Engage the Public: Poll Question Generation for Social Media Posts

Zexin Lu1  Keyang Ding1  Yuji Zhang1  Jing Li1  Baolin Peng2  Lemao Liu3
1Department of Computing, The Hong Kong Polytechnic University, HKSAR, China
2Microsoft Research, Redmond, WA  3Tencent AI Lab, Shenzhen, China
1{zexin.lu, keyang.ding, yu-ji.zhang}@connect.polyu.hk
1jing-amelia.li@polyu.edu.hk  2bapeng@microsoft.com  3redmondliu@tencent.com

Abstract

This paper presents a novel task to generate poll questions for social media posts. It offers an easy way to hear the voice of the public and learn from their feelings to important social topics. While most related work tackles formally-written texts (e.g., exam papers), we generate poll 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 poll choices (answers). For experiments, we collect a large-scale Chinese dataset from Sina Weibo containing over 20K polls. The results show that our model outperforms the popular S2S models without exploiting topics from comments and the dual decoder design can further benefit the prediction of both questions and answers. Human evaluations further exhibit our superiority in yielding high-quality polls helpful to draw user engagements.

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 1 shows two example poll questions on Sina Weibo1, henceforth Weibo, a popular Chinese microblog. The goal of our task is to output an opinion question, such as Q1 and Q2, and invite other users to engage in the discussion to a source post (e.g., P1 and P2); poll choices (answers like A1 and A2) 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 (Du et al., 2017). Despite of the extensive efforts made in this line (Sun et al., 2018; Yao et al., 2018; Chai and Wan, 2020; Sun et al., 2020), most previous work focus on the processing of formally-written texts, such as exam questions

\[ \text{P1: } \ldots \text{市场价值超过爱奇艺 (The market value of B site exceeds iQiyi).}\]
\[ \text{Q1: } \text{你们平时常用哪个app看视频? (Which app do you usually use to watch videos?)}\]
\[ \text{A1: } \text{腾讯视频 (Tencent Video); 优酷 (Youku); 爱奇艺 (iQiyi); B站 (B site)}\]

\[ \text{P2: } \ldots \text{理性分析一下赵粤和希林娜依高: 希林vocal确实厉害，但是...舞蹈实力有点不够看: 赵粤呢舞蹈厉害...但是唱歌实力较弱些... (A rational analysis of Akira and Curley G: Curley's vocal is indeed great, but ... her dancing is not that good; Akira dances well ... but her singing is weaker...)}\]
\[ \text{Q2: 谁更适合当C位? (Who should take the center position?)}\]
\[ \text{A2: 赵粤 (Akira); 希林娜依高 (Curley G)}\]

Figure 1: Example polls from Sina Weibo. P1, Q1, and A1 (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 (Wang et al., 2019b), 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 “爱” (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 consisting around 20K Weibo polls associated with their source posts and user comments. 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 (Liu et al., 2019) 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 shown that joint prediction of questions and their answers can benefit both tasks.

2 Study Design

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.

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.

Our dataset and code are publicly available in https://github.com/polyusmart/Poll-Question-Generation
Table 1: Statistics of our dataset. Num: number; Len: average number per post; Qs: question; Ans: answer.

| Post Num | Len | Comment Num | Len | Qs | Ans Choice Num | Len | Voter Num |
|----------|-----|-------------|-----|----|----------------|-----|-----------|
| 20,252   | 54.0| 173         | 16.9|   | 14             | 5.9 | 13,004    |

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

Data Analysis. The statistics of the dataset is displayed in Table 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 noisier than original posts; both questions and answers are short, which follow the fashion of user-generated contents on social media.

To further investigate the data sparsity in social media contents, we sample some texts from LDC news corpus (formally-written texts) — 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. 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 2(a) shows the count of polls over 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 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.

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 3.

3.1 Source Posts and Comments Encoding

Following the common practice in S2S (Du et al., 2017), we encode a source post $P$ in the form of word sequence $\langle w_1, w_2, \ldots, w_|P| \rangle$, 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 (BiGRU) (Cho et al., 2014) is adopted. For the $i$-th word $w_i \in P$, we first convert it into an embedding vector $\nu_i$, which is later processed into hidden
which is further taken to produce the topic composition of comments \( \theta \) with softmax transformation.

At last, the decoder reconstructs comments and produces a BOW vector \( C_{bow}^t \) (conditioned on the latent topic \( \theta \)) through another neural perception.

