTM models in hierarchical structure, one for content modeling and the other for structure modeling.

Content Modeling. The content representations are captured on turn level with a BiLSTM encoder, namely content encoder. Given the conversation turn $\mathbf{x}_{i,t}$, each word $x_{i,t,w}$ is represented as a embedding vector $\mathbf{v}_{i,t,w}$ with an embedding layer $I(\cdot)$, which is initialized by pre-trained embeddings and updated in the training. $\mathbf{v}_{i,t,w}$ is then fed into the content encoder and the learned representation is denoted as $\mathbf{h}^c_{i,t}$.

Structure Modeling. To learn structure representations for $\mathbf{x}_i$, which indicate the interaction between adjacent turns in its history, our model applies another BiLSTM, namely structure encoder. Its t-th state takes the content representation of the t-th turn $\mathbf{x}_{i,t}$ as input and the learned structure representation is denoted as $\mathbf{h}^s_i$.

Joint Prediction. Inspired by [28] (applying a multi-task learner for keyphrase extraction), our model owns two types of outputs in prediction layer and jointly tackles two tasks, one predicts whether there will be new turns and the other estimates the fine-grained residual life category. In other words, in addition to our final output y_i to produce residual life category, our model uses a binary output $b_i \in \{\textsf{ended}, \textsf{ongoing}\}$ to indicate whether $\mathbf{x}_i$ will carry on. In doing so, b_i would benefit to the prediction for conversations with many future turns, such as the example in Figure 1, because the prediction of $b_i = \textsf{ongoing}$ can strengthen the confidence of y_i to predict very long residual life for such conversations. For the similar reason, b_i can also help in predicting conversations with very short residual life. Formally,

$$b_i = softmax(\mathbf{h}^s_i) \tag{1}$$

where $\mathbf{h}^s_i$ is the structure representation of $\mathbf{x}_i$. To coordinate the two outputs, we first let $\mathbf{h}^s_i$ to serve as the input for the third BiLSTM to explore the hidden states $\mathbf{h}^o_i$. Then, we compute the final output by:

$$y_i = softmax(\mathbf{h}^o_i) \tag{2}$$

To further combine the joint effects of our two outputs, we define our final objective function as:

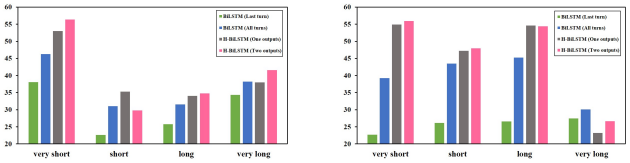
$$\mathcal{L}(\Theta) = \alpha \sum_{i=1}^{N} d(b_i, \hat{b}_i) + (1-\alpha) \sum_{i=1}^{N} d(y_i, \hat{y}_i) \tag{3}$$

where $\mathcal{L}(\cdot)$ is our loss function, Θ is the set of parameters, α is a hyperparameter for trading off the two effects, N denotes the count of instances, $d(\mathbf{x}, \mathbf{y})$ is the divergence measure between $\mathbf{x}$ and $\mathbf{y}$ (here we use cross entropy), and $\hat{b}_i$ and $\hat{y}_i$ denote the gold-standard category labels.

4. EXPERIMENTS

Setup. Here we describe how we setup our experiment.

Data Preprocessing. For English datasets, i.e., Twitter, Movie, and Wiki, we used Stanford NLP toolkit [29] for tokenization and lemmatization.[9] For Chinese (ChatbotCN), we applied NLPIR tool [30] for Chinese word segmentation.[10]



(a) Twitter (b) ChatbotCN

Fig. 4: F1 scores for the conversations with varying residual life categories. Horizontal axis: residual life categories from very short to very long. Vertical axis: F1 scores (%). For each category, from left to right shows the results of BiLSTM (*Last turn*), BiLSTM (*All turns*), H-BiLSTM (*One output*), and H-BiLSTM (*Two outputs*).

Comparison Models. We first consider a weak baseline majority vote (assigning the major labels in training set to all the test instances). Then, we employ logistic regression (LR) [31] and support vector machine (SVM) [32] with features proposed in [4] and [5]. For LR and SVM, we test two versions: one with features extracted from the last turn (henceforth LR (*Last turn*) and SVM (*Last turn*)), and the other from the entire history (henceforth LR (*All turns*) and SVM (*All turns*)). Similar models are built with BiLSTM: BiLSTM (*last turn*) and BiLSTM (*All turns*), where the latter model takes a long word sequence constructed by chronologically ordered turns in conversation history.

