MACHINE-AIDED ONLINE USER ENGAGEMENTS

LU ZEXIN

PhD

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University
Department of Computing

Machine-Aided Online User Engagements

Lu Zexin

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
March 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Lu Zexin

Abstract

In view of the widespread use of social platforms, interpersonal communications have come to play an increasingly crucial role in our daily activities. Nevertheless, although every individual is part of our society, many of them have not yet gathered the ability to socialize well with others. For these people, they may unintentionally ruin a conversation or be reluctant to voice their opinions. To help them perform in social interactions better, this thesis proposes novel solutions to employ data-driven natural language processing (NLP) methods to provide support and guidance to users for them to better engage in online social interactions.

To that end, we first measure the understanding ability of NLP models on the user-generated content on social media. Specifically, we present the first benchmark to investigate how well the state-of-the-art natural language understanding (NLU) models tackle social media tasks, where the texts usually exhibit the inherent noise (e.g., informal writings) underlying the user-generated contents. To build the benchmark, we gather two large-scale Chinese datasets from Weibo — 80K posts with crowd-sourcing annotations and 3K posts with expert annotations for three fundamental tasks (Chinese word segmentation, part-of-speech tagging, and named-entity recognition) to examine how well models gain the generic language understanding. In addition, model performance on popular social media applications, such as rumor detection, emoji prediction, sentiment analysis, and hashtag classification, are examined to investigate NLU models' capability of capturing specific semantics from social media messages. The experimental results demonstrate the

effectiveness of trendy language encoders from the BERT family to to understand social media messages, which even obtained better results than human readers.

Then, we examine user participants' behavior in conversations via estimating their effects on the residual life for conversations, which is defined as the count of new turns to occur in a conversation thread. While most previous work focuses on the 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 the potential NLP models to evaluate the engaging degree of user discussions.

At last, we research how to actively draw the engagement of users who prefer not to comment with words. A novel task is proposed to generate vote questions for social media posts. It offers an easy way to hear the voice of the public and learn from their feelings about important social topics. While most related work tackles formal languages (e.g., exam papers), we generate vote questions for short and colloquial social media messages exhibiting severe data sparsity. To deal with that, we propose to encode user comments and discover latent topics therein as contexts. They are then incorporated into a sequence-to-sequence (S2S) architecture for question generation and its extension with dual decoders to additionally yield vote answers. For experiments, we collect a large-scale Chinese dataset from Sina Weibo. The results show that our model outperforms the popular S2S models without leveraging topics from comments and the dual decoder design can further benefit the prediction of both questions and answers.

Publications arising from the thesis

1. **Zexin Lu**, Keyang Ding, Yuji Zhang, Jing Li, Baolin Peng, and Lemao Liu. "Engage the Public: Poll Question Generation for Social Media Posts", In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, Association for Computational Linguistics.

2. **Zexin Lu**, Jing Li, Yingyi Zhang, and Haisong Zhang. "Getting your conversation on track: Estimation of residual life for conversations", In IEEE Spoken Language Technology Workshop, SLT 2021, IEEE.

3. **Zexin Lu**, Keyang Ding, Zhe Wang, Jing Li, Lemao Liu. "A Chinese Benchmark for Social Media Language Understanding", under review, submitted to The 29th International Conference on Computational Linguistics, Coling 2022.

Acknowledgements

Firstly I would like to express heartfelt thanks to my supervisor Dr. LI Jing for her outstanding support and guidance during my Ph.D. study. Her personality (kind-hearted and incredible patience) and profound knowledge have had a lifelong impact on me.

Also, I would like to thank my co-supervisor (Prof. Li Qing), my defense BOE chair (Prof. Li Wenjie), confirmation panel member (Dr. Luo Xiapu), ex-supervisor (Prof. Guo Song), my intern mentor in Tencent Ai Lab (Dr. Liu Lemao), and the external defense reviewers for their effort in my road to graduation.

Then, my sincere thanks go to the professors and teachers in our department, Dr. Wu Xiaoming, Dr. Yang bo, Dr. Yang Lei, etc., for their suggestions and help in my study and life.

Besides, I'd like to take this opportunity to thank my co-authors and my group members (Dr. Peng Baolin, Dr. Xiang Rong, Mr. Ding Keyang, Ms. Zhang Yuji, Mr. Zhang Yubo, Mr. Tan Hanzhuo, Mr. Wang Zhe, and Ms. Pan Yalu. Throughout the journey to becoming a Ph.D., I have received a great deal of your support and assistance.

Last but not least, I would like to thank my dear friends. All of you make a beautiful memory in this Ph.D. journey. Dr. Li Zhe, Prof. Zhang Yinyan, Dr. Li Yanran, Dr. Chen Jiaxin, Dr. Xu Zhou, Dr. Wang Fang, Dr. An Zhenlin, Dr. Lin Qiongzhen, Ms. Pan Qinrui, Dr. Guo Jingcai, Dr. Hou NingNing, Ms. Cui Kaiyan, Ms. Zhang Jie, Mr. Wu Haotian, Mr. Yang Qiang, Mr. He Qingqiang, Dr. Ma Chenlin, Dr. Li Zhuo, Dr.

Tsz Nam Chan, Dr. Yang Yanni, Mr. Li Qimai, Mr.Zhan Liming, Mr. Liu Bo, Dr. Li Lida, Mr. Zhang Xindong, Ms.Cheng Haiming, Dr. Xue Lei, Dr. Yu Le, Mr.Xiang Haodong, Mr. Jin A long, Dr. Xu Wenchao, Ms. Liu Xiaofei, Mr. Zhou Qihua, Prof. Milos, Dr. Xie Xin, Dr. Qu Zhihao, Dr. Tang Bin, Dr. Wang Haozhao, Dr. Zhan Yufeng, Mr.Hu Shengdun, Dr. Wang Yue, Mr. Meng Ziqiao, Mr. Liu Yibin, Ms. Chen Yi, Mr. Zhang Yuhan, Mr. Du Xuefeng, Mr. Liu Ziquan, Mr. Chen Yi, Dr. Chen Tao, Mr. Wang Bin, Dr. Chen Junyang, Dr. Zhang Tong, Mr. Lan Fengbo, Mr. Yin Tengfei, Mr. Feng Yiwen, Dr. Sun Ruqi, Dr. Wen Jiaxin, Dr. Xu Chufan, Dr. Jiang Yuechi, Ms. Wen Xi, Dr. Zheng Zimu, Mr. Zhu Qingling, Mr. He Xingwei, Mr. Fang Xianghong, Ms. Luo Yan, etc.

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

The last decade has witnessed the flourish of Internet. It broadly affects people's daily life attributed to the substantial advances made in both the infrastructure (e.g., the speed-up of mobile networks) and applications (e.g., the technology of instant messaging). Consequently, many of people's everyday activities and interpersonal communications have been gradually moved to the online world, such as meetings and chitchats, largely benefit from the development and implementation of the Internet.

In context of the worldwide expansion of the Internet, social media platforms play essential roles in our daily connections to others, such as the opinion exchange from diverse educational and cultural backgrounds and the sharing of information that is breaking or widely discussed. Social media has essentially revolutionized our living manners. Millions of users are turning to micro-blogging platforms, such as Sina Weibo,

Twitter, and Facebook, to share ideas with friends and voice viewpoints to the public; especially during the lockdown period of COVID-19 crisis, people have to stay at home and social media is the only outlet to allow individuals to stay in touch with this world.

1.1 Motivation

Although every individual is a part of our society, many of them have not yet gained the capability to socialize with others well. Some people are willing to discuss but fail to engage appropriately. For example, they may unintentionally kill a conversation because of the improper behavior in the discussion, e.g., raising an irrelevant point or dropping wordy and boring messages. Meanwhile, many others, for various reasons (e.g., the introverted personality), may not be well-motivated to explicitly voice their opinions, rendering it difficult for them to navigate social life online from the very beginning. All such cases would result in unwanted social experiences and as a worse consequence, those people may refuse to socialize with others afterwards ascribed to the negative effects of these unpleasant experiences.

Under this circumstance, it is vital to encourage the effective and engaging interpersonal communications on social media while many individuals have not been well trained to gain the useful people skills.

The possible reason is that in our education system, especially in the Asian area, limited attention has been paid to teach students how to socialize with others well.

On the other hand, social media platforms also provide us with rich resources to learn how to enable positive user engagements, such as the engaging conversations with successful outcomes. It is potential to make good use of those materials, such as other people's chatting history and high-quality comments, to help individuals gain a better engagement in social media discussions. Considering the good capability of Natural Language Processing (NLP) models in digesting big data and distilling salient content therein, in this thesis, we propose to employ NLP technology to leverage large-scale social media data, automatically learn from large volume of users' social behavior, and automatically assist individuals's engagement in social media conversations.

1.2 Challenges

In this subsection, we discuss the challenges of using NLP techniques in the modeling of user engagements on social media. First of all, models should help users understand others' messages, whereas the inherent noise exhibited in the user-generated data on online social network platforms may inevitably hinder NLP models from understanding the context

and thus worsen the performance of the applications. Second, the online social environment is complex and uncontrollable because participants might vary in their personality and purposes. For instance, good interactions may become out of control when a conversation killer appears. Third, some people may inherently not want to explicitly voice in online discussions for many reasons, e.g., the introverted personality, and how to engage them into social interactions hence becomes a concrete challenge. In the following, we will discuss the challenges in details with the corresponding solutions presented in Chapter (3, 4, 5) and the contributions summarized in 1.3.

1.2.1 Special Tokens and Data Noisiness

Social media exhibits an informal and colloquial language style, where we can observe the prominent use of special social media tokens (e.g., emojis, slangs, and special punctuation). These tokens, either intentionally (e.g., abbreviation, slang, and exaggeration) or unintentionally used (e.g., typo and missing blank), present the concrete challenge for the language understanding models to capture the essential meanings well, because it usually requires the prior knowledge (e.g., common sense and culture understanding) probably absent in the limited context of a short piece of text.

The cutesssst DOG ever lol, RT if U likeit

Figure 1.1: An example tweet with informal language styles. The user types *cutesssst* to emphasize the degree of cuteness for the dog. *lol* is a slang that means happy. *RT* stands for re-tweet, *U* is slang or abbreviation of YOU. *likeit* could be the missing blank.

In Figure 1.1, as can be seen, to understand the meaning of the tweet, one should priorly know how to interpret the abbreviation, slang, exaggeration, missing blank, etc. All these factors will result in a deviated and much larger vocabulary compared to that of formal texts, e.g., news articles. In other words, as the data sample increases, the vocabulary table will dramatically expanded and the word distribution over the vocabulary will be sparse, whereas the existing NLP models largely rely on the rich and dense context to gain the generic language understanding capability.

A: I got 99 points in the Calculus class!

B: 

Figure 1.2: A sample tweet requiring cultural background for understanding. The emoji smile is a polysemy, even antonym, in Chinese culture.

What's more, the understanding of some social media tokens (e.g., emojis or memes) further requires background knowledge. For example, in Figure 1.2, the emoji  conventionally has the meaning of smile or happy. However, many young people in China today prefer to use

it for expressing sarcasm and scoff. Conditioned on the age of user B, the reply may convey very different meanings – older person may simply want to congratulate user A's achievement while younger person may sneer at user A because of jealousy. Therefore, without the prior knowledge of such culture, the NLP models might be confused about how to correctly understand the smile face in the reply.

To examine this challenge, we build the first social media benchmark to investigate the NLP model's understanding ability and the discussion will be present in Chapter 3.

1.2.2 Uncontrollable Environments in Conversations

In online conversations, it is very likely that discussion may go awry and the environments may become out of control. For example, the occurrence of conversation killers may end a heated debate because of the unappropriated wordings. On the other hand, the rise of red herrings probably distracts the focus of a meeting and further results in a lengthy and unpleasant discussion.

As the toy example shown in Figure 1.3, the boy is a conversation killer in the discussion with a girl he wants to date. The girl shared her personal experience at the beginning and raised a question to move the conversation forward. However, the boy simply said "I don't know",

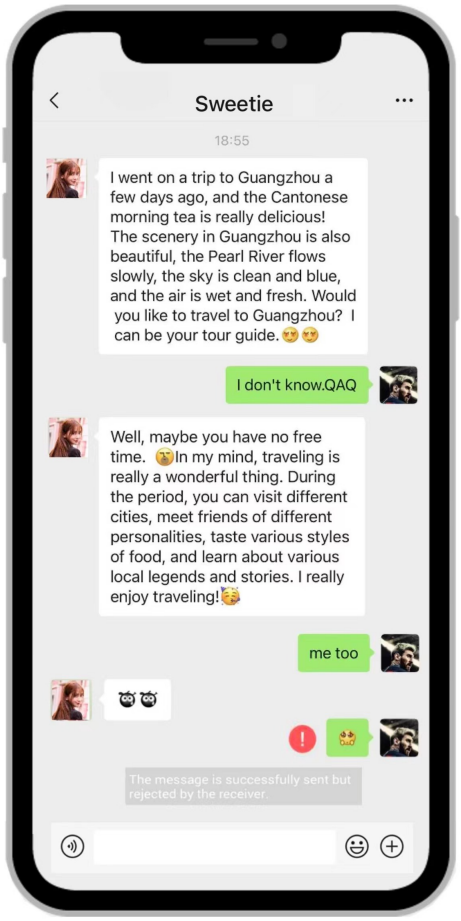


Figure 1.3: An example conversation with a boy participant mistakenly kill a conversation because of the inappropriate response. He does not know how to well respond the girl's prompts and then turned the girl away.

which tends to end up the conversation. When the conversation continues with this pattern, the girl started to become bored and did not want to talk with him anymore. It is because the girl starts to feel uncomfortable in the conversation discourse. Few people would like to continue the talk in this unpleasant environment. However, the boy probably wanted to talk more with the girl. Because of his introverted personality, he didn't know how to carry on the discussion without turning the girl away.

As can be seen above, the conversation environments might be complicated and uncontrollable, which might be challenge the models to be on track of what is happening in the context. To address this problem, we propose to estimate the residual life (which is defined as how many new messages a conversation thread will continue) for conversations in Chapter 4 and guide users to communicate in a proper way.

1.2.3 Reluctance to Online User Engagements

In the previous subsection, we discussed the scenarios where users attempted to engage in online conversations yet somehow faced some difficulties in keeping the conversation on track. Here we focused on a more challenging scenario where users may be reluctant to explicitly voice their opinions in online discussions. For example, people with the physiological preference of extroversion are willing to spend time on communication and interaction with the outside world. On the contrary, introverted people are described as reserved and solitary. They're energetic on reflection or thinking, i.e., focusing on their inner world, whereas would be uncomfortable when communicating with others.

On social media platforms, people do not need to talk face-to-face. It gives an excellent chance for introverted people to express their ideas and communicate with others behind their computers or cell phones.

However, some introverted people may still be reluctant to drop messages explicitly. Instead, these people may prefer to read others' posts or information rather than share viewpoints with words.

How shall we engage those users who are unwilling to voice their thoughts on social media platforms, and encourage them to be involved in interpersonal communications? To address this challenge, we propose a solution to automatically generate poll questions for social media posts, which allows us to actively engage users who are reluctant to explicitly talk. The details will be discussed in Chapter 5, where we will show that a poll question not only offers the channel for users to voice opinions via easy voting, but also demonstrates the potential to draw later user engagements in online conversations.

1.3 Contributions

As discussed above, to help users better engage in social activities, we propose employing natural language processing methods to assist users to understand others' posts, get on track of the conversation context for those involved in discussions, and draw user engagements for those reluctant to talk.