### 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 \( \theta \) from user comment modeling. The words in \( Q \) are predicted sequentially with the following formula:

\[
Pr(Q|P, C_{bow}) = \prod_{j=1}^{\text{int}} Pr(q_j|q_{<j}, M, \theta) \tag{1}
\]

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

\[
\alpha_{ij} = \frac{\exp(f_n(h_i, s_j, \theta))}{\sum_{i'=1}^{\text{int}} \exp(f_n(h_{i'}, s_j, \theta))} \tag{2}
\]

\( s_j \) is the GRU decoder’s \( j \)-th hidden states and:

\[
f_n(h_i, s_j, \theta) = v^T_i \tanh(W_n[h_i; s_j; \theta] + b_n) \tag{3}
\]

In addition, we adopt copy mechanism (See et al., 2017) to allow the generated questions to contain the keywords from the source posts:

\[
p_j = \lambda_j \cdot p_{gen} + (1 - \lambda_j) \cdot p_{copy} \tag{4}
\]

\( p_{gen} \) 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:

\[
\lambda_j = \text{sigmoid}(W_\lambda [u_j; s_j; t_j; \theta] + b_\lambda) \tag{5}
\]

\( t_j \) is the context vector (weighted sum) of the attention to predict the \( Q \)’s \( j \)-th word, whose embedding is \( u_j \). \( W_\lambda \) and \( b_\lambda \) 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. 2) as the question decoder (denoted as \( \beta_{ij} \) here) and copy mechanisms (Eq. 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 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.

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:

\[
L_{NTM} = D_{KL}(p(z) \parallel q(z|C)) - E_{q(z|C)}[p(C|z)]
\]

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:

\[
L_{QG} = \sum_{n=1}^{N} \log(P_{Q}(Q_{n}|P_{n}, \theta_{n}))
\]

\( N \) is the number of training samples; \( Q_{n}, P_{n}, \) and \( \theta_{n} \) are the target poll question, source post, and topic distribution of the \( n \)-th training sample. Answer generation loss \( L_{AG} \) is defined similarly. The training loss of the entire model are defined as:

\[
L = L_{NTM} + \gamma_{Q} \cdot L_{QG} + \gamma_{A} \cdot L_{AG}
\]

where \( \gamma_{Q} \) and \( \gamma_{A} \) balance the weights over NTM and the two decoders.

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 (Sutskever et al., 2014) (i.e., BASE); also compared are the S2S with pre-trained models from the BERT family — tiny ERINE (Sun et al., 2019) (i.e., ERINE), BERT (Devlin et al., 2019) (i.e., BERT), and RoBERTa (Liu et al., 2019) (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 mechanism (i.e., COPY) (Meng et al., 2017), topic modeling from posts (i.e., TOPIC) (Wang et al., 2019a), and bidirectional attentions over posts and comments (i.e., CMT (BIATT)) (Wang et al., 2019b). 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 architecture 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.

\( ^8 \)https://github.com/fxsjy/jieba
\( ^9 \)https://www.paddlepaddle.org.cn/hub
\( ^10 \)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 (Gururangan et al., 2020), which is beyond the scope of this paper and will be explored in future work.
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 $\gamma_Q = 1$ (Eq. 8).

**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 (Lin, 2004) and BLEU (Papineni et al., 2002). 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 Experimental Results

In this section, we first show the main comparison results on poll question generation involving both automatic evaluations and human ratings (in §5.1). Then, model sensitivity to varying lengths of source posts and poll questions are discussed in §5.2, followed by the analyses of models’ capability to handle poll questions exhibiting varying degrees of user engagements (§5.3). Next, §5.4 discusses the performance of dual decoders that jointly generate questions and answers. A case study is presented at last (in §5.5) to interpret the sample outputs.

### 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.

**Automatic Evaluations.** Table 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 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 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 3 shows the average ratings over the four annotators.

All the models are rated worse than the gold standard, which means 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 “you think?” (What’s your viewpoint?) based on our observation. More...
over, 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.

5.2 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 4. Here, we compare BASE and RoBERTA, TOPIC, and our CMT (NTM).11

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

5.3 Polls Questions vs. User Engagements

As shown in the human ratings (§5.1), comments might help to generate engaging poll questions. For a further discussion, Figure 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.

Figure 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).

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 6 shows 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 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 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

| Gold Standard | Relevance | Fluency | Engagingness |
|---------------|-----------|---------|--------------|
| 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 3: Average human ratings. Higher scores indicate better results. DUAL DEC exhibits good potential to generate questions likely to draw user engagements.

---

11In §5.2 and §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 §5.4.
Table 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, ϕed means the predicted questions are employed as input while for gold, we adopt gold standard questions (they are assumed to be unavailable for test).