In addition, we compare with a variant of our model, i.e., H-BiLSTM (*One output*), which contains only one output for predicting a conversation's residual life category. For convenience, our full model with two outputs (b_i and y_i) will be referred to as H-BiLSTM (*Two outputs*).

Model Settings. All hyperparameters are turned on development sets by grid search. For BiLSTM models, we set their state size of each direction to 150, RMSProp [33] as the optimizer for parameter updating, and the trade-off parameter α to 0.5 for balancing b_i and y_i. Pre-trained embeddings are used. For Twitter, we employ the embeddings learned from a collection of 99M tweets. For Wiki and Movie, we use the embeddings released by [34].[11]. For Chinese ChatbotCN dataset, word embeddings are pre-trained on 467M posts from Weibo (a Chinese social media platform). We also tested embeddings pre-trained with standard RoberTa, which didn't provide much performance gain. It is probably because non-trivial designs are needed to adapt them to social media data (noisy and colloquial), which is beyond the scope of this paper and we leave the adaption work to future studies.

Residual Life Estimation Results. We show the main comparison results for residual life categorization in Table 2, where we report accuracy and average F1 scores for the four

[9] https://github.com/stanfordnlp/CoreNLP
[10] https://github.com/NLPIR-team/NLPIR
[11] https://spinningbytes.com/resources/word-embeddings/

	Twitter		Movie		Wiki		ChatbotCN	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Comparison models								
Majority Vote	13.3	36.3	11.2	28.8	10.9	27.8	12.0	31.6
LR (*Last turn*)	22.0	25.8	24.8	28.1	27.8	31.9	24.1	25.2
LR (*All turns*)	26.7	27.2	28.4	31.1	32.7	36.9	30.2	35.3
SVM (*Last turn*)	17.5	36.8	9.7	23.8	17.4	29.7	21.1	24.4
SVM (*All turns*)	25.1	35.0	28.9	33.1	32.8	39.9	24.6	26.9
BiLSTM (*Last turn*)	30.2	30.3	28.6	29.3	33.2	36.3	25.7	25.9
BiLSTM (*All turns*)	36.7	35.6	36.2	37.4	42.0	45.2	39.5	39.6
Our models								
H-BiLSTM (*One output*)	40.1	41.0	44.2	49.0	45.6	54.5	45.0	47.5
H-BiLSTM (*Two outputs*)	**41.1**	**42.1**	**46.9**	**49.3**	**53.2**	**64.3**	**46.2**	**48.0**

Table 2: Classification results of the four categories of residual life, where Acc refers to accuracy and F1 denotes the average F1 scores over the four residual life categories (%).

possible outcomes. The following observations are drawn:

• ***Manually-crafted features are not enough.*** SVM or LR models with manually-crafted features yield generally worse results than neural models. It means that conversations' residual life estimation is challenging and impossible to rely on hand-coded features or rules.

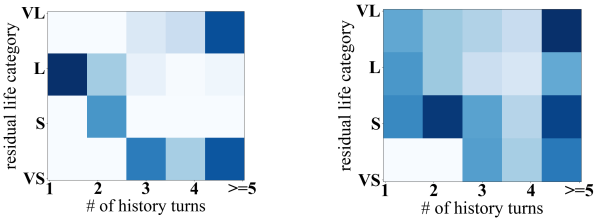
• ***History information is important.*** LR and SVM perform better when they are combined with rich history features. Similar observations can be seen for neural models where H-BiLSTM and BiLSTM (*All turns*) produce better results than BiLSTM (*Last turn*), which only relies on the content of the last turn.

• ***Jointly modeling of content and structure is effective.*** By jointly learning representations from turn content and conversation structure, the H-BiLSTM models achieve better results than the BiLSTM (*All turns*) model. This demonstrates that both turn content and conversation structure are useful in indicating residual life of conversations.

• ***Multi-task learning helps each other.*** The results of H-BiLSTM (*Two outputs*) are better than H-BiLSTM (*One output*) on all datasets. This indicates the effectiveness to simultaneously tackle the two tasks with shared parameters, because they are highly related to each other.

To further investigate the model performance over varying residual life categories, we select four models: BiLSTM (*Last turn*), BiLSTM (*All turns*), H-BiLSTM (*One output*), and our H-BiLSTM (*Two outputs*), given relatively better performance in Table 2. Their F1 scores in predicting residual life ranging from very short to very long are shown in Figure 4. We see the two H-BiLSTM models have consistently better performance than others, which again shows the joint effects of content and structure to conversations' residual life. We also find that the H-BiLSTM (*Two outputs*) tends to outperform H-BiLSTM (*One output*) for conversations with very short and very long residual life. The possible reason is that the "yes" prediction of new turns (b_i = *ongoing*) helps increase model's confidence to predict very long residual life for conversations, so does the very short cases.