To tackle these tasks, we first investigate NLP models' understanding ability on noisy social media text over a newly built benchmark and

found out that the performance of cutting-edge NLP models based on pre-trained BERT [24] is competitive to human beings, which suggests that NLP models are able to well understand social media posts despite of the noise after examining abundant data. Based on that, we propose two applications to help people's engagement on social media. One is for those who have already been in a conversation and signal the possible results of their replies; the other is for people who do not want to engage in a discussion explicitly and we can generate a poll question for them to hear their voice.

1.3.1 Evaluation of Social Media Language Understanding Models

To comprehensive examine models' capability in handling noisy data from social media, we present the first benchmark for social media language understanding (SMLU). While most existing benchmarks concern formal texts in common domains (e.g., news), we explore social media language exhibiting informal writings and noisy data, presenting the challenges to capture valuable features. To build the benchmark, we gather three large-scale Chinese datasets from Weibo — 80K posts with manual annotations and 3K posts with expert annotations for three fundamental tasks (Chinese word segmentation, part-of-speech tagging,

and named-entity recognition) and 46K involving user-generated emoji labels to explore a popular social media application: emoji prediction. Rumor detection, sentiment analysis, and hashtag classification are also incorporated into our SMLU benchmark with existing datasets. The experimental results demonstrate the effectiveness of trendy language encoders from the BERT family to handle fundamental tasks well and their limited capability to master post-level context understanding in handling some social media applications. Two conclusions have been drawn. First, the overall understanding ability of the NLP machine is competitive to human beings. Second, in some scenarios, social media posts should be put in richer context of user interactions to allow better understanding.

1.3.2 Conversation Context Modeling for Residual Life Estimation

To enable models to handle uncontrollable conversation environments and offer help to participants involved, we provide a solution to quantify user replies' quality with a new concept of "*residual life*", which is defined to be the possible number of new turns that will be followed by a response. For the modeling of conversation context, 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.

While previous work focus on coarse grained categorization of whether or not a conversation will end, our model is not only able to answer the "yes-or-no" question of conversation ending, but also estimate the fine-grained remaining life to be "very short", "short", "long", or "very long". Extensive experiments on various social media conversations demonstrate the superiority of our proposed model in measuring conversation environments. Moreover, we experiment with human-machine conversations and pointed out our potential to potential helpfulness in chatbot response selection, while previous studies concern human-human conversations only.

1.3.3 Poll Question Generation in External Context

To engage users reluctant to talk in online interactions, we propose a new task to automatically generate poll questions. Previous work mainly concerns question generation in formal language, such as essay questions or exam questions. Because of the potential data noisiness and context sparseness of social media posts, richer context might be needed to allow a better understanding on post level, which has also been shown in Section 1.3.1. We therefore propose to encode user comments and

discover latent topics therein as context. They are then incorporated into a sequence-to-sequence (S2S) architecture for question generation and its extension with dual decoders to additionally yield poll answers. For experiments, we collect a large-scale Chinese dataset from Sina Weibo. The results show that our model outperforms the popular S2S models without leveraging topics from comments and the dual decoder design can further benefit the prediction of both questions and answers.

Chapter 2

Background Study

In this section, we introduce our research background knowledge by reviewing related work and cut-edge papers, which mainly focus on user engagement prediction, evaluation of natural language understanding, pre-training technology, question generation, muti-task learning, etc.

We first explain the concept of user engagement and its scope in our research. Two applications – residual life prediction and question generation – are introduced respectively in the subsection, which aims to improve user engagement. Additionally, since both two applications will explore multi-task learning technique, we study this point in section 2.2. Also, we will introduce the neural topic model and pre-training techniques in section 2.3 and section 2.4 respectively. Both of them would provide some help in our generation task. Besides, we investigate the benchmark domain in section 2.5 because we need to evaluate the

understanding ability of our natural language model on social media data, which is a foundation to build an application.

2.1 User Engagement

To begin with, we would like to introduce what is online user engagement. The term online user engagement can be applied in many activities, such as human online shopping [42] and gaming applications [60]. Here it refers to helping people better socialize with others. In other words, we would like to improve people's experiences when socializing with other human beings through technology [63]. And thus everyone can have better engagement in our online social activity, such as chitchat or interaction on social media platforms.

To this end, we propose to predict the conversation residual life for participants and to generate questions for reluctant interaction users. We will introduce the background and related work for these two applications in section 2.1.1 and section 2.1.2.

2.1.1 Residual Prediction

Conversation residual life is defined as how many new messages a conversation thread will continue. We would like to predict the conversation residual life for the users and give them an early warning if the com-

ing result is not what they want. With the help of such early warning, users can avoid being conversation killers by rewriting messages before sending them out.

In previous work, there are studies analyzing the number of retweets or replies for social media messages [5, 6, 36, 71, 82], 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 [41], they only investigate the 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 [64], e.g., the prediction of user engagement degree [48, 106, 107, 108]. 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.

2.1.2 Question Generation

Another application is poll question generation for social media posts. This work aims to draw reluctant people's attention and boost them to express their opinion starting by voting or discussing this poll question.

Our work is in line with question generation, where most prior efforts focus on how to ask good exam questions given an article and the pre-defined answers. Some adopt manually-crafted rules or features [25, 28, 35, 40, 47, 55], largely relying on the labor-intensive process for rule design or feature engineering. To simply the training, automatic feature learning hence becomes increasingly popular. For example, [13] first employs a Bayesian model to learn topic features and then leverages them to yield questions. These pipeline methods require the expertise involvement to manually customize the model inference algorithms, while our neural network design allows end-to-end training of topic modeling and question generation.

Recently, built upon the success the encoder-decoder framework, S2S-based question generation architecture has demonstrated promising results [12, 27]. To better encode the input, researchers adopt suc-

cessful training design from other tasks, such as self-attention mechanism [75, 116], language model pre-training [65], variational inference [105], and reinforcement learning [65, 109]. Heuristic features, e.g., the answers' positions in the article [43, 56, 83, 117] are sometimes considered. For question decoding, certian contraints are added to control the generation, such as some aspects to be contained [39], varying levels of difficulty [29] and specificity [11].

2.2 Multi-task Learning

Multi-task learning, also named joint learning, is mainly used for improving model generalization ability. Usually, it can share parameters or model blocks by hard parameter sharing [8] or soft parameter sharing [2].

Predicting conversation residual life is our first application attempt to leverage multi-task learning. we combine fine-grained categorization with coarse-grained classification to find that two tasks can help each other reach a better optimum.

Our question generation is also related with previous work handling the generation of questions and answers in a multi-task learning setting [84, 88, 96]. Nonetheless, none of the aforementioned research concern vote questions and answers on social media, which exhibit very

different language styles compared with any existing studies and has been extensively explored.

2.3 Topic Modeling

In the question generation task, we leverage the neural topic model to extract latent topic information and fed this feature to enhance the s2s framework.

Topic models aim to discover topic words from word co-occurrence at a document level. The most famous traditional topic model is latent Dirichlet allocation (LDA), which is based on Bayesian graphical models [9]. However, these models rely on the expertise‘s participation to customize the model based on specific situations [15].

Since machine learning stepped into the era of deep learning, the neural topic model [62] emerged in response to this right moment. Recent works use the neural topic model to infer latent topics and then further to facilitate s2s framework training, which no longer requires expert effort.

The neural topic model has proven useful for downstream tasks, such as citation recommendation [7], keyphrase generation [97], and conversation understanding [110]. Different from them, our topic model is to learn from comment data and explore the question generation task in a multi-task learning setting.

2.4 Pre-training

With the advent of pre-training and fine-tuning paradigm, the performance of various natural language processing tasks has been revolutionized. In Chapter 5, we will investigate whether the pre-trained models can boost our question quality in a social media setting. Here we have a preliminary study on pre-training techniques.

Trendy pre-trained models are ELMo[67], ULMFiT[38], BERT[24], ALBERT[49], ERNIE[115], RoBERTa[57], XLNet[104], GPT-3[10], etc.

Autoencoding training. Google's BERT[24] is a pre-trained encoder from a Transformer[91] model with unlabeled text. One additional output layer can be introduced to create state-of-the-art models for a wide range of tasks. The researchers reported that eleven NLP tasks surpassed previous results of accuracy, such as pushing the GLUE score to 80.5% (7.7% point absolute improvement). Baidu's ERNIE[115] incorporates knowledge graphs to model external knowledge and thus provide rich structured knowledge facts for better language understanding, resulting in significant improvements on various knowledge-driven tasks. Facebook's RoBERTa[57] is an optimized method for BERT's training, such as masking strategy, key hyper-parameters, and data size, which leads

to better downstream task performance and thus illustrates that BERT was significantly undertrained. Google's ALBERT is a lite BERT, an upgrade to BERT that advances the state-of-the-art performance on twelve NLP tasks. ALBERT uses 89% fewer parameters than the BERT model, with two optimizations to reduce model size — (1)factorization of the embedding layer and parameter, (2)sharing across the hidden layers of the network.

Autoregressive training. Based on the bidirectional LSTM language model, ELMo[67] is pre-trained on a large text corpus to extract deep contextualized word representation which can be easily added to existing models and significantly improve state of the art across six challenging NLP problems, including question answering, textual entailment, and sentiment analysis. ULMFiT[38] is another pre-trained model in the style of the language model, which is inspired by inductive transfer learning and aims to be applied for any general task in NLP. Based on the Transformer-XL model[22], XLNet[104] has been pretreated on a generalized autoregressive method and has surpassed BERT in 20 tasks. By scaling up the transformer-based language model to 175 billion parameters, 10x more than any previous non-sparse language model, GPT-3[10] is applied without any gradient updates or fine-tuning. This

giant model significantly improves task-agnostic, few-shot performance. In addition, the researcher reported that GPT-3 sometimes even reaches competitiveness with prior state-of-the-art fine-tuning approaches.

2.5 Benchmark Evaluation

As we mentioned above, model understanding social media data is the foundation of building applications. In this subsection, we study the benchmark for evaluating the understanding ability of the NLP model.

The evaluation of the natural language understanding (NLU) models is drawing growing attention in the NLP community. However, how to compare the capabilities of the various models is still challenging. To that end, SENTEVAL [21] benchmark presents seven tasks to compare the language representation abilities of various word embeddings. Based on that, GLUE [93] provides new NLU tasks to ensure consistency and comparability between different models. Afterward, as the advances achieved by pre-trained models, the results of BERT [24] family models are found to be better than human performance on GLUE. In view of that, the SUPERGLUE [92] benchmark recently comes out with more reasonable tasks to probe the NLU abilities of modern models. Besides English, benchmarks in various languages are proposed, e.g., Polish [72], Korean [33], Indonesian [44], and Chinese [102]. Nevertheless, none of

the existing benchmarks examine the informal styles and data sparsity in social media language and how they affect models' NLU abilities.

Our work is also related with handcrafted benchmark Twitter datasets for tasks like POS tagging (1,827 tweets) [30] and NER (2,400 tweets) [68]. These small datasets are unable to examine the trendy NLU models based on deep learning (concerning overfitting). Moreover, they focus on specific tasks while we study general SMLU involving multiple tasks.

Chapter 3

A Chinese Benchmark for Social Media Language Understanding

As we mentioned in 1.2.1, social media exhibits an informal and colloquial language style, where the problem of special tokens and data noisiness is prominent. In this section, we propose a natural language understanding benchmark to investigate models' understanding ability over social media text.

3.1 Introduction

In view of the growing popularity of social media, the last decade has witnessed a large revolution of individuals' everyday communication manners. As more and more people turn to the online world to voice

[$P1$]:挺好看的 #鬼吹灯之黄皮子坟# 第04集 阮经天携探险团绝地求生 一起来看O网页链接 (来自@腾讯视频) ⟨*Nice #TheWeaselGrave# episode 4 Ethan Juan and his expedition survived from the danger. Oh, let's watch this together. URL (From @TencentVideo)*⟩

[$L1$]:挺\AD 好 看\VA 的\DEC #鬼 吹 灯 之 黄 皮 子 坟#\HASH 第04\OD 集\M 阮经天\NR$PER 携\VV 探险团\NN 绝地\NN 求生\VV 一起\AD 来\VV 看\VV O\FW 网 页\NN 链 接\NN (\PU 来 自\VV @腾 讯 视 频\MENT)\PU

[$P2$]:我了宰宰到底经历了什么⟨*What happened to our beloved Jae Jae*⟩

[$L2$]:🐱 doge ⟨*tease*⟩

Figure 3.1: Two example Weibo posts. $P1$ and $P2$ are originally in Chinese with its translation put in ⟨⟩. $L1$ indicates the annotation for Chinese word segmentation (separated by space), POS tagging (after \), and NER (after $). $L2$ shows the emoji added by the author.

opinions or exchange ideas, they jointly contribute to the formation of a new language genre — social media language, which is widely adopted by today's social media users to initiate or engage in discussions. It exhibits short and colloquial styles, which is beneficial to carry user-generated content for wide broadcast and easy communications.

In this chapter, we study social media language understanding (SMLU), which focuses on developing the automatic ways to encode social media messages and discover the essential features therein for handling downstream tasks. It has been shown in previous work that models' SMLU abilities will largely benefit various social media applications,

such as emotion analysis [1, 4] and keyphrase prediction [97], all helpful in facilitating people's decision makings and allowing quick accesses to the salient contents from massive amounts of social media data.

Nevertheless, the informal styles of social media language present concrete challenges for SMLU models to capture meaningful representations [97]. To better illustrate that, Figure 3.1 shows two posts from Weibo[1], P_1 to advertise a TV series (*The Weasel Grave*) and P_2 to flirt the idol celebrity (*Jae Jae*). As can be seen, there appear many fresh words and slangs, such as 我了 (a fandom slang as an intimate "our"[2]), which may substantially hinder NLU models' capability to make sense of them. It is because the dynamically evolving words and language patterns on social media might result in the severe data sparsity problem, where NLU models are unable to gain the essential semantics from limited features.

To evaluate the existing NLU models' capabilities to handle social media language, we present the first Chinese benchmark with six datasets from Weibo. We contribute one 80K large-scale dataset and one 3K tiny version, all of which are manually annotated for three fundamental Chinese processing tasks — Chinese word segmentation, part-of-speech

[1] `weibo.com`. A popular Chinese social media platform, exhibiting Twitter alike styles.

[2] The Chinese word "我了" is derived from the Korean word "wuli" through the similar pronunciation, as Chinese fandom culture is largely influenced by South Korea.

(POS) tagging, and named-entity recognition (NER). We also contribute 46K posts with user-tagged emoji for a popular social media application — emoji prediction. Another 3 social media application datasets are adopted form existing works to investigate SMLU in several aspects, i.e. rumor detection, sentiment anylyse and hashtag classification. Compared with the existing benchmark focusing on formal languages from common domains (e.g., news articles and Wikipedia), our benchmark concerns texts on social media exhibiting different language styles attributed to the informal writings and noisy data.

To the best of our knowledge, *we are the first to study social media language understanding (SMLU) benchmark* with four large-scale and two small datasets presented for the easy comparison and evaluation of how models understand social media language.

Extensive experiments are carried out on our benchmark. We examine the SMLU capabilities of the state-of-the-art pre-trained models from BERT family in both separate training and multi-task learning settings. The results show that trendy pre-trained models can capture a good understanding of social media language via fine-tuning on large task-specific and well-annotated datasets. Nevertheless, they are still unable to perform applications well, involving more noisy user-annotated labels. Multi-task training can boost the performance of baseline models

but their performance gain over the advanced pre-trained models are very limited. Then, we quantify model performances over varying post length in the NER task (the most challenging one among the three fundamental tasks), where the additional layers or multi-task training consistently present the performance gain. Finally, we discuss the existing model's limitation on SMLU to provide insights for future work.