6 Related Work

Our work is in the 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 (Labutov et al., 2015; Dhole and Manning, 2020; Fabbri et al., 2020), largely relying on the labor-intensive process for rule design or feature engineering. To simplify the training, automatic feature learning hence becomes popular. For example, Chali and Hasan (2015) 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, S2S-based question generation architecture has demonstrated promising results (Du et al., 2017; Chai and Wan, 2020). To better encode the input, researchers adopt successful training design from other tasks, such as self-attention mechanism (Zhao et al., 2018; Scialom et al., 2019), language model pre-training (Pan et al., 2019), variational inference (Yao et al., 2018), and reinforcement learning (Yuan et al., 2017; Pan et al., 2019). Heuristic features, e.g., the answers’ positions in the article (Zhou et al., 2017; Sun et al., 2018;...
We have presented a novel task to generate social media poll questions and answers. Extensive experiments on a large-scale dataset newly collected from Weibo have demonstrated the effectiveness of our proposed model.

7 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.

Acknowledgments

This work was partially done when Zexin Lu was an intern at Tencent AI Lab under CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (R-ZDCJ). The research is also supported by NSFC Young Scientists Fund (62006203) and PolyU internal funds (1-BE2W, 4-ZZKM, and 1-ZVRH). The authors would like to thank Lida Li, Yue Wang, Yubo Zhang, Zhe Wang, and anonymous reviewers from ACL-IJCNLP 2021 for their insightful suggestions on various aspects of this work.

Ethical Considerations

The task will not pose ethical problems. First, the polls are open access to the public users (so as to collect their opinions). Second, Weibo allows any users to report suspicious cases with ethical concerns and the reported contents will be removed immediately. Third, the polls are running in an anonymous way to protect the privacy of voters.

The dataset is collected through the official APIs of Weibo and is consistent with the Weibo terms of use. We also manually examined the data to ensure the following points. First, we conduct data anonymization and manually examined the data to ensure there are no privacy and ethical concerns, e.g., personal information, toxic language, and hate speech. In the generated polls, we didn’t spot any cases that might have the concern. Second, the involved Weibo users are all public ones. To that end, we automatically filtered out personal users without the official confirmation of Weibo (the confirmed public users can be identified with a “VIP” tag). The user list is manually checked again to mitigate the ethical concern.

For the annotation, we recruited part-time research assistants to work with the pay 15.7 USD/hour and at most 20 hours per week.

Table 5: Questions generated for the source posts in Figure 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.

| BASE | RoBERTA | Topic | CMT (NTM) | DUAL DEC |
|------|---------|-------|-----------|----------|
| 你会看电影 (Would you watch) & 你平时喜欢哪个视频(Which video do you usually use) & 你平时常看哪个视频 (Which video do you usually watch) & 你平时用哪个视频 app (Which video app do you usually use) |
| 你会选择 (Who do you think is better) & 你觉得谁更好 (Who do you think is better) & 你认为谁最适合 (Who do you think is most suitable for the center position) & 你认为你更喜欢谁 (Who do you prefer, Akira or Carley G) |
| 你会选择 Akira or Carley G & 你会选择 (Who do you think is better) & 你认为谁更适合 (Who do you think is more suitable for the center position) & 你认为你更喜欢谁 (Who do you prefer, Akira or Carley G) |
| >bilibili 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) |

References

Eleftheria Ahtaridis, Christopher Cieri, and Denise DiPersio. 2012. LDC language resource database: Building a bibliographic database. In 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).

Yang Trista Cao, Sudha Rao, and Hal Daumé III. 2019. 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.

Zi Chai and Xiaojun Wan. 2020. 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.

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

Kyunghyun Cho, Bart van Merrienboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Kausubh D. Dhole and Christopher D. Manning. 2020. Syntax 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.

Xinya Du, Junru Shao, and Claire Cardie. 2017. 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.

Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. 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.

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

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Bellagay, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In 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.

Wenpeng Hu, Bing Liu, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2018. 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.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. 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.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. 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.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. arXiv preprint, abs/1907.11692.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. 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 582–592. Association for Computational Linguistics.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. 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.
Boyu Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. 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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. 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 1073–1083. Association for Computational Linguistics.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. 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.

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

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. 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.

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

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

Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu, and Shuming Shi. 2019a. Topic-aware neural keyphrase generation for social media language. 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 2516–2526. Association for Computational Linguistics.

Yue Wang, Jing Li, Irwin King, Michael R. Lyu, and Shuming Shi. 2019b. 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.

Kaichun Yao, Libo Zhang, Tiejian Luo, Lili Tao, and Yanjun Wu. 2018. 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.

Xingdi Yuan, Tong Wang, Çaglar Gülçehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. 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.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. 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.

Yao Zhao, Xiaochuan Ni, Yuan yuan Ding, and Qifa Ke. 2018. 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.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1: Long and Short Papers, pages 3901–3910. Association for Computational Linguistics.
Table 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.