Effects of Conversation History. Results in the previous discussions show the usefulness of conversation history. Here we take Twitter conversations as an example to further analyze how it affects the residual life.



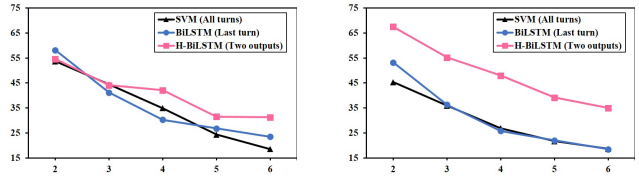
(a) Gold-standard Results (b) Predicted Results

Fig. 5: Heatmaps showing the turn number correlations of history and future in Twitter conversations. The left one shows the gold-standard results and the right one shows the predictions. Horizontal axis: the # of history turns. Vertical axis: the category of residual life where VS, S, L, and VL indicates very short, short, long, and very long residual life. Darker color in (x, y) indicates more conversation instances containing x turns in the history and y turns in the future.

First, we quantitatively analyze the residual life distributions for conversations with varying turns in their history. Such distributions are visualized via the heatmaps in Figure 5. The left one shows the gold-standard distributions and the right the results predicted by our H-BiLSTM (*Two outputs*) model. Their similar color patterns demonstrate that our predicted distributions roughly fit with the real. We also observe that for a dark grid in the gold-standard heatmap, its upper or lower neighbor of the corresponding grid in the predicted heatmap tends to be highlighted. This shows the particular challenge to distinguish adjacent categories, such as short and very short residual life.

From the gold-standard heatmap, we also find something interesting. For the conversation history ≥ 5 turns, there are two possible outcomes indicated by the darkest grids: very short or very long residual life. This implies that most conversations with a long history either end soon (maybe because users get tired of being engaged) or they would possibly grow into heated debates and thus have very long residual life. Differently, for the conversations with only one history turn, they tend to have long residual life because these conversations are in relatively early stages.

To better understand how the history and residual life are related, we conduct a case study on the Twitter conversation in Figure 1. Recall that the conversation ,with a tenor of argument, does not end until nine turns later, whose residual life should be categorized as very long. The BiLSTM (*All turns*) outputs short for it because it is unable to explore conversation structure and capture the argumentative fashion presented by turn interactions. BiLSTM (*Last turn*) yields a closer answer with long residual life. It may notices the rhetorical question in the last turn "Don't you get that?". Such content is likely to move a discussion forward and ignored by BiLSTM (*All turns*) entrapped with other information. By exam-



(a) Twitter (b) ChatbotCN

Fig. 6: Estimation accuracies given varying granularity of residual life (%). Historical axis: the total number of categories. Vertical axis: the corresponding accuracy. H-BiLSTM (*Two outputs*) performs consistently better.

ining turn interactions, H-BiLSTM (*One output*) also predict long residual life as it learns useful features from conversation structure. Nevertheless, only H-BiLSTM (*Two outputs*) successfully predicts very long residual life, because the prediction of ongoing from b_i makes it become more confident to predict very long residual life.

Residual Life in Varying Granularity. In the aforementioned discussions, we focus on residual life with four categories. Here we discuss the estimation results on varying granularity of residual life categories. To this end, we test the performance of SVM (*All turns*), BiLSTM (*Last turn*), and our H-BiLSTM (*Two outputs*) when estimating residual life with two to six categories (defined similarly in preliminaries). The accuracies on Twitter and ChatbotCN are shown in Figure 6, where the parallel decrease curves indicate the increasing difficulty to estimate residual life categories with finer granularity. We also find that our H-BiLSTM (*Two outputs*) produces consistently better accuracies and shows its effectiveness to estimate varying residual life granularity.

Residual Life vs. Response Selection. To provide more insights, here we present a preliminary discussion on the correlation between the estimated residual life and the manually annotated quality of responses. To this end, we first follow the procedure in [35] to collect a 10K prompt-response pairs from Weibo, where a prompt-reply pair refers to a Weibo post and one of its replies. We then invite two experienced annotators to label the quality of each reply as bad and good, where bad replies are off-topic or incoherent to the prompt, and good should be assigned to on-topic and interesting responses. Later, for each quality level, we sample 1K responses with the corresponding label agreed by both annotators. Based on these selected data, we apply our H-BiLSTM (*Two outputs*) trained on the ChatbotCN dataset to estimate the residual life of conversations with two turns in history, i.e., a prompt and its reply. We then measure the proportions of bad and good responses with varying predicted categories for their residual life. In the results, less than 10% of the instances are predicted to have long or short residual life. It may be ascribed to the difficulty to distinguish these two categories from others given such short history with two turns. For the rest two categories, i.e., very short and very long residual life, we show the results in Figure 7. As can be seen, our model tends