3.2 Study Design

This section firstly introduced seven tasks for our SMLU benchmark. And then, we show the datasets, including preprocessing method, datasets statistics information, and preliminary data analysis.

3.2.1 Tasks

We designed seven tasks to evaluate model understanding ability, which can be mainly divided into two types: traditional tasks and social media applications. Traditional tasks are Chinese Word Segmentation (SEG), Part-of-Speech Tagging (TAG), and Named Entity Recognition (NER), which are syntactic tasks based on character-level hidden state classification in a local perspective. These tasks are denoted as STAR for abbreviations. As for social media applications, it includes Rumor Detection, Emoji Prediction, Sentiment Analysis, and Hashtag Classification,

henceforth RESH, which are classification tasks based on the final state of the encoder in a global perspective. All tasks are introduced as follows.

Chinese Word Segmentation (SEG). For this task, the goal is to delimit word boundaries in a Weibo post. The motivation behind the task is that Chinese (like many Asian languages) do not have word delimiters and they are widely-used text units for Chinese processing [51].

Part-of-Speech Tagging (TAG). Given a word sequence, the target is to predict the part-of-speech (POS) tag for each word. The Weibo POS tagset is customized following CTB standard (for Chinese POS tagging) [103] with new POS tags introduced by the previous Twitter POS dataset [30], resulting in 40 tags.

Named Entity Recognition (NER). We aim to extract words from a post to form named entities and label their NER type, e.g., person names. The NER tagset (with 10 tags) is designed following the Twitter NER dataset [68].

Rumor Detection (RUM). Rumors spread dramatically fast through online social media platforms [81]. To automatically detect rumors before they cause severe social disruption poses a high requirement on text understanding. This task is mainly to identify rumors from posts.

Emoji Prediction (EMO). Weibo users may label emojis to posts to express feelings. Here we focus on the 24 official tags [26] and train models to predict the emoji given a post. It can be considered a 24-class classification task.

Sentiment Anylyse (SEN). Social posts can be used as sensors to perceive users' feelings, which can be beneficial to collecting mood swings for the government. We aim to identify labels for posts out of 6 classes varying from "surprise" to "neutral" and to "fear", which is based on a discourse-level understanding of posts.

Hashtag Classification (HAS). A hashtag is a form of user-generated text label for their own social media message, which is usually prefaced by the hash symbol #. Hashtags enable users to search cross-user posts if related posts have been tagged with that hashtag. This hashtag task is designed to classify posts out of predefined 50 hashtags.

3.2.2 Dataset

In this subsection, we firstly introduce our datasets, then show the data processing, and finally, data analysis. Our SMLU benchmark provides 6 Weibo datasets for evaluation, among which are two datasets with manual annotation labels for STAR tasks and four datasets for RESH

tasks. The statistics of our datasets are displayed in Table 3.1.

Task	Class	Dataset		Size	Post Len
SEG	4	STAR	Large	80,000	47.77
TAG	41				
NER	21		Tiny	3,000	53.15
RUM	2	CED		3,300	110.15
EMO	24	EMO		46,022	59.28
SEN	6	EWECT		34,768	42.78
HAS	50	HASHTAG		94,732	10.14

Table 3.1: Dataset statistics. Task: evaluation task of SMLU; Class: type of classification label; Dataset: dataset name; Size: size of dataset; Post Len: average count of characters per post.

STAR-Large is a dataset at a large scale of 80K with crowdsourcing manual annotation labels for SEG, TAG, and NER, while STAR-Tiny is a lite version, a refined dataset by experts with a 3K sample. SEG and NER are defined as 4, 21 classification tasks following BIOES and BIO rules respectively. As for TAG, one additional label ”unknown” is introduced to be compatible with the noisy labels (annotation typo), which finally results in 41 classes. EMO is a dataset for emoji prediction, with post and user-generated labels.

The rest datasets named CED, EWECT, and JUNE for their corresponding task are reused from existing publications and competitions. CED is adopted form [81], which provides the true or false result for

Weibo post. Probably posts for rumor detection tend to tell breaking news. CED has longer posts on average, almost double as many as others. EWECT was released in a competition named The Evaluation of Weibo Emotion Classification Technology (SMP2020-EWECT). JUNE is a dataset with posts under 50 topics in the month of June 2014, which comes from [52]. As can be seen, JUNE posts exhibit much fewer Chinese characters on average because stop words are removed away in the original dataset.

Data Collection and Processing. Here we firstly provide the process to build the dataset STAR and EMO. Then shows the methods to split datasets to train, valid and test subsets.

To build STAR, the 83K raw data is gathered with Weibo search API [3]fed with the hashtags (queries) trended in Sep-Dec 2018 .[4] We recruited experienced annotators to conduct manual annotation. Here we separated the annotators into two groups to work independently. The inter-annotator agreement is measured on SEG, TAG, and NER based on data from two sides, which results in 0.873, 0.782, and 0.686 Cohen's Kappa [99], indicating fair and moderate agreement.

Then we randomly selected 3K posts for experts to further fine

[3]https://open.weibo.com/wiki/C/2/search/statuses/limited

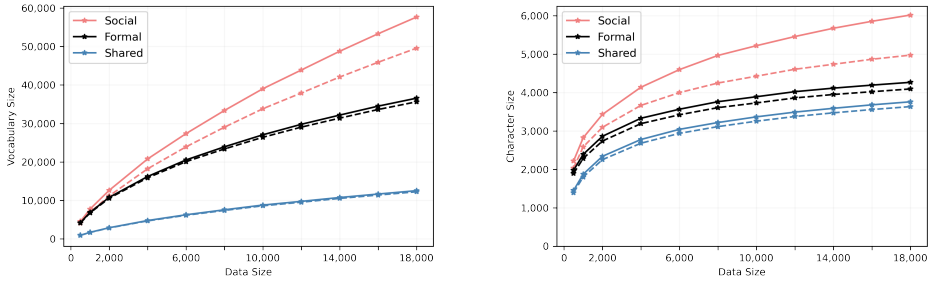
[4]https://open.weibo.com/wiki/Trends/en

trimming. For the disagreed labels, the third annotator group (experts) reviewed and picked a side. At last, we obtained 80K posts with noisy-labeled (STAR-Large) and 3K posts with fine-trimmed labeled (STAR-Tiny) for STAR, all with high-quality tags. We measure the vocabulary frequency over 83K STAR and observe that almost 83.4% of the vocabulary appears less than five times, whereas 59.2% of them appear only once. The sparse distribution of vocabulary patterns in STAR sheds light on the challenges of SMLU.

For the second dataset (EMO), its raw data collection process is similar to STAR, though the data was collected in 2020. Then we selected the posts containing the 24 official emojis and adopt majority vote to handle posts with multiple emoji labels. We further analyze the emoji label distribution on EMO. The top four emojis (❤ *(heart)*, 😩 *(facepalm)*, 😭 *(cry)*, and 😼 *(doge)*) take 53.1%. This indicates users' diverse preferences over emojis and the label imbalance challenge in EMO task.

Sparsity Analysis. To investigate the difference and tendency between social media and formal text in a quantified and visualized perspective, we firstly sample a subset in the same data size from STAR-Large and Chinese Treebank 9.0 (henceforth, CTB) respectively for statistics, vary-

ing the size of the subset from 500 to 18K. CTB is a formal text dataset which mainly collected from news and magazine. And then, we show the vocabulary distribution of shared words under 18K samples from each side.



(a) Vocabulary Size (b) Character Size

Figure 3.2: The left figure shows vocabulary size grows with data size. The dashed lines mean to filter out non-Chinese characters from solid lines in the same color respectively. The right one display the size of basic character set in corresponding data size.

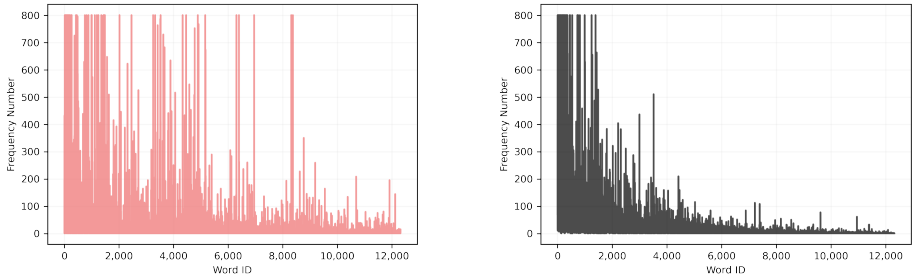
In Figure 3.2(a), it shows the size of vocabulary set in social media posts and formal text respectively, varying the sampling data size from 500 to 18K. The number of shared words between both vocabularies is also plotted to distinguish the gap. The dashed line in the same color as a solid line is in the same setting but filtering out non-Chinese symbols. In contrast, the Figure 3.2(b) shows these statistics based on single Chinese character level. All statistical values with star symbol * are average values with five repeated sampling from the corresponding full datasets.

Observations are summarized as follows.

For vocabulary, 1) Social media have richer and distinctive vocabularies. Vocabulary size increase with the data size, while the tendency of social media is more dramatic compared with formal text but shared word grows obvious slowly. The main reason comes from fresh words, typos, and word-symbol combinations that are created every day, which may not be used in the news; 2) Non-Chinese symbols accelerate the explosion of social media vocabulary. Chinese social language is usually decorated with symbol emoticon(e.g., TAT denotes cry) and initials alphabet (e.g., YYDS means eternal god in Chinese pinyin of ”永远的神”).

For character, 1) Character sizes in two domains are both relatively small. For example, the number of social language characters is one-tenth that of social vocabulary under 18K samples; 2) Social media and formal text consistently share a high proportion of characters with respect to different data sizes, although Non-Chinese symbols continue to prop up the gap between statistics of social media.

The comparison of vocabulary and characters indicate that vocabulary space can be sharply reduced by leveraging character tables, although the same character has a different meaning in a different context. Thus, Bert's family [24] shows great potential to solve this dilemma,



(a) Social Media (b) Formal Text

Figure 3.3: This figure shows the vocabulary frequency numbers over shared words. The left figure is social media vocabulary distribution while the left one is a formal text version.

with the pre-trained models' embedding for each character based on context.

Figure 3.3 shows the vocabulary frequency numbers over shared words which mean appear in both social media and formal datasets, with the X-axis being the shared word id. We set the ceiling as 800 for frequency numbers to zoom in on the distribution because the span of the range is quite large. The left figure 3.3(a) shows the social media data distribution while the right figure 3.3(b) is for formal text version.

It can be obviously observed that the tendency of the distribution of social media does not always overlap with formal text. We print the top 10 frequency word in both side and find this following observation. "企业" (enterprise) highly appear in informal text but seldom in social media while "视频" (video) reverse. What is more, the informal text is more concentrated while the social media distribution trend is decentralized.

This tendency poses a great challenge for the social media model to handle every word well.

3.3 Experiments

This section firstly shows the experimental setup, which includes model setup, training setup, and evaluation then exhibits the main SMLU benchmark results. And finally, give some auxiliary analysis which includes control-variable analysis, cross-domain analysis, error analysis, and case study.

3.3.1 Experimental Setup

Comparison Models. We first adopt the RNN encoder as the baseline [19]. Then, two state-of-the-art pre-trained encoders from BERT family — BERT and RoBERTa [57] — are considered in SMLU comparison. A two-layer biGRU is adopted as an instance of RNN and the BERT family is implemented with Transformers toolkit [100].

Training Setup. All datasets are split into 80% for training, 10% for validation, and 10% for the test, except for EWECT (split by the official competition organizer) and the STAR-Tiny. One thousand data points are sampled from the STAR-Tiny to build a test set for fair evaluation over all experiments related to STAR tasks. We train and save the model

by epoch and use the early stops strategy if the average loss of that epoch is not decreasing. In each epoch, a mini-batch iteration is used for parameter updating. To optimize parameters, we apply the Adam optimizer and set gradient clipping as 1.0. The initial learning rate for the pre-trained model is 1e-5, and that for RNN is 0.001. In STAR training, we both separately explore single-task training for a task-specific model and multi-task training. To further explore the effects on data scale, we examine two different training datasets: LARGE and TINY.

Evaluation. All results of STAR are measured in F1-score, using the standard conlleval script[5]which is commonly used for measuring the quality of sequence labeling prediction [58]. As for RASH classification tasks, we adopt accuracy as the evaluation metric.

3.3.2 Experimental Results

Our SMLU benchmark contains STAR and RESH tasks. The STAR result is shown in table 3.2 and the RESH result is displayed in table 3.3.

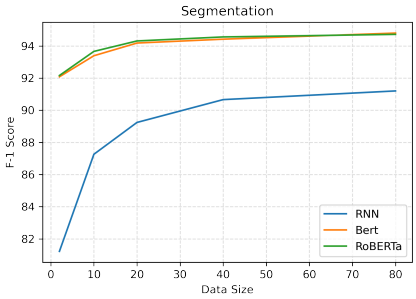
From the table 3.2 of the STAR results, we draw the following observations. First, models all rely on the scale of data and all models perform worse on TINY than LARGE. Especially for NER, almost

[5]https://github.com/sighsmile/conlleval

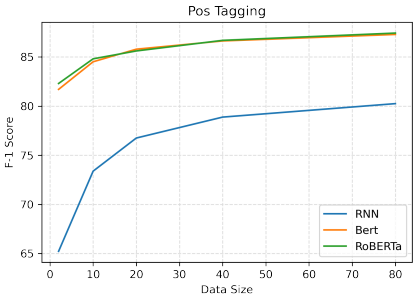
MODEL	SEG		TAG		NER	
	LARGE	TINY	LARGE	TINY	LARGE	TINY
SINGLE-TASK TRAINING						
RNN	91.20	81.22	80.25	65.22	62.35	37.09
BERT	94.80	92.08	87.29	81.70	74.63	53.41
ROBERTA	94.72	92.16	87.43	82.30	76.40	46.65
MULTI-TASK TRAINING						
RNN	91.25	82.28	80.31	65.97	62.04	38.42
BERT	94.87	92.30	87.20	81.40	76.55	52.60
ROBERTA	94.88	95.57	87.14	81.44	76.26	53.25

Table 3.2: STAR results. Models conduct sequence tagging and neural ones employ MLP-based output layer. We report performance in two training settings, i.e. single task and multi-task learning. These experiments are tested in the same testing set (1K randomly selected from expert data) and measured in F1-score.

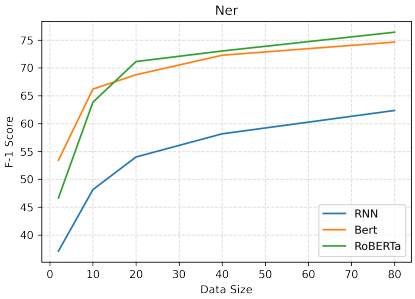
all models have 20% gap between two datasets. Second, pre-trained models result in better results. On a small data scale, the pre-trained model help to boost the performance significantly. As for large data size, the pre-trained model still provides a significant contribution. In the TAG and NER task, pre-trained models exhibit 6% to 10% relative improvement in general. Third, the pre-trained model help to reduce the performance gap between large and tiny dataset in Segmentation and Pos Tagging while providing limited help for NER. Surprisingly, a pre-trained model trained on tiny data is competitive with that on Large data in Segmentation, and this ”pre-trained” plus ”tiny data” setting even



(a) Segmentation.