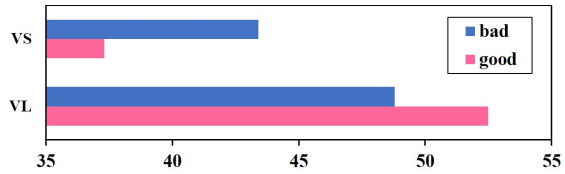


Fig. 7: The proportions of bad and good responses predicted to have very long (VL) and very short (VS) residual life (%). For each category, the upper and lower bar shows the results for bad and good responses, respectively.

to estimate longer residual life for responses with better quality. The observation implies that the estimated residual life may serve as automatic annotations for response quality and useful features to train dialogue systems.

5. RELATED WORK

In previous work, there are studies analyzing the number of retweets or replies for social media messages [4, 15, 14, 36, 37], which focus on human engagements on social media and measuring various of features. Distinguished from these studies, our work does not rely on a labor-intensive process of feature engineering and provides an alternative with neural models for this task. More importantly, in addition to human-human conversations, we also investigate the residual life for human-machine conversations as well as its application on dialogue response selection for chatbots, which is, to our best knowledge, the very first research of its kind. Although there are recent studies on thread ending posts on social media [5], they only investigate binary prediction of ended conversations. Different from them, we focus on fine-grained categorization of future turn numbers, which is beyond a simple "yes or no" answer to whether new turns will be received.

Our work is also related to state tracking in conversations [38], e.g., the prediction of user engagement degree [39, 40, 41, 42]. These studies measure speech features and involve human annotation for engagement degree. Instead, our approach does not require such features and can be conducted without manually annotated labels, which enables its ability to be scaled for large datasets.

6. CONCLUSION

We have presented a framework for estimating four categories of conversations' residual life, corresponding to varying stages in conversation progress. To tackle this task, a hierarchical neural model has been proposed to jointly learn representations from the content of each turn and the structure of turn interactions. Experimental results on both human-human and human-machine conversations show that our model is able to capture indicative features from conversation history and thus give superior performance. A further study shows the potential of the predicted residual life in benefiting response quality annotation for chatbots.

7. REFERENCES

[1] Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou, "A contextual hierarchical attention network with adaptive objective for dialogue state tracking," in *ACL*, 2020, pp. 6322–6333.

[2] Sanuj Sharma, Prafulla Kumar Choubey, and Ruihong Huang, "Improving dialogue state tracking by discerning the relevant context," in *NAACL*, 2019, pp. 576–581.

[3] Bo-Hsiang Tseng, Marek Rei, Pawel Budzianowski, Richard E. Turner, Bill Byrne, and Anna Korhonen, "Semi-supervised bootstrapping of dialogue state trackers for task-oriented modelling," in *EMNLP*, 2019, pp. 1273–1278.

[4] Lars Backstrom, Jon M. Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil, "Characterizing and curating conversation threads: expansion, focus, volume, re-entry," in *WSDM*, 2013, pp. 13–22.

[5] Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei, "Find the conversation killers: A predictive study of thread-ending posts," in *WWW*, 2018, pp. 1145–1154.

[6] Hao Cheng, Hao Fang, and Mari Ostendorf, "A dynamic speaker model for conversational interactions," in *NAACL*. 2019, pp. 2772–2785, Association for Computational Linguistics.

[7] Anne Tissen, "A case-based architecture for A dialogue manager for information seeking," in *SIGIR*, 1991, pp. 152–161.

[8] Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura, "Conversational response re-ranking based on event causality and role factored tensor event embedding," *CoRR*, vol. abs/1906.09795, 2019.

[9] Bo Zhang and Xiaoming Zhang, "Hierarchy response learning for neural conversation generation," in *EMNLP*, 2019, pp. 1772–1781.

[10] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma, "Topic aware neural response generation," in *AAAI*, 2017, pp. 3351–3357.

[11] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil, "Generating high-quality and informative conversation responses with sequence-to-sequence models," in *EMNLP*, 2017, pp. 2210–2219.

[12] Philippe Laban, John Canny, and Marti A. Hearst, "What's the latest? A question-driven news chatbot," in *ACL*, 2020, pp. 380–387.

[13] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazaré, and Jason Weston, "Learning from dialogue after deployment: Feed yourself, chatbot!," in *ACL*, 2019, pp. 3667–3684.