(b) Pos Tagging.



(c) NER.

Figure 3.4: This figure shows the F1-score (y-axis) of Segmentation, Pos Tagging, and NER tasks varying with data size (x-axis). These experiments are tested in the same testing set and measured in F1-score. In the tasks of Segmentation and Pos Tagging, the performance of Bert is comparable to that of RoBERTa. Fine-tuning with Tiny data (3K) is sufficient for the pre-trained models to thoroughly beat RNN with a large amount of data (80K). In the third subfigure, RoBERTa turns defeat into victory to Bert as data size increases. Fine-tuning over small data in NER has a limited advantage when compared to RNN trained over big data.

can defeat the performance of RNN on Large data in both SEG and TAG. Fourth, RoBERTa is competitive with Bert but fails to follow up with Bert in NER under tiny data. We will have a detailed discussion on the last two surprising findings in the next section.

Inspired by the previous findings [93], individual and related tasks

(SEG, TAG and NER) might help each other and thus multi-task training may deliver gains over separate training. Interesting, we see from Table 3.2 (the lower half) that multiple-task training indeed boosts performance on TINY, especially for RoBERTa to do SEG and NER tasks. However, for TAG, multiple-task training provides limited help in both small and big data size. As for big data setting, single-task training overall do as well as multi-task training. The reason may be that the classification size of the TAG task is much larger than that of SEG and NER, with 41 vs 4 and 21. So it is harder for the TAG task to converge in multitask learning.

MODEL	RUM	EMO	SEN	HAS
RNN	59.09	34.30	70.26	52.11
BERT	65.45	38.89	77.22	55.88
RoBERTa	62.12	38.97	77.78	55.72

Table 3.3: RESH results. All these applications are formulated as classification tasks. Rumor Detection (RUM), Emoji Prediction (EMO), Sentiment Anylyse (SEN), and Hashtag Classification (HAS) have 2, 24, 6, and 50 classes respectively. Accuracy is adopted as evaluation metrics.

We further examine the RESH tasks and report the results in Table 3.3. Observations are summarized as follows. First, although pre-trained models obtain substantial gains, all models perform poorly, which implies that social media post is beyond the capability of existing NLU models to well understand. For example, as mentioned above, Rumor Detection

is formulated binary classification problem, which picks a side "True" that can reach 60% accuracy while the best performance of models is only 65.45%. Second, user-generated are much noisier than expert annotations. The labels for EMO and Hash tasks are user-generated and the labels for RUM and SEN are annotated by a specialist. Obviously, the accuracy of EMO and HAS is relatively lower. Third, similar to the last observation in STAR, RoBERTa is competitive with Bert but fails behind Bert in RUM tasks. Again, the data size of RUM is relatively small, which can give a preliminary sense that RoBERTa is not good as Bert at a small data size. This conclusion will be verified in the next section.

Further Discussions on EMO. As shown above, advanced models can perform well on fundamental tasks, yet the EMO task is still very challenging for them. Even for humans, it is sometimes hard to predict which emoji authors tend to use from a short and informal post. In Figure 3.1, P_2, "我了宰宰" (*our beloved Jae Jae*) is crucial to predict the emoji, while as an uncommon slang, catching its semantics may be hard. Besides, the emoji 🐱 (*doge)* has ambiguous meanings (here it means flirting) and our models wrongly predict 🤦 *(facepalm)*.

Furthermore, we probe into the model outputs and analyze the errors.

First, the prominence of fresh words largely affects the model results. For instance, the uncommon named-entity "泪花灰" (a series of cosmetic contact lenses) cannot be recognized by our models. Second, for EMO, the label imbalance discussed in (§5.2.2) presents concrete challenges. The existing models can hardly predict some uncommon emojis, such as 😲 *(wow)*.

Human Rating. Here we sample 100 source posts for each task and invite four native Chinese speakers to do the social media task, which will compete with the output of NLP models.

	Human	RoBERTa
EMO	16%	41%
HASH	20%	50%

Table 3.4: Average human rating accuracy. Higher scores indicate better results. Models exhibit the good potential to outperform humans under noisy data.

Table 3.4 shows the average accuracy of the four annotators and the model RoBERTa. We can observe that machines can outperform human beings in social media applications. This may be because machine models read more training data and master more background and knowledge.

Here gold standard labels is user-generated. The ground truth label for the corresponding post is user-generated, while in human rating

evaluation, the post is not generated from the participant. The prediction for the social media task is very challenging for human because there is a gap between the blogger and reader. For example, the blogger may use 'dog' emotion to express unhappy, while participant prefer to use 'cry' emotion to rate the unhappy context.

3.4 Conclusion

We present the first benchmark for Chinese social media language understanding (SMLU), with two large-scale datasets. The empirical results show that trendy pre-trained NLU models can perform well via fine-tuning on large-scale, task-specific, and well-annotated data, though the understanding of noisy emoji labels tagged by social media users is still beyond their capability. These findings suggest that SMLU is still challenging and we believe the availability of our benchmark will advance forward future work on social media analysis.

3.5 Appendix

宋茜	吴亦凡	快乐大本营	重返20岁	朴有天
exo快乐大本营	伯贤	郑爽	世界杯	韩庚
爸爸回来了	父亲节	金曲奖	郑秀晶	穆勒
鹿晗	天天向上	戚薇	金泰妍	tfboys
iphone6	不一样的美男子	周笔畅	尼坤	爸爸去哪儿
c罗	刘烨	郑容和	西班牙	赵丽颖
泰妍	梅西	陆毅	花儿与少年	摩纳哥王妃
柯震东	王俊凯	时间煮雨	何以笙箫默	我是女王
佑荣	言承旭	男神顾长官	佟丽娅	金秀贤
蔡依林	鬼鬼吴映洁	小苹果	宁财神	钟汉良

Table 3.5: Hashtag labels.

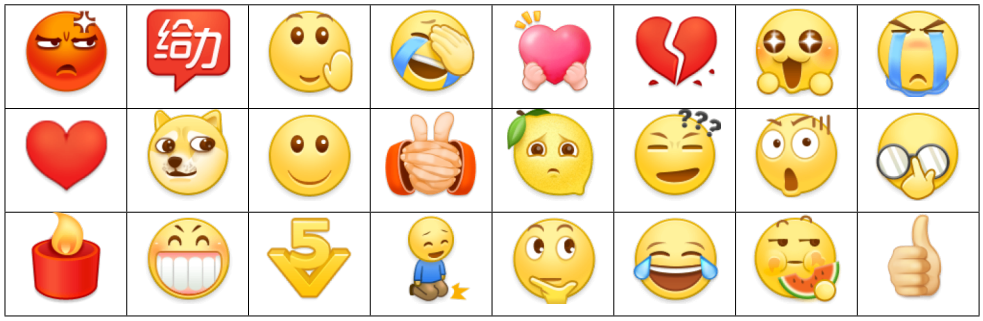


Table 3.6: Emotion labels.

b-time	i-time	b-loc	i-loc	b-per	i-per
b-math	i-math	b-org	i-org	b-book	i-book
b-brnd	i-brnd	b-ent	i-ent	b-money	i-money
b-song	i-song	no-ner			

Table 3.7: NER labels.

nr	lb	nt	url	ij	ment	nn	vc	email	p
emoj	hash	lc	cd	fw	msp	sb	as	ve	on
pn	ad	va	sp	ba	dev	od	m	vv	deg
dec	cs	emot	der	dt	jj	pu	phone	cc	unk

Table 3.8: TAG labels.

Chapter 4

Getting Your Conversation on Track: Estimation of Residual Life for Conversations

Based on Chapter 3, which indicates NLP models possess the ability to understand social media language, we propose two applications to improve user engagement. This chapter is the first one to improve the experience when users engage in the conversation, as we mentioned in Challenge 1.2.2.

4.1 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 [77, 79, 90] and advancing the user interaction experience [6, 41, 16]. 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* [6]. Specifically, following previous work [89] 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

[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
...

Figure 4.1: A Twitter conversation snippet. [T_i]: The i-th turn in the conversation snippet. There are nine new turns to occur.

will help one in doing the right things at the right time. For instance, when curating conversations, it is inappropriate to recommend ending discussions for users to be involved. Another promising application is on response selection [87, 112], 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" [101, 78, 46, 34], and thus turn human participants away.

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 [36, 6] and ended chats [41]; 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 [82], message diffusion patterns [45, 50, 95], 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.

To illustrate how the history contexts can affect residual life of conversations, Figure 4.1 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 [50, 95], simple lexical statistics [82, 70], and social networks of users [36, 45, 6], 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 hand-crafted 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 [6]. 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.

4.2 Preliminaries

4.2.1 Basic Notions for Conversations

We follow the definitions for common concepts of conversations from previous studies [73, 80]. The unit of a conversation is a **turn**, defined as an utterance given by one participant. Specifically, for most social media conversations [74], 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 [52], 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.,

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

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

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.

To track the progress of a conversation, we follow previous study [89] 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.

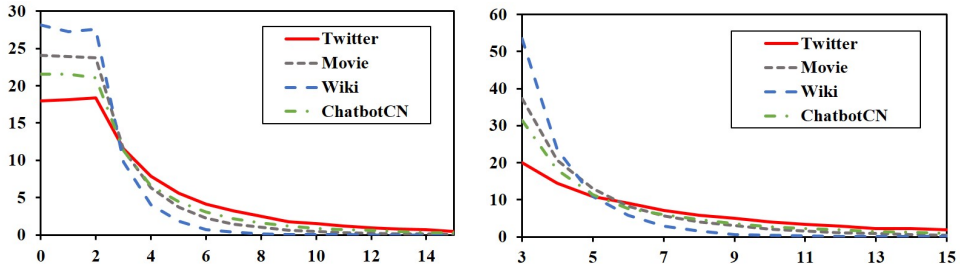
4.2.2 A Study on Conversation Data

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 4.1: 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 using Twitter search API[4]recursively with "in-reply-to" relations (the HTML pages only display partial conversations). The Movie dataset is released by [23], which contains fictional conversations held between two characters from movie scripts. It is close to off-line conversations held in our daily life [41]. The Wiki dataset is released by [6] consist-

[2]https://en.wikipedia.org/wiki/Help:Talk_pages
[3]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[4]https://developer.twitter.com/en/docs/tweets/search/overview/standard



(a) Residual Life (b) All Turns

Figure 4.2: 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 (%).

ing 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 [41].

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]

[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 [69], one with smaller mean for inter-conversation spans, and the other intra-conversation spans.

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 4.2(a). We can observe a severe imbalance for varying numbers of future turns. To understand the cause of such imbalance, in Figure 4.2(b), 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 [41, 14]. Based on the data distributions, we further determine our four residual life categories in the similar manner of [6] (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.

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

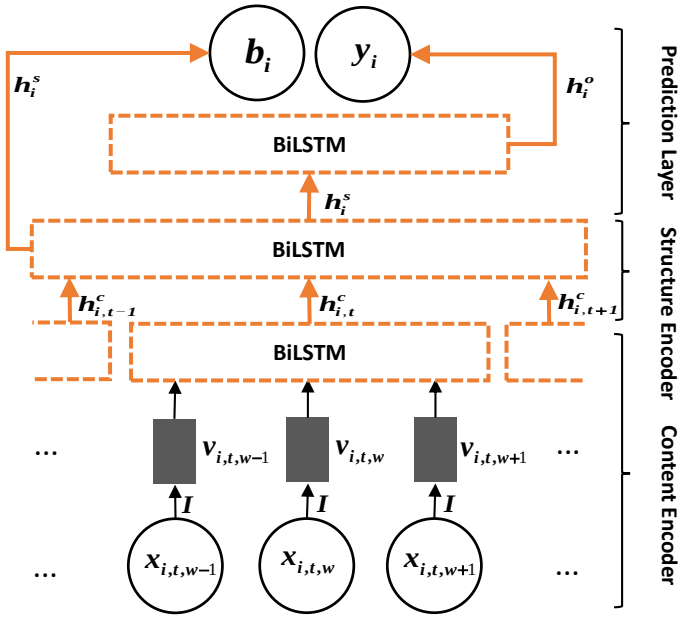


Figure 4.3: The hierarchical BiLSTM model for estimating residual life categories of conversations. Here h refers to final state of their corresponding encoder cell.

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 4.2(a)). For boundary cases, we assign them to one side of categories if more instances are found in the corresponding quantile.[8]

4.3 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

[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.

(BiLSTM) network [53] and jointly explores the content of turns and the structure of conversations. Our overall architecture is illustrated in Figure 4.3.

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$\}$.

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

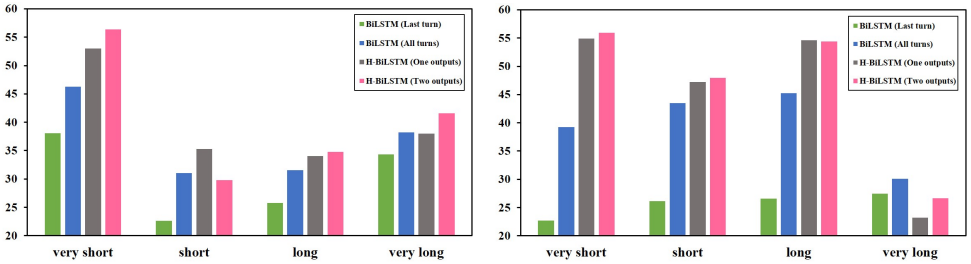
applies another BiLSTM, namely structure encoder. Its t-th state takes the content representation of the t-th turn $\mathbf{x}_{i,t}$ as input and the learned structure representation is denoted as $\mathbf{h}_i^s$.

Joint Prediction. Inspired by [114] (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 $b_i \in \{\text{ended}, \text{ongoing}\}$ 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 4.1, because the prediction of $b_i = \text{ongoing}$ 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,

$$b_i = softmax(\mathbf{h}_i^s) \tag{4.1}$$

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

$$y_i = softmax(\mathbf{h}_i^o) \tag{4.2}$$



(a) Twitter (b) ChatbotCN

Figure 4.4: 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*).

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

$$\mathcal{L}(\Theta) = \alpha \sum_{i=1}^{N} d(b_i, \hat{b}_i) + (1-\alpha) \sum_{i=1}^{N} d(y_i, \hat{y}_i) \tag{4.3}$$

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.

4.4 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 [59] for tokenization and lemma-

tization.[9]For Chinese (ChatbotCN), we applied NLPIR tool [113] for Chinese word segmentation.[10]

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) [37] and support vector machine (SVM) [17] with features proposed in [6] and [41]. 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 [31] as the optimizer for parameter updating, and the trade-off parameter α to 0.5 for balancing b_i and y_i. Pre-trained

[9]https://github.com/stanfordnlp/CoreNLP
[10]https://github.com/NLPIR-team/NLPIR

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 [20].[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 thesis and we leave the adaption work to future studies.

	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 4.2: 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 (%).

Residual Life Estimation Results. We show the main comparison results for residual life categorization in Table 4.2, where we report accuracy and average F1 scores for the four possible outcomes. The

[11]https://spinningbytes.com/resources/word-embeddings/

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.