[14] Liangjie Hong, Ovidiu Dan, and Brian D. Davison, "Predicting popular messages in twitter," in *WWW*, 2011, pp. 57–58.

[15] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *SocialCom*, 2010, pp. 177–184.

[16] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev, "Prediction of retweet cascade size over time," in *CIKM*, 2012, pp. 2335–2338.

[17] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei, "Deepcas: An end-to-end predictor of information cascades," in *WWW*, 2017, pp. 577–586.

[18] Jia Wang, Vincent W. Zheng, Zemin Liu, and Kevin Chen-Chuan Chang, "Topological recurrent neural network for diffusion prediction," in *ICDM*, 2017, pp. 475–484.

[19] Matthew Rowe, Sofia Angeletou, and Harith Alani, "Anticipating discussion activity on community forums," in *SocialCom*, 2011, pp. 315–322.

[20] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the organization of conversational interaction*, pp. 7–55. Elsevier, 1978.

[21] Lei Shen, Yang Feng, and Haolan Zhan, "Modeling semantic relationship in multi-turn conversations with hierarchical latent variables," in *ACL*, 2019, pp. 5497–5502.

[22] Tatjana Scheffler, Berfin Aktaş, Debopam Das, and Manfred Stede, "Annotating shallow discourse relations in twitter conversations," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking*, 2019.

[23] Jing Li, Ming Liao, Wei Gao, Yulan He, and Kam-Fai Wong, "Topic extraction from microblog posts using conversation structures," in *ACL*, 2016.

[24] Cristian Danescu-Niculescu-Mizil and Lillian Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *CMCL@ACL*, 2011, pp. 76–87.

[25] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *International Conference on Acoustics, Speech, and Signal Processing*, Apr 1990, pp. 293–296 vol.1.

[26] Yi Chang, Xuanhui Wang, Qiaozhu Mei, and Yan Liu, "Towards twitter context summarization with user influence models," in *WSDM*, 2013, pp. 527–536.

[27] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," in *ACL*, 2015, pp. 1106–1115.

[28] Qi Zhang, Yang Wang, Yeyun Gong, and Xuanjing Huang, "Keyphrase extraction using deep recurrent neural networks on twitter," in *EMNLP*, 2016, pp. 836–845.

[29] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky, "The Stanford CoreNLP natural language processing toolkit," in *ACL*, 2014, pp. 55–60.

[30] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu, "HHMM-based Chinese Lexical Analyzer ICTCLAS," in *SIGHAN*, 2003, pp. 184–187.

[31] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant, *Applied logistic regression*, vol. 398, John Wiley & Sons, 2013.

[32] Vladimir Cherkassky, "The nature of statistical learning theory," *IEEE Trans. Neural Networks*, vol. 8, no. 6, pp. 1564, 1997.

[33] Alex Graves, "Generating Sequences With Recurrent Neural Networks," *arXiv pre-print*, vol. abs/1308.0850, 2013.

[34] Mark Cieliebak, Jan Deriu, Dominic Egger, and Fatih Uzdilli, "A twitter corpus and benchmark resources for german sentiment analysis," in *SocialNLP@EACL*, 2017, pp. 45–51.

[35] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen, "A dataset for research on short-text conversations," in *EMNLP*, 2013, pp. 935–945.

[36] Yoav Artzi, Patrick Pantel, and Michael Gamon, "Predicting responses to microblog posts," in *NAACL*, 2012, pp. 602–606.

[37] Matthew Rowe, Sofia Angeletou, and Harith Alani, "Predicting discussions on the social semantic web," in *ESWC*, 2011, pp. 405–420.

[38] Yawen Ouyang, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujian Huang, and Jiajun Chen, "Dialogue state tracking with explicit slot connection modeling," in *ACL*, 2020, pp. 34–40.

[39] Zhou Yu, Dan Bohus, and Eric Horvitz, "Incremental coordination: Attention-centric speech production in a physically situated conversational agent," in *SIGDIAL*, 2015, pp. 402–406.

[40] Zhou Yu, Leah Nicolich-Henkin, Alan W. Black, and Alexander I. Rudnicky, "A wizard-of-oz study on A non-task-oriented dialog systems that reacts to user engagement," in *SIGDIAL*, 2016, pp. 55–63.

[41] Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft, "An open-source dialog system with real-time engagement tracking for job interview training applications," *Proceedings of IWSDS*, 2017.

[42] Divesh Lala, Koji Inoue, Pierrick Milhorat, and Tatsuya Kawahara, "Detection of social signals for recognizing engagement in human-robot interaction," *arXiv preprint*, vol. abs/1709.10257, 2017.