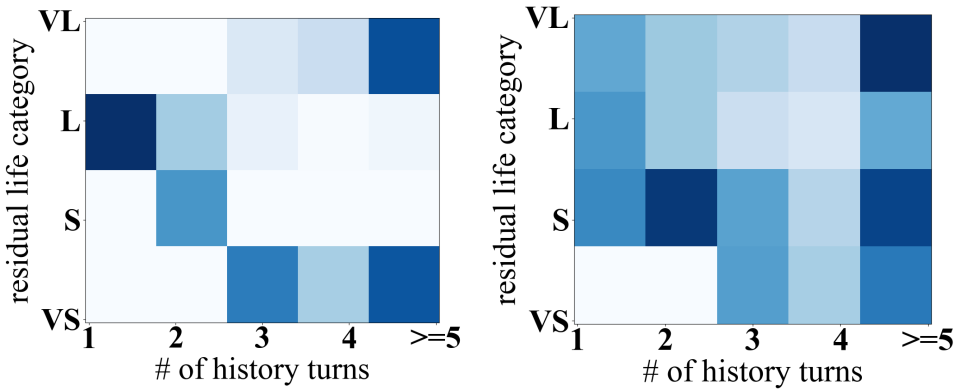
• ***Multi-task learning helps each other.*** The results of H-BiLSTM (*Two outputs*) are better than H-BiLSTM (*One output*) on all datasets. This indicates the effectiveness to simultaneously tackle the two tasks with shared parameters, because they are highly related to each other.

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 4.2. Their F1 scores in predicting residual life ranging from very short to very long are shown in Figure 4.4. 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 ($b_i = \textit{ongoing}$) 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.

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 4.5. 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



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

Figure 4.5: 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.

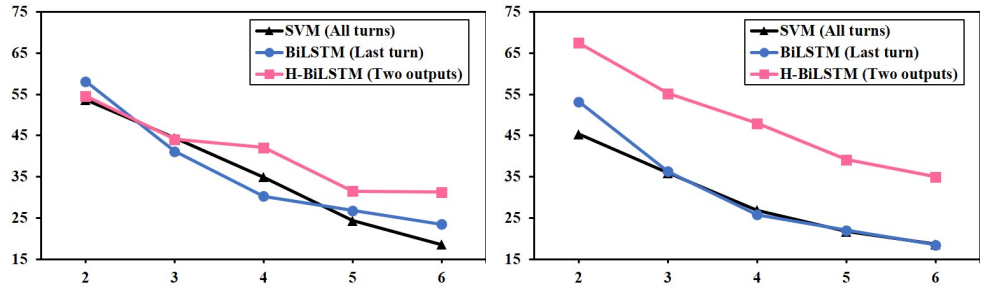
that for a dark grid in the gold-standard heatmap, its upper or lower neighbor of the corresponding grid in the predicted heatmap tends to be highlighted. This shows the particular challenge to distinguish adjacent categories, such as short and very short residual life.

From the gold-standard heatmap, we also find something interesting. For the conversation history $\geqslant 5$ 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 4.1. 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 examining 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.



(a) Twitter (b) ChatbotCN

Figure 4.6: 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.

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 4.6, 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 [94] 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 4.7. As can be seen, our model tends 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.

4.5 Conclusion

We have presented a framework for estimating four categories of conversations' residual life, corresponding to varying stages in conversation

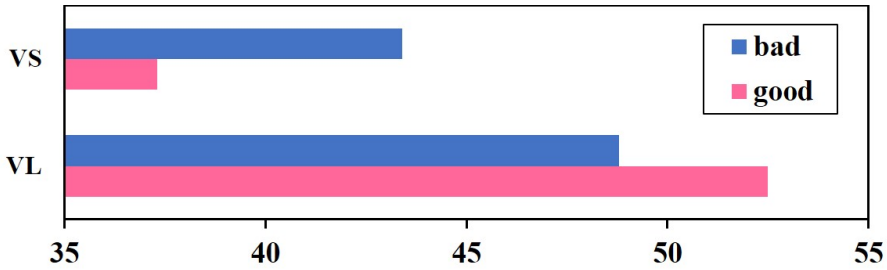


Figure 4.7: The proportions of bad and good responses predicted to have very long (VL) and very short (VS) residual life (%). For each category, the upper and lower bar shows the results for bad and good responses, respectively.

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.

Chapter 5

Engage the Public: Poll Question Generation for Social Media Posts

In the last chapter, we discuss the first attempt to help the users keep track of conversations and predict corresponding residual life to avoid some out-of-control situations, such as being conversation killers. As for this chapter, we propose to generate questions for social media posts that aim to help those people that are reluctant to voice their opinion (Challenge 1.2.3) and then draw their attention to engage in discussion.

5.1 Introduction

Social media is a crucial outlet for people to exchange ideas, share viewpoints, and keep connected with the world. It allows us to hear the

public voice for decision making and better understanding our society. Nevertheless, for the silent majority, they tend to read others' messages instead of voicing their own opinions with words, possibly because of the introvert personality, busy schedule, and others. How shall we better engage them into the discussions and learn from their thoughts?

In this work, we present a novel application to automatically generate a poll question for a social media post. It will encourage public users, especially those reluctant to comment with words, to input their reflections via voting. For example, the statistics of our dataset show that 13K users on average engaged in a poll compared with 173 commented to a post. For a better illustration of the task, Figure 5.1 shows two example poll questions on Sina Weibo[1], henceforth Weibo, a popular Chinese microblog. The goal of our task is to output an opinion question, such as Q_1 and Q_2, and invite other users to engage in the discussion to a source post (e.g., P_1 and P_2); poll choices (answers like A_1 and A_2) can be produced together to allow easy public engagement (via voting).

To date, most progress made in question generation is built upon the success of encoder-decoder frameworks [27]. Despite of the extensive efforts made in this line [83, 105, 12, 84], most previous work focus on the processing of formally-written texts, such as exam questions

[1] `weibo.com`

$[P_1]$: ...B站市值超过爱奇艺 (*The market value of B site exceeds iQiyi*)...

$[Q_1]$: 你们平时常用那个app看视频？ (*Which app do you usually use to watch videos?*)

$[A_1]$: 腾讯视频 (*Tencent Video*); 优酷 (*Youku*); 爱奇艺 (*iQiyi*); B站 (*B site*)

$[P_2]$: ...理性分析一下赵粤和希林娜依高：希林vocal确实厉害，但是...舞蹈实力有点不够看；赵粤呢舞蹈厉害...但是唱歌实力较弱些... (*A rational analysis of Akira and Curley G: Curley's vocal is indeed great, but ... her dancing is not that good; Akira dances well ... but her singing is weaker...*)

$[Q_2]$: 谁更适合当c位？ (*Who should take the center position?*)

$[A_2]$: 赵粤 (*Akira*); 希林娜依高 (*Curley G*)

Figure 5.1: Example polls from Sina Weibo. P_i, Q_i, and A_i ($i = 1, 2$) refer to the i-th source post, its poll question, and the corresponding poll choices (answers). Different choices are separated by the ";". Italic words in "()" are the English translation of the original Chinese texts on their left. In the source posts, we fold the words irrelevant to polls in "..." for easy reading.

in reading comprehension tests. The existing methods are therefore suboptimal to handle social media languages with short nature and informal styles, which might present challenges to make sense of the source posts and decide what to ask. For example, from the limited words in P_1, it is hard to capture the meanings of "B站" (*B site*) and "爱奇艺" (*iQiyi*) as video apps, which is nevertheless crucial to predict Q_1. Moreover, the question itself, being in social media fashion, is likely to contain fresh words, such as "c位" (*center position*) in Q_2, which

may further hinder the models' capability to predict the poll questions in social media style.

To tackle these challenges, we first enrich the short contexts of source posts with other users' comments; a neural topic model is employed to discover topic words therein and help identify the key points made in source posts. It is based on the assumption that the salient words in a source post are likely to be echoed in its comments [98], potentially useful to learn the map from posts to poll questions. For example, the core words in Q_1 — "app" and "视频" (*video*) — co-occur frequently in the comments with "B站" (*B site*) and "爱奇艺" (*iQiyi*), which may help the model to link their meanings together. The topic representations are then incorporated into a sequence-to-sequence (S2S) architecture to decode poll questions word by word. Furthermore, we extend the basic S2S to a version with dual decoders to generate questions and answers in a multi-task learning setting and further exploit their correlations. For example, modeling answers in A_2 might help indicate that P_2 centers around "赵粤" (*Akira*) and "希林娜依高" (*Curley G*), two celebrities.

To the best of our knowledge, *this work is the first to study poll questions on social media, where their interactions among answer choices, source posts, and reader users' comments are comprehensively explored*. As a pilot study over social media polls, we also contribute the very first

dataset containing around 20K Weibo polls associated with their source posts and user comments.[2]We believe our dataset, being the first of its kind, will largely benefit the research on social media polls and how they help promote the public engagements.

On our dataset, we first compare the model performance on poll question generation in terms of automatic evaluation and human evaluation. The automatic evaluation results show that the latent topics learned from the first few pieces of user comments is already helpful — they result in our models' significantly better performance than the S2S baselines and their trendy extensions proposed for other tasks. For example, our full model achieves 38.24 ROUGE-1 while S2S with RoBERTa [57] yields 34.08. Human evaluation further demonstrates our models' capability to generate poll questions relevant to the source post, fluent in language, and particularly engaging to draw user attentions for discussions. We then quantify models' sensitivities to the length of varying source posts and poll questions, where the scores of our model are consistently better. Next, we find our model exhibits an increasing trend in predicting poll questions that will engage more comments in the future, which suggests the potential helpfulness of comments to indicate engaging questions. At last, the performance of dual decoder designs are discussed and it is

[2]Our dataset and code are publicly available in https://github.com/polyusmart/Poll-Question-Generation

shown that joint prediction of questions and their answers can benefit both tasks.

5.2 Study Design

5.2.1 Task Formulation

Our major input is a social media post (i.e., **source post**) and the main output a **poll question** that continue the senses of the source post and encourage public users to voice opinions. For each question, possible answer choices (i.e., **answers**) may also be yielded as a side product to enable participants to easily input their thoughts. To enrich the contexts of source posts, their reply messages (i.e., **user comments**) are also encoded as external features.

5.2.2 Data Description

Here we describe the dataset we collect to empirically study social media polls.

Data Collection. Weibo allows users to create polls, asking questions to the public and inviting others to share their thoughts via voting. It enables the construction of a dataset with user-generated polls. At the beginning, we gathered around 100K random Weibo posts, whereas less than 0.1% of them contain polls. The sparse distribution of polls presents

the challenge to scale up the dataset. To deal with that, we looked in to the sampled polls and draw two interesting points: first, many polls carry trendy hashtags (user-annotated topic labels like #COVID19) to draw user attentions; second, a user who once created a poll is likely to do it again.

Inspired by these observations, we first obtained the popular hashtags since Nov 2019.[3]Then, we gathered the posts under the hashtag through the Weibo search API, from which the ones containing polls are picked out.[4]Next, we examined the authors of these polls and access their posting history to gather more polls they created from Weibo user timeline API.[5] Afterwards, for each post, we crawled its comments via the comment API.[6]Finally, 20,252 polls were obtained from 1,860 users.

Data Analysis. The statistics of the dataset is displayed in Table 5.1. As can be seen, comments are shorter than posts, probably because users tend to put more efforts in crafting original posts than replying to others and hence comments may be relatively nosier than original posts; both questions and answers are short, which follow the fashion of user-generated contents on social media.

[3]https://open.weibo.com/wiki/Trends/en

[4]https://open.weibo.com/wiki/C/2/search/statuses/limited

[5]https://open.weibo.com/wiki/C/2/statuses/user_timeline_batch

[6]https://open.weibo.com/wiki/2/comments/show

Post		Comment		Qs	Ans Choice		Voter
Num	Len	$\overline{\text{Num}}$	Len	Len	$\overline{\text{Num}}$	Len	$\overline{\text{Num}}$
20,252	54.0	173	16.9	11.0	3.4	5.9	13,004

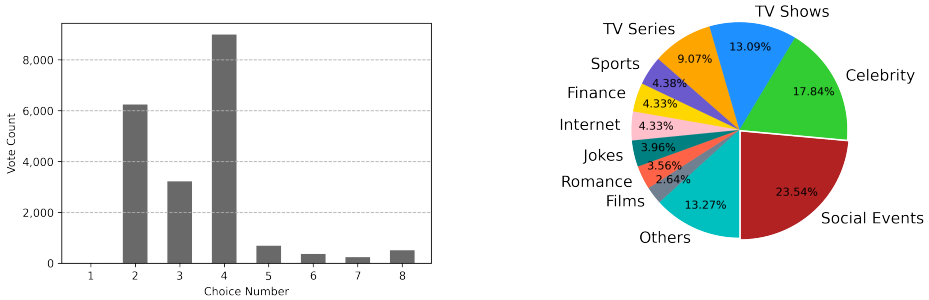
Table 5.1: Statistics of our dataset. Num: number; $\overline{\text{Num}}$: average number per post. Len: average count of words per post; Qs: question; Ans: answer.

To further investigate the data sparsity in social media contents, we sample some texts from LDC news corpus (formally-written texts) [3] — the samples contain the same token number as our social media texts. Our corpus's vocabulary size and entropy are 24,884 and 7.46, while those for news corpus are 9,891 and 5.98. This suggests the sparsity of social media data.

We also observe that each post exhibits more voters than comments, implying that users may prefer to voice opinions via voting, which is easier than commenting with words. We further analyze the effects of polls on user engagements and draw an interesting finding. For the same author, their posts with polls exhibit 1.65, 22.2, and 1.80 times comments, likes, and reposts on average compared to posts without polls.[7] This implies that adding polls indeed help to draw user engagements to a post.

For each poll, there are less than 4 answer choices on average. To further characterize that, Figure 5.2(a) shows the count of polls over

[7]For each author, we additionally sample 500 posts without polls for comparison.



(a) Choice Number Statistics (b) Topic Categories

Figure 5.2: The left figure shows the count of polls over varying choice number in their answers (x-axis: choice number; y-axis: vote count). The right one displays the distribution of the polls' topic categories.

varying numbers of answer choices appearing in them and the statistics suggest that most users are not willing to craft over 5 poll choices, which, interestingly, exhibit similar statistics in exam questions. In addition, we probe into what types of topics are more likely to contain polls. To that end, we examined source posts with hashtags and manually categorized the hashtags into 11 topics. Figure 5.2(b) shows the poll distribution over topics. Most polls fall in "social events" category, which mostly concern public emergency and in our dataset tremendous posts focus on the outbreak of COVID-19. There are also a large proportion of polls concern entertainment topics such as celebrities and TV shows, probably initiated for advertising purpose.

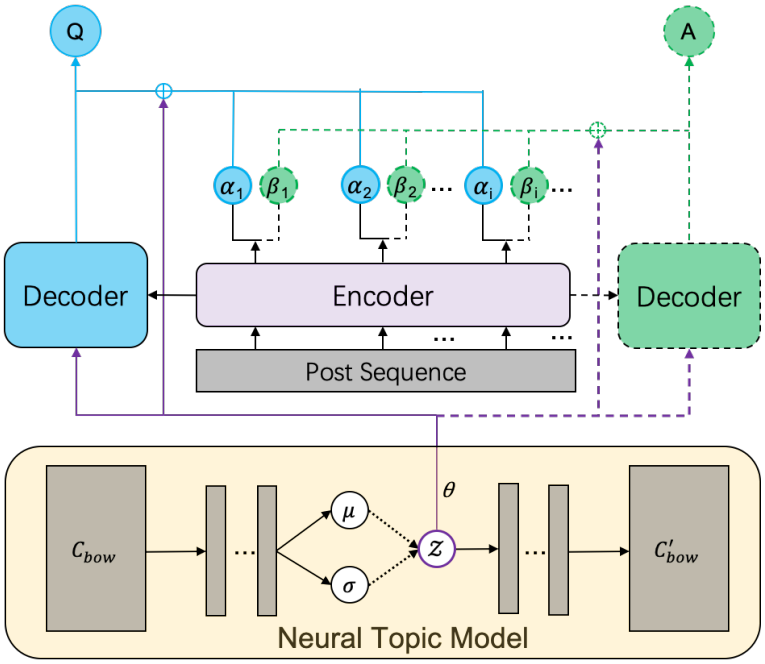


Figure 5.3: The architecture of the dual decoder S2S (sequence-to-sequence) model to jointly generate questions and answers. It contains a neural topic model for context modeling (in the bottom), a sequence encoder fed with the source post (in the center), and two sequence decoders to handle the output, where the left one predicts questions (Q) and the right answers (A).

5.3 Poll Question Generation Framework

This section introduces our framework with two variants: one based on a basic S2S (single decoder) and the other is its extension with dual decoders to predict poll questions and answer choices in a multitask learning setting. The model architecture of the dual decoder model is shown in Figure 5.3.

5.3.1 Source Posts and Comments Encoding

Following the common practice in S2S [27], we encode a source post P in the form of word sequence $\langle w_1, w_2, ..., w_{|P|} \rangle$, where $|P|$ is the number of words in the post. For user comments C, bag of words (BOW) representations are employed for topic modeling, henceforth C_{bow} over BoW vocabulary. More details are provided below.

Source Post Encoding. To encode the post sequence P, a bidirectional gated recurrent unit (Bi-GRU) [18] is adopted. For the i-th word $w_i \in P$, we first convert it into an embedding vector ν_i, which is later processed into hidden states in the forward ($\overrightarrow{\mathbf{h}_i}$) and backward ($\overleftarrow{\mathbf{h}_i}$) directions, respectively. They are then concatenated as $\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}]$ and sequentially put into a memory bank $\mathbf{M} = \langle \mathbf{h}_1, \mathbf{h}_1, ..., \mathbf{h}_{|P|} \rangle$, which will be further delivered to decoders for their attentive retrieval.

User Comments Modeling. Considering the noisy nature of user comments, latent topics are employed to recognize the salient contents therein. They are explored based on word statistics and represented as clusters of words tending to co-occur in the comments of some posts (probably concerning similar topics), such as the names of video apps in Figure 5.1. In topic modeling, we assume there are K topics and each topic k is represented with a topic-word distribution over the BoW

vocabulary. A post P has a topic mixture θ, which is learned from the words appearing in its comments C_{bow}.

Our topic learning methods (from comments) are inspired by the neural topic model (NTM) based on variational auto-encoder (VAE) [62, 111], which allows the end-to-end training of NTM with other modules in an unified neural architecture. It employs an encoder and a decoder to resemble the data reconstruction process of the comment words in BoW.

Concretely, the input C_{bow} is first encoded into prior parameters μ and σ using neural perceptrons. Then, through Gaussian transformation, they are applied to draw a latent variable: $\mathbf{z} = \mathcal{N}(\mu, \sigma^2)$, which is further taken to produce the topic composition of comments (θ) with softmax transformation. At last, the decoder reconstructs comments and produces a BOW vector C'_{bow} (conditioned on the latent topic θ) through another neural perception.

5.3.2 Poll Decoding

Here we further describe how we generate questions (and answers in the dual decoders settings) with the encoded source posts and comments.

Question Generation. To handle the output of a question Q, the corresponding decoder (i.e., **question decoder**) is formed with a uni-directional GRU and fed with the memory bank M from source post

encoding and the topic distribution θ from user comment modeling. The words in Q are predicted sequentially with the following formula:

$$Pr(Q \mid P, C_{bow}) = \prod_{j=1}^{|\mathbf{q}|} Pr\left(q_j \mid \mathbf{q}_{<j}, \mathbf{M}, \theta\right) \tag{5.1}$$

where q_j means the j-th word in Q and $\mathbf{q}_{<j}$ refers to Q's predicted word sequence from slot 1 to $j-1$. To leverage comment modeling results in the decoding, we incorporate θ into the attention weights (defined below) over source posts and concentrate on topic words therein for question generation.

$$\alpha_{ij} = \frac{\exp\left(f_\alpha\left(\mathbf{h}_i, \mathbf{s}_j, \theta\right)\right)}{\sum_{i'=1}^{|P|} \exp\left(f_\alpha\left(\mathbf{h}_{i'}, \mathbf{s}_j, \theta\right)\right)} \tag{5.2}$$

$\mathbf{s}_j$ is the GRU decoder's j-th hidden states and:

$$f_\alpha\left(\mathbf{h}_i, \mathbf{s}_j, \theta\right) = \mathbf{v}_\alpha^T \tanh\left(\mathbf{W}_\alpha\left[\mathbf{h}_i; \mathbf{s}_j; \theta\right] + \mathbf{b}_\alpha\right) \tag{5.3}$$

In addition, we adopt copy mechanism [76] to allow the generated questions to contain the keywords from the source posts:

$$p_j = \lambda_j \cdot p_{gen} + (1 - \lambda_j) \cdot p_{copy} \tag{5.4}$$

p_{gen} refers to the likelihood to generate a word while p_{copy} is the extractive distribution derived from the attention weights over the source input. The soft switcher $\lambda_j \in [0, 1]$ can determine whether to copy a word or generate a new one in aware of the comments' topics:

$$\lambda_j = sigmoid\left(\mathbf{W}_\lambda\left[\mathbf{u}_j; \mathbf{s}_j; \mathbf{t}_j; \theta\right] + \mathbf{b}_\lambda\right) \tag{5.5}$$

$\mathbf{t}_j$ is the context vector (weighted sum) of the attention to predict the Q's j-th word, whose embedding is $\mathbf{u}_j$. $\mathbf{W}_\lambda$ and $\mathbf{b}_\lambda$ are both learnable parameters.

Answer Generation. To further explore the relations between questions (Q) and answers (A), we "replicate" the question decoder's architecture and form another decoder to handle answer generation (**answer decoder**). The answer choices are concatenated to form an answer sequence and neighboring choices are separated with a special token *"<sep>"*. The answer decoder also adopts the same topic-aware attentions (Eq. 5.2) as the question decoder (denoted as β_{ij} here) and copy mechanisms (Eq. 5.4) to be able to put topic words from the source into the answer choices, such as "赵粤" (*Akira*) and "希林娜依高" (*Curley G*) in Figure 5.1.

Question decoder and answer decoder work together in a dual decoders setting, whose parameters are updated simultaneously to exploit the essential correlations of poll questions and their answers.

5.3.3 Model Training

This subsection describes how we jointly train the neural topic model (henceforth NTM) for comment modeling and the decoders for question and answer generation with multi-task learning. The loss function for

NTM is defined as:

$$\mathcal{L}_{NTM} = D_{KL}(p(\mathbf{z}) \,||\, q(\mathbf{z} \,|\, C)) - E_{q(\mathbf{z}|C)}[p(C|\mathbf{z})] \tag{5.6}$$

The C above refers to C_{bow}. The first term is the KL divergence loss and the second is the reconstruction loss in VAE. For question generation, the loss is:

$$\mathcal{L}_{QG} = -\sum_{n=1}^{N} \log\left(Pr\left(Q_n \mid P_n, \theta_n\right)\right) \tag{5.7}$$

N is the number of training samples; Q_n, P_n, and θ_n are the target poll question, source post, and topic distribution of the n-th training sample.

$$\mathcal{L}_{AG} = -\sum_{n=1}^{N} \log\left(Pr\left(A_n \mid P_n, \theta_n\right)\right) \tag{5.8}$$

Answer generation loss $\mathcal{L}_{AG}$ is defined similarly. The training loss of the entire model are defined as:

$$\mathcal{L} = \mathcal{L}_{NTM} + \gamma_Q \cdot \mathcal{L}_{QG} + \gamma_A \cdot \mathcal{L}_{AG} \tag{5.9}$$

where γ_Q and γ_A balance the weights over NTM and the two decoders.

5.4 Experimental Setup

Data Preprocessing. First, we removed meta data (e.g., author's locations and emoji labels) and replaced links, mentions (@username),

and digits with generic tags "URL", "MENT", and "DIGIT". Then, for some poll questions echoed in the source posts, we took them away for fair experiments. Next, an open-source toolkit `jieba` is employed for Chinese word segmentation.[8]Afterwards, we filtered out stop words and for the remaining, we maintained two vocabularies with the most frequent 50K words for sequences (input and output) and another 100K words for BoW. Finally, comments are capped at the first 100 words to examine poll question generation with the early comments and their potential to draw future user engagements.

In evaluations, we split our data into 80% for training, 10% for validation and 10% for test.

Baselines and Comparisons. For baselines, we first consider the basic S2S [86] (i.e., BASE); also compared are the S2S with pre-trained models from the BERT family — tiny ERINE [85] (i.e., ERINE), BERT [24] (i.e., BERT), and RoBERTa [57] (i.e., ROBERTA), which were implemented with the paddle hub platform.[9] For all S2S with pre-trained models, their pre-trained parameters were further fine-tuned on our training data.

Then, we consider the following S2S extensions with copy mecha-

[8]`https://github.com/fxsjy/jieba`
[9]`https://www.paddlepaddle.org.cn/hub`

nism (i.e., COPY) [61], topic modeling from posts (i.e., TOPIC) [97], and bidirectional attentions over posts and comments (i.e., CMT (BIATT)) [98]. All of them were proposed for keyphrase generation tasks and set up following their original papers.

For our models, we consider two variants — CMT (NTM) in the single decoder archetecture and its dual decoder version DUAL DEC.[10]

Model Settings. All the hyperparameters are tuned on the validation set via grid search. For NTM, it is pre-trained for 50 epochs before joint training and afterwards different modules take turns to update parameters. We adopt two-layers bidirectional GRU to build source post encoder and one-layer unidirectional GRU question and answer decoders. The hidden size of each GRU is 300. For a word embedding, the size is set to 150 and randomly initialized. In training, we apply Adam optimizer with initial learning rate as 1e-3, gradient clipping as 1.0, and early-stopping strategy adopted. The weights to trade off losses in multi-task learning is set to $\gamma_Q = \gamma_A = 1$ (Eq. 5.9).

[10]We also finetuned BERT with our models yet cannot observe much performance gain. It is because NTM is able to learn essential features from the input and BERT cannot provide additional benefits. Another possible reason is that social media BERT is unavailable in Chinese and that trained on out-domain data (e.g., news) might not fit well with Weibo languages. Large-scale Weibo data might be acquired for continue pre-training [32], which is beyond the scope of this thesis and will be explored in future work.

Evaluation Metrics. We adopt both automatic measures and human ratings for evaluations. For the former, we examine two popular metrics for language generation tasks — ROUGE [54] and BLEU [66]. For the latter, human annotators rates with 4 point Likert scale (i.e., $\{0, 1, 2, 3\}$) and over three criteria are considered: the relevance to the source posts (**relevance**), how fluent the generated language reads (**fluency**), the attractiveness degree of the questions in drawing people's engagements (**engagingness**).

5.5 Experimental Results

In this section, we first show the main comparison results on poll question generation involving both automatic evaluations and human ratings (in §5.5.1). Then, model sensitivity to varying lengths of source posts and poll questions are discussed in §5.5.2, followed by the analyses of models' capability to handle poll questions exhibiting varying degrees of user engagements (§5.5.3). Next, §5.5.4 discusses the performance of dual decoders that jointly generate questions and answers. A case study is presented at last (in §5.5.5) to interpret the sample outputs.

5.5.1 Comparison on Poll Question Generation

We first show the comparison results on poll question generation, where we will discuss automatic evaluations and human ratings in turn below.

MODEL	ROUGE-1	ROUGE-L	BLEU-1	BLEU-3
S2S Baselines				
BASE	21.62±0.7	20.64±0.7	20.35±0.7	2.11±0.5
+ERNIE	29.62±0.5	27.82±0.4	21.66±0.5	3.25±0.4
+BERT	33.62±1.2	31.57±1.1	24.43±0.7	4.54±0.4
+ROBERTA	34.08±1.3	31.98±1.2	24.88±1.0	4.85±0.5
S2S Extensions				
+COPY	35.13±0.4	33.20±0.4	30.27±0.4	7.95±0.3
+TOPIC	36.65±0.6	34.70±0.6	31.11±0.5	8.66±0.5
+CMT (BIATT)	27.74±0.4	26.21±0.4	23.97±0.3	4.15±0.2
Our Models				
+CMT (NTM)	37.95±0.4	35.97±0.3	32.07±0.2	8.89±0.3
+DUAL DEC	38.24±0.3	36.14±0.3	32.27±0.4	9.04±0.3

Table 5.2: Main comparison results for poll question generation. The underlined scores are the best in each column. Average scores are before ± and the numbers after are the standard deviation over 5 runs initialized with different seeds. Our models CMT (NTM) and DUAL DEC significantly outperforms all the other comparison models (paired t-test; p-value < 0.05).

Automatic Evaluations. Table 5.2 reports the automatic measured results on question generation.

As can be seen, our task is challenging and basic S2S performs poorly. Pre-trained models from the BERT family can offer some help though limited. It is probably because the pre-training data is from other domains (e.g., news and online encyclopedia), where the representations learned cannot fully reflect the styles of social media languages.

We then observe copy mechanism and latent topics (learn from posts) are both useful, where the former allows the keyword extracted from the post to form a question while the latter further helps find topic words to be copied. On the contrary, user comments, though able to provide useful information, are noisy (also implied by Table 5.1). So, it is important to encode the comments in an appropriate way — CMT (NTM) captures salient topic features from the comments and performs much better than CMT (BIATT), which might be hindered by the noise and exhibit the second worst results.

In addition, we notice DUAL DEC slightly outperforms its single decoder variant CMT(NTM), though the gain is small. To better examine their prediction results, we conduct human evaluations.

Human Ratings. Here we sampled 400 source posts (and their outputs), and invited four PhD students (native Chinese speakers) to rate the poll questions in a 4 point Likert scale — 0 for extremely bad, 1 for bad, 2 for good, and 3 for extremely good — without knowing where the results come from. Each annotator reviews 100 samples and one's assignments vary with others' and Table 5.3 shows the average ratings over the four annotators.

All the models are rated worse than the gold standard, which means

	Relevance	Fluency	Engagingness
Gold Standard	2.79	2.84	2.74
BASE	1.26	2.14	1.35
ROBERTA	1.33	1.06	0.96
TOPIC	1.81	1.66	1.50
CMT (NTM)	1.91	1.67	1.55
DUAL DEC	2.02	1.87	1.67

Table 5.3: Average human ratings. Higher scores indicate better results. DUAL DEC exhibits good potential generate questions likely to draw user engagements.

automatic poll question generation still has a long way to go. We also observe that models with latent topics exhibit relatively better relevance. This may be because topic models allow the capture of salient contents from the input and detail injection to the output. Besides, CMT (NTM) and DUAL DEC perform the best in engagingness, probably because user comments and poll answers might provide implicit clues (e.g., fresh words) helpful to predict engaging questions. For fluency, BASE outperforms our models by a small margin, as it tends to yield short and generic questions, such as "你怎么看" (*What's your viewpoint?*) based on our observation. Moreover, we measure the length of questions generated by BASE and DUAL (our full model) and find that 11.0% questions generated by BASE contain less than 5 words whereas the number for DUAL is only 1.6%. This again demonstrates our potential to generate longer questions with richer details.

5.5.2 Effects of Post and Question Length

We further quantify the question generation results over varying lengths of source posts and poll questions and show the corresponding ROUGE-1 scores in Figure 5.4. Here, we compare BASE and ROBERTA, TOPIC, and our CMT (NTM).[11]

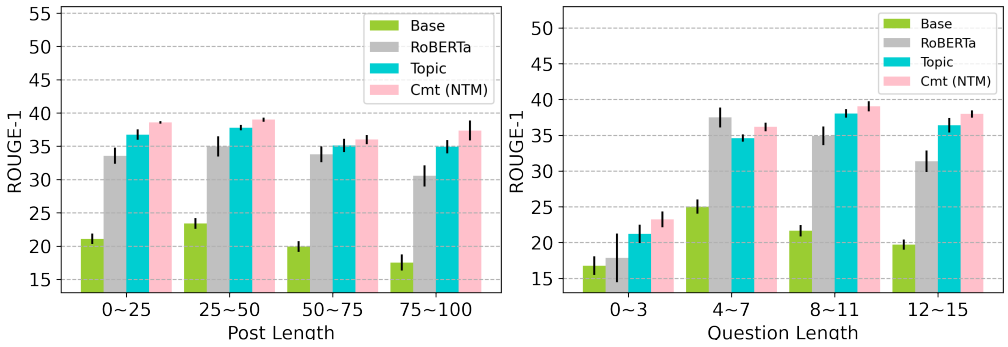


Figure 5.4: ROUGE-1 scores (y-axis) over varying length (word count in x-axis) of source posts (on the left) and poll questions (on the right). For both subfigures, the bars from the left to right shows the results of BASE, ROBERTA, TOPIC, and CMT (NTM).

Post length seems not to affect much on the models' performance, probably attributed to the length limitation in Weibo — even the relatively longer posts contain limited words. On the contrary, for the question length, the two S2S baselines both exhibit obvious performance drops when generating long questions, while TOPIC and CMT (NTM) perform steadily. This suggests that latent topics, either captured from posts or comments, may have the potential to enrich questions with

[11] In §5.5.2 and §5.5.3, we experiment in the single decoder settings so as to focus on the quality of generated questions. We will further discuss the dual decoders in §5.5.4.

detailed descriptions, and hence can better tackle long questions. Nevertheless, CMT (NTM) presents consistently better ROUGE-1 in diverse scenarios.

5.5.3 Polls Questions vs. User Engagements

As shown in the human ratings (§5.5.1), comments might help to generate engaging poll questions. For a further discussion, Figure 5.5 shows the ROUGE-1 of ROBERTA, TOPIC, and CMT (NTM) in handling questions for polls that later engage varying user comment numbers. Interestingly, CMT (NTM) performs better when predicting questions that engage more comments at the end. This means that early comments might provide useful clues for models to distinguish attractive questions with the potential to draw more public engagements in the future. Lacking the ability to learn from comments, TOPIC exhibits relatively more stable trends.

5.5.4 Discussion on Dual Decoders

The previous two subsections are discussed in the single decoder setting and here we further examine the effectiveness to jointly predict questions and answers. BASE, COPY, TOPIC, and CMT (NTM) with single and dual decoders are discussed.

We first compare question generation results and Figure 5.6 shows

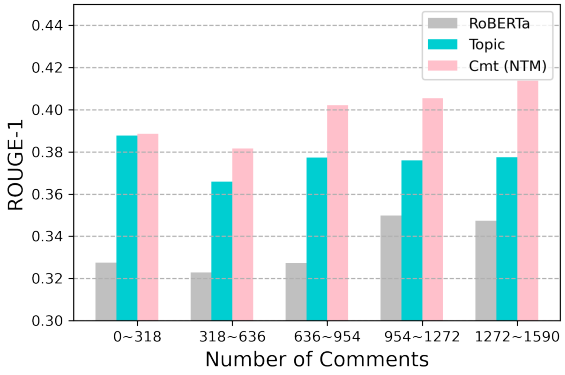


Figure 5.5: Model performance in handling polls that result in varying comment numbers (x-axis). Y-axis: ROUGE-1. Bars from left to right represent RoBERTa, Topic, and Cmt (NTM).

the ROUGE-1 scores. It is seen that dual decoders can boost the results of Base and Copy, implying that questions and answers are indeed related and exploiting their interactions can successfully bring performance gain. However, we cannot observe large-margin improvements in Topic and Cmt (NTM), probably because many words in answers, such as "赵粤" (*Akira*) and "希林娜依高" (*Curley G*) in Figure 5.1, are also topic words that can be discovered with topic models. Therefore, jointly generating answers only provides limited help to their question generation results.

Then, we analyze how the multitask learning ability of dual decoders influence the prediction of poll answers. Table 5.4 displays the comparison results with pipeline models that sequentially generate questions and then answers. By examining the pipeline results, we first find that source posts are helpful in answer generation, which results in the out-

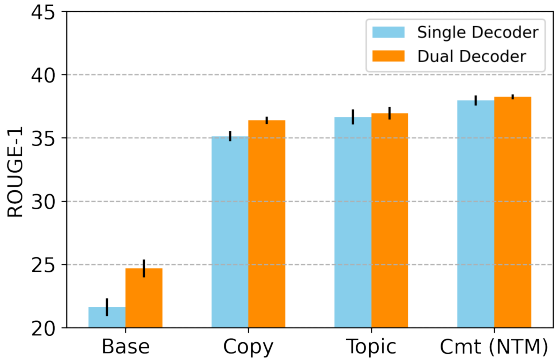


Figure 5.6: ROUGE-1 scores of BASE, COPY, TOPIC, and CMT (NTM) from left to right. For each model, left bars (in blue) shows them in single decoder setting while the right bars (in orange) dual decoders.

performance of PT+QS over QS ONLY. Besides, answer generation trained with predicted questions or the gold standards do not make much difference. Gold standard questions might exhibit higher quality while predicted questions may better fit the tests (answer choices should be predicted without knowing the human-crafted questions).

For dual decoders, CMT (NTM) still performs the best, implying that latent topics from user comments can also contribute to better prediction of poll answers. In comparison with the best pipeline model (PT+QS), the scores from CMT (NTM) are competitive, though the dual decoder allows end-to-end training and is easier to be used (with less manual efforts in model training and application).

MODEL	ROUGE-1	ROUGE-L	BLEU-1	BLEU-3
Pipeline Models				
QS ONLY (PRED)	26.65±0.2	25.09±0.2	22.50±0.8	4.27±0.5
QS ONLY (GOLD)	25.51±0.5	24.17±0.4	22.43±0.3	3.76±0.3
PT+QS (PRED)	31.29±0.6	29.18±0.5	26.35±0.1	8.15±0.3
PT+QS (GOLD)	31.78±0.6	29.63±0.6	26.39±0.6	8.14±0.3
Dual Decoders				
BASE	24.68±0.7	22.59±0.5	21.38±0.3	3.22±0.4
+COPY	30.03±0.5	28.02±0.5	25.55±0.5	8.28±0.3
+TOPIC	30.56±0.8	28.49±0.8	26.00±0.5	8.26±0.4
+CMT (NTM)	31.72±0.7	29.54±0.7	26.55±0.2	8.65±0.2

Table 5.4: The comparison results of models with dual decoders (on the bottom half) and pipeline models (on the top). For the pipeline models, we first produce questions (QS) using CMT (NTM), from which we further generate answers with the S2S model. QS ONLY is fed with QS only while PT+QS the concatenated sequence of posts (PT) and QS. In the training of answer generation, PRED means the predicted questions are employed as input while for GOLD, we adopt gold standard questions (they are assumed to be unavailable for test).

5.5.5 Case Study

To provide more insights, we further take the two Weibo posts in Figure 5.1 as the input cases and examine the output of varying models in Table5.5.[12]

Unsurprisingly, BASE tends to yield generic questions as limited features are encoded from the noisy source. ROBERTA sometimes produces repeated words (e.g., its output to P_1), hindering its capability to generate fluent language (also indicated by Table 5.3). This is possibly caused by the overfitting problem as RoBERTa might rely on large-scale

[12]Here we analyze the case with two examples while similar observations can be drawn from many output cases. More cases will be discussed in Figure 5.6 (in the Appendix).

BASE	你会看吗 (*Would you watch*)
ROBERTA	你平时喜欢哪个视频频频 (Which videooooo do you usually like)
TOPIC	你平时常用哪个视频 (*Which video do you usually use*)
CMT (NTM)	你平时在哪个视频网站 (*Which video site are you on*)
DUAL DEC	你平时用哪个视频 app (*Which video app do you usually use*) >bili 哔哩 (*Bilibili*); 爱奇艺 (*iQiyi*); 腾讯视频 (*Tencent Video*); 芒果tv (*Mango TV*); 优酷 (*Youku*); 其他评论区补充 (*Comment with other choices*)
BASE	你觉得谁的表现更强 (*Who do you think is better*)
ROBERTA	你觉得谁更好 (*Who do you think is better*)
TOPIC	你觉得谁出道了 (*Who do you think debuted*)
CMT (NTM)	你觉得谁更适合c位 (*Who do you think is more suitable for the center position*)
DUAL DEC	你觉得赵粤和希林娜依高谁更可 (*Who do you prefer, Akira or Curley G*) >赵粤 (*Akira*); 希林娜依高 (*Curley G*)

Table 5.5: Questions generated for the source posts in Figure 5.1: P_1 (top) and P_2 (bottom). For DUAL DEC (i.e., CMT (NTM) with dual decoders), the question is followed by the answer in the next row.

in-domain data for fine-tuning.

We also find that modeling topics and user comments may enable the output to contain trendy wordings, making it more engaging, such as "c位" (*center point*) in CMT (NTM)'s output question for P_2 and the names of many new video apps in DUAL DEC's generated answer choices for P_1. Furthermore, the dual decoders might learn the cohesive relations between questions and answers, such as the *Akira* and *Curley G* occurring in both the generated questions and answer choices (P_2).

5.6 Conclusion

We have presented a novel task to generate social media poll questions. User comments encoded with a neural topic model are leveraged in a S2S framework; dual decoder architecture is further adopted to explore the interactions between questions and answers. Extensive experiments on a large-scale dataset newly collected from Weibo have demonstrated the effectiveness of our proposed model.

5.7 Appendix

[$Post$]: #2020百大最美女星#刘亦菲和迪丽热巴都上榜啦！！！都是天然美女啊～两个人一个人演过电影版的三生三世，一个演过剧版的三生三世。 (*#100 Most Beautiful Women in the World 2020# Liu Yifei and Dilraba Dilmurat are both on the list!!! Both of them are natural beauties˜One of them played in the movie Eternal Love while the other played in its TV series version*)
[$Question$]: 谁的颜让你心动呢 (*Whose face makes you heart flip*)
[$Answer$]: 刘亦菲 (*Liu Yifei*); 迪丽热巴 (*Dilraba Dilmurat*)
[$Base$]: 你最喜欢谁 (*Who do you like the best*)
[$RoBERTa$]: 你更喜欢谁 (*Who do you prefer*)
[$Topic$]: 你更喜欢哪一个(*Which one do you prefer*)
[$Cmt(NTM)$]: 你更喜欢谁的造型 (*Whose look do you prefer*)
[$DualDec$]: 你觉得谁更有cp感 (*Who do you think is better coupled with the leading man*)
>刘亦菲 (*Liu Yifei*); 迪丽热巴 (*Dilraba Dilmurat*)

[$Post$]: 有意见建议同性婚姻合法化写入民法典 (*Some people suggest that same-sex marriage be legalized into the Civil Code*)
[$Question$]: 你支持同性恋结果合法化吗 (*Do you support the legalization of same-sex marriage*)
[$Answer$]: 同意 (*Agree*); 不同意 (*Disagree*)
[$Base$]: 你怎么看 (*What do you think*)
[$RoBERTa$]: 你支持同性结婚化吗 (*Do you support the same-sex marriage*)
[$Topic$]: 你支持同性恋合法化吗 (*Do you support the legalization of homosexuality*)
[$Cmt(NTM)$]: 你支持同性恋婚姻合法化吗 (*Do you support the legalization of the same-sex marriage*)
[$DualDec$]: 你支持同性恋婚姻合法化吗(*Do you support the legalization of the same-sex marriage*)
>支持 (*Support*); 不支持 (*Objection*)

[$Post$]: #瑞幸咖啡伪造交易22亿# 在否认业绩造假两个月后，瑞幸今日盘前发布公告：内部调查显示，从2019年第二季度到2019年第四季度与虚假交易相关的总销售金额约为22亿元。于是，#瑞幸暴跌# 。 (*#Ruixing Coffee forged 2.2 billion transactions# Two months after denying fraud, Luckin released an announcement before the market today: An internal investigation showed that total sales related to invalid transactions from the second quarter of 2019 to the fourth quarter of 2019 amounted to about 2.2 billion Yuan. Consequently, #Luckin Coffee stock plummet#*)
[$Question$]: 你还会喝瑞幸咖啡吗 (*Will you still drink Luckin coffee*)
[$Answer$]: 会，我券还没用完呢 (*Yes. I still have the coupons to use*); 不会，没券就不喝 (*No. No coupon, no coffee.*); 从来就没有喝过 (*I've never drunk the coffee there*); 不管如何都是死忠粉 (*Die-hard fan no matter what*)
[$Base$]: 你会买 iphone 吗 (*Would you buy an iphone*)
[$RoBERTa$]: 你喝过瑞幸咖啡吗 (*Have you ever drunk Luckin coffee*)
[$Topic$]: 你会买瑞幸咖啡吗 (*Would you buy Luckin coffee*)
[$Cmt(NTM)$]: 你觉得瑞幸咖啡合理吗 (*Do you think Luckin Coffee is reasonable*)
[$DualDec$]: 你还会买瑞幸咖啡吗 (*Will you still buy Luckin coffee*)
>会 (*Yes*); 不会 (*No*); 看情况 (*It depends*)

[$Post$]: 杨丽萍因为没有结婚生孩子，过着与花草舞蹈为伴的生活，被网友diss是一个失败的范例，真正的女人应该要儿孙满堂，才是幸福的。 (*Yang Liping, who has no marriage or children, lives a life with flowers and dancing. However, she has been ridiculed by netizens and viewed as a typical loser — a real woman should have a large family of children and grandchildren to live in happiness.*)
[$Question$]: 如何定义成功女性(*How to define a successful woman*)
[$Answer$]: 事业有成 (*Success in career*); 儿孙满堂 (*Have children and grandchildren*); 家庭事业双丰收(*Success in family and career*); 充实的灵魂 (*Interesting soul*)
[$Base$]: 你觉得哪种行为有问题 (*What kind of behavior do you think is problematic*)
[$RoBERTa$]: 女女是女人是女人是什么 (*What is woman is woman*)
[$Topic$]: 你觉得结婚应该定义成功吗 (*Do you think marriage should come to define success*)
[$Cmt(NTM)$]: 你怎么看待成功的女性杨丽萍 (*How do you think of the successful woman Yang Liping*)
[$DualDec$]: 你觉得如何定义成功女性 (*How would you define successful women*)
> 应该 (*Should*); 不支持(*Objection*); 评论区补充 (*Add more details in comments*)

[$Post$]: #杨幂魏大勋恋情实锤# 杨幂魏大勋恋情再次被实锤，现在已经成了圈子内外不是秘密的秘密了。 (*#Smoking gun of Yang Mi and Wei Daxun# Yang Mi and Wei Daxun's love affair has been verified again, and it has now become a secret inside and outside the circle.*) [$Question$]: 你看好杨幂魏大勋的恋情吗(*Are you optimistic about Yang Mi's romantic relationship with Wei Daxun*) [$Answer$]: 看好 (*Optimistic*); 不看好 (*Pessimistic*); 有波折终能修成正果 (*There will be twists and turns but the ending will be good*) [$Base$]: 你觉得这个做法怎么样 (*What do you think of this approach*) [$RoBERTa$]: 你觉得魏魏勋勋恋爱吗(*Do you think Wei Wei Xun Xun is in love*) [$Topic$]: 你觉得谁更渣 (*Who do you think is more scummy*) [$Cmt(NTM)$]: 你怎么看待这恋情的 (*What do you think of the romantic relationship*) [$DualDec$]: 你觉得杨幂魏大勋有必要吗 (*Do you think Yang Mi and Daxun Wei are necessary to do so*) $>$杨幂 (*Yang Mi*); 魏大勋 (*Wei Daxun*); 都不喜欢 (*Do not like either of them*); 吃瓜 (*I'm an onlooker*)

Table 5.6: Five additional cases. One block refers to one case, including its source post ($Post$), ground truth question ($Question$) and answer ($Answer$), followed by and the results generated by varying models (model names are in []). For answers, different choices are separated by ";" and the outputs of $DualDec$ appear after a $>$. Italic words in "()" are the English translation of the original Chinese texts on their left.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

With the worldwide expansion of the Internet, interpersonal communications have been gradually moved to the virtual world. Despite the essential need to talk with others on social media, some individuals may not experience good online interactions because of the informal language styles adopted by other online users, uncontrollable environments in online discussions, and the reluctance to explicitly engage in social media conversations. To allow a better online engagement experience in social interactions, we have proposed to employ natural language processing methods to assist users to understand others' posts, get on the track of the conversation context for those involved in discussions, and draw user

engagement for those reluctant to talk.

This thesis mainly consists of three research work.

In the first work, we build a social media understanding benchmark to investigate the NLP models' ability to handle the social media language. We are the first to study the social media language understanding benchmark with datasets on both fundamental and popular social media tasks presented to evaluate how models understand social media language. The experimental result shows that the overall understanding ability of the NLP machine is better than human beings while it also uncovers the limitation of state-of-the-art models in handling post-level context.

In the second and third work, we present two application scenarios to help the online user engagement — one is for users who have been involved in a discussion and the other is for those reluctant to explicitly interact with others.

For those involved in discussions, we measure the quality of user replies with a novel concept of conversation residual life, which reflects the number of coming turns to occur in the future. To model the conversation discourse and measure the environments, we leverage a hierarchical neural model that jointly explores indicative representations from the content in turns and the structure of conversations in an end-to-end man-

ner. It can be used to keep track of dialogue and thus help response selection to move the discussion forward.

Another application is poll question generation for engaging people who may be reluctant to join a discussion and thus help them to interact more actively, which is a new task on social media. We explore new features from comments by the neural topic model and demonstrate that it can help the S2S model improve the quality of generating questions. We also have explored multitask learning on this model with the question and answer jointly trained and demonstrated that multitask learning works for this task.

6.2 Future Work

This thesis has studied how to use NLP methods to improve user engagements on social media. This is an important topic worth long-term exploration and we will point out some interesting directions of the future work here.

For the work of social media language understanding benchmark, more social media applications can be incorporated to test the understanding ability from a more comprehensive point of view. Specifically, we only consider classification tasks in this work yet ignore generation tasks. Since the generation ability is downstream of understanding, we

can consider testing the encoder from the decoder's standpoint, such as hashtag generation and question generation. Also, some techniques and external knowledge can be proposed or introduced to design or improve the understanding model's inference ability. Limited efforts have been made to explore how the pre-trained models can learn useful language representations from the noisy social media data. Especially a dynamic and scalable vocabulary table can be considered in this design because of the limited attention paid to this issue as far aw we know.

For the estimation of residual life for conversations, we can explore a better modeling structure in the future. The existing model simulates the thread and reads the message by turns. However, sometimes one sentence will kill the conversation but nothing with the chatting history. So, separately modeling the history turns and the current message is a potential attempt to improve the performance. Besides, the preprocessing of the dataset can be improved. We randomly split the chatting log and sort by the number of the rest turns. But how to map it to long or short residual life requires a more wise method, rather than equal segments according to rest turns. What is more, to make suggestion for the improper behavior users is also a meaningful topic.

For the question generation task, some other techniques can be considered to control the sentence diversity or difficulty, such as GANs and

reinforcement learning. And then the next step, where the question is better for the post to improve people's engagement. This question can be cooperated with 'repost' and 'like' data to model a more charming question. Besides, multi-modal is a new way to explore the question since picture usually provides more vivid information, thus illustrating pictures from the post as input is a potential way to enhance the understanding ability. On the other side, generating a question and a funny illustration picture simultaneously is also a more attractive way. From a practice standpoint, semi-supervising is a potential method because question label is not always easily collected in many scenarios.

Bibliography

[1] Muhammad Abdul-Mageed and Lyle H. Ungar. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 718–728. Association for Computational Linguistics, 2017.

[2] Yaser S. Abu-Mostafa. Learning from hints in neural networks. *J. Complex.*, 6(2):192–198, 1990.

[3] Eleftheria Ahtaridis, Christopher Cieri, and Denise DiPersio. LDC language resource database: Building a bibliographic database. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 1723–1728. European Language Resources Association (ELRA), 2012.

[4] Mario Ezra Aragón, Adrián Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y-Gómez. Detecting depression in social media using fine-grained emotions. In Jill

Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1481–1486. Association for Computational Linguistics, 2019.

[5] Yoav Artzi, Patrick Pantel, and Michael Gamon. Predicting responses to microblog posts. In *NAACL*, pages 602–606, 2012.

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

[7] Haoli Bai, Zhuangbin Chen, Michael R. Lyu, Irwin King, and Zenglin Xu. Neural relational topic models for scientific article analysis. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 27–36. ACM, 2018.

[8] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Mach. Learn.*, 28(1):7–39, 1997.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information*

Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada], pages 601–608. MIT Press, 2001.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[11] Yang Trista Cao, Sudha Rao, and Hal Daumé III. Controlling the specificity of clarification question generation. In *Proceedings of the 2019 Workshop on Widening NLP@ACL 2019, Florence, Italy, July 28, 2019*, pages 53–56. Association for Computational Linguistics, 2019.

[12] Zi Chai and Xiaojun Wan. Learning to ask more: Semi-autoregressive sequential question generation under dual-graph interaction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online,*

July 5-10, 2020, pages 225–237. Association for Computational Linguistics, 2020.

[13] Yllias Chali and Sadid A. Hasan. Towards topic-to-question generation. *Comput. Linguistics*, 41(1):1–20, 2015.

[14] Yi Chang, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. Towards twitter context summarization with user influence models. In *WSDM*, pages 527–536, 2013.

[15] Uttam Chauhan and Apurva Shah. Topic modeling using latent dirichlet allocation: A survey. *ACM Comput. Surv.*, 54(7):145:1–145:35, 2022.

[16] Hao Cheng, Hao Fang, and Mari Ostendorf. A dynamic speaker model for conversational interactions. In *NAACL*, pages 2772–2785. Association for Computational Linguistics, 2019.

[17] Vladimir Cherkassky. The nature of statistical learning theory. *IEEE Trans. Neural Networks*, 8(6):1564, 1997.

[18] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014.

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

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

[21] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018.

[22] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics, 2019.

[23] 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*, pages 76–87, 2011.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.

[25] Kaustubh D. Dhole and Christopher D. Manning. Syn-qg: Syntactic and shallow semantic rules for question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 752–765. Association for Computational Linguistics, 2020.

[26] Keyang Ding, Jing Li, and Yuji Zhang. Hashtags, emotions, and comments: A large-scale dataset to understand fine-grained social emotions to online topics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1376–1382, 2020.

[27] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1342–1352. Association for Computational Linguistics, 2017.

[28] Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*,

pages 4508–4513. Association for Computational Linguistics, 2020.

[29] Yifan Gao, Jianan Wang, Lidong Bing, Irwin King, and Michael R. Lyu. Difficulty controllable question generation for reading comprehension. *CoRR*, abs/1807.03586, 2018.

[30] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 42–47. The Association for Computer Linguistics, 2011.

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

[32] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics, 2020.

[33] Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. Kornli and korsts: New benchmark datasets for korean natural language understanding. In Trevor Cohn, Yulan

He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 422–430. Association for Computational Linguistics, 2020.

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

[35] Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 609–617. The Association for Computational Linguistics, 2010.

[36] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *WWW*, pages 57–58, 2011.

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

[38] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics, 2018.

[39] Wenpeng Hu, Bing Liu, Jinwen Ma, Dongyan Zhao, and Rui Yan. Aspect-based question generation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver,*

BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings. OpenReview.net, 2018.

[40] Hafedh Hussein, Mohammed Elmogy, and Shawkat Guirguis. Automatic english question generation system based on template driven scheme. *International Journal of Computer Science Issues (IJCSI)*, 11(6):45, 2014.

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

[42] Mohamed Khalifa and Vanessa Liu. Online consumer retention: contingent effects of online shopping habit and online shopping experience. *Eur. J. Inf. Syst.*, 16(6):780–792, 2007.

[43] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. Improving neural question generation using answer separation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6602–6609. AAAI Press, 2019.

[44] Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian NLP. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 757–770. International Committee on Computational Linguistics, 2020.

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

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

[47] Igor Labutov, Sumit Basu, and Lucy Vanderwende. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 889–898. The Association for Computer Linguistics, 2015.

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

[49] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

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

[51] Haizhou Li and Baosheng Yuan. Chinese word segmentation. In *Proceedings of the 12th Pacific Asia conference on language, information and computation*, pages 212–217, 1998.

[52] Jing Li, Ming Liao, Wei Gao, Yulan He, and Kam-Fai Wong. Topic extraction from microblog posts using conversation structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[53] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *ACL*, pages 1106–1115, 2015.

[54] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[55] David Lindberg, Fred Popowich, John C. Nesbit, and Philip H. Winne. Generating natural language questions to support learning on-line. In *ENLG 2013 - Proceedings of the 14th European Workshop on Natural Language Generation, August 8-9, 2013, Sofia, Bulgaria*, pages 105–114. The Association for Computer Linguistics, 2013.

[56] Bingran Liu. Neural question generation based on seq2seq. In *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence*, pages 119–123, 2020.

[57] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and

Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, abs/1907.11692, 2019.

[58] Ying Luo, Fengshun Xiao, and Hai Zhao. Hierarchical contextualized representation for named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8441–8448. AAAI Press, 2020.

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

[60] Graham McAllister and Gareth R. White. Video game development and user experience. In Regina Bernhaupt, editor, *Game User Experience Evaluation*, Human-Computer Interaction Series, pages 11–35. Springer, 2015.

[61] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep keyphrase generation. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 582–592. Association for Computational Linguistics, 2017.

[62] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In

Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR, 2017.

[63] Heather L. O'Brien and Elaine G. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *J. Assoc. Inf. Sci. Technol.*, 59(6):938–955, 2008.

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

[65] Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. Reinforced dynamic reasoning for conversational question generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2114–2124. Association for Computational Linguistics, 2019.

[66] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.

[67] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*

2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 2227–2237. Association for Computational Linguistics, 2018.

[68] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP*, 2011.

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

[70] Matthew Rowe, Sofia Angeletou, and Harith Alani. Anticipating discussion activity on community forums. In *SocialCom*, pages 315–322, 2011.

[71] Matthew Rowe, Sofia Angeletou, and Harith Alani. Predicting discussions on the social semantic web. In *ESWC*, pages 405–420, 2011.

[72] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. KLEJ: comprehensive benchmark for polish language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1191–1201. Association for Computational Linguistics, 2020.

[73] 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*, pages 7–55. Elsevier, 1978.

[74] 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.

[75] Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6027–6032. Association for Computational Linguistics, 2019.

[76] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics, 2017.

[77] 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*, pages 6322–6333, 2020.

[78] 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*, pages 2210–2219, 2017.

[79] Sanuj Sharma, Prafulla Kumar Choubey, and Ruihong Huang. Im-

proving dialogue state tracking by discerning the relevant context. In *NAACL*, pages 576–581, 2019.

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

[81] Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. CED: credible early detection of social media rumors. *IEEE Trans. Knowl. Data Eng.*, 33(8):3035–3047, 2021.

[82] 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*, pages 177–184, 2010.

[83] Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3930–3939. Association for Computational Linguistics, 2018.

[84] Yibo Sun, Duyu Tang, Nan Duan, Tao Qin, Shujie Liu, Zhao Yan, Ming Zhou, Yuanhua Lv, Wenpeng Yin, Xiaocheng Feng, Bing Qin, and Ting Liu. Joint learning of question answering and question generation. *IEEE Trans. Knowl. Data Eng.*, 32(5):971–982, 2020.

[85] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, 2019.

[86] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.

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

[88] Duyu Tang, Nan Duan, Tao Qin, and Ming Zhou. Question answering and question generation as dual tasks. *CoRR*, abs/1706.02027, 2017.

[89] Anne Tissen. A case-based architecture for A dialogue manager for information seeking. In *SIGIR*, pages 152–161, 1991.

[90] 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*, pages 1273–1278, 2019.

[91] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[92] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019.

[93] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupala, and Afra Alishahi, editors, *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355. Association for Computational Linguistics, 2018.

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

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

[96] Tong Wang, Xingdi Yuan, and Adam Trischler. A joint model for question answering and question generation. *CoRR*, abs/1706.01450, 2017.

[97] Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu,

and Shuming Shi. Topic-aware neural keyphrase generation for social media language. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2516–2526. Association for Computational Linguistics, 2019.

[98] Yue Wang, Jing Li, Irwin King, Michael R. Lyu, and Shuming Shi. Microblog hashtag generation via encoding conversation contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1624–1633. Association for Computational Linguistics, 2019.

[99] Matthijs J. Warrens. Cohen's linearly weighted kappa is a weighted average. *Adv. Data Anal. Classif.*, 6(1):67–79, 2012.

[100] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics, 2020.

[101] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou,

and Wei-Ying Ma. Topic aware neural response generation. In *AAAI*, pages 3351–3357, 2017.

[102] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A chinese language understanding evaluation benchmark. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4762–4772. International Committee on Computational Linguistics, 2020.

[103] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. The penn chinese treebank: Phrase structure annotation of a large corpus. *Nat. Lang. Eng.*, 11(2):207–238, 2005.

[104] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019.

[105] Kaichun Yao, Libo Zhang, Tiejian Luo, Lili Tao, and Yanjun Wu. Teaching machines to ask questions. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4546–4552. ijcai.org, 2018.

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

[107] 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*, pages 55–63, 2016.

[108] 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.

[109] Xingdi Yuan, Tong Wang, Çaglar Gülçehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 15–25. Association for Computational Linguistics, 2017.

[110] Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. What you say and how you say it: Joint modeling of topics and discourse in microblog conversations. *Trans. Assoc. Comput. Linguistics*, 7:267–281, 2019.

[111] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3120–3131. Association for Computational Linguistics, 2018.

[112] Bo Zhang and Xiaoming Zhang. Hierarchy response learning for neural conversation generation. In *EMNLP*, pages 1772–1781, 2019.

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

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

[115] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics, 2019.

[116] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels,*

Belgium, October 31 - November 4, 2018, pages 3901–3910. Association for Computational Linguistics, 2018.

[117] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 662–671. Springer, 2017.