Topic-Guided Self-Introduction Generation for Social Media Users

Chunpu Xu[1], **Jing Li**[1]*, **Piji Li**[2], **Min Yang**[3]
[1] Department of Computing, The Hong Kong Polytechnic University
[2] College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics
[3] Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[1]`chun-pu.xu@connect.polyu.hk`; [1]`jing-amelia.li@polyu.edu.hk`;
[2]`pjli@nuaa.edu.cn`; [3]`min.yang@siat.ac.cn`

Abstract

Millions of users are active on social media. To allow users to better showcase themselves and network with others, we explore the auto-generation of social media *self-introduction*, a short sentence outlining a user's personal interests. While most prior work profiles users with tags (e.g., ages), we investigate sentence-level self-introductions to provide a more natural and engaging way for users to know each other. Here we exploit a user's tweeting history to generate their self-introduction. The task is non-trivial because the history content may be lengthy, noisy, and exhibit various personal interests. To address this challenge, we propose a novel unified topic-guided encoder-decoder (UTGED) framework; it models latent topics to reflect salient user interest, whose topic mixture then guides encoding a user's history and topic words control decoding their self-introduction. For experiments, we collect a large-scale Twitter dataset, and extensive results show the superiority of our UTGED to the advanced encoder-decoder models without topic modeling. [1]

Self-introduction:

Invertebrate Paleontologist and Collection Manager at the Delaware Museum of Natural History.

User previously published tweets (user history):

- How Delaware are you? New book on the 'secret' First State may stump you httpurl **(Delaware)**
- Duck! Octopuses caught on camera throwing things at each other **(invertebrates)**
- Rare fossil #clam discovered alive httpurl **(paleontology)**
- 'A labor of love' | Revamped Delaware Museum of Nature and Science opens its doors to the public again **(Delaware; museum)**
- Delaware's close to naming an official state dinosaur! **(Delaware; paleontology)**
- She's back: Museum of Nature and Science sets reopening events **(museum)**
- Rafinesque, Ready for a Close-Up httpurl **(others)**
- Researchers have unlocked the secret to pearls' incredible symmetry **(invertebrates)**
- New Jersey is a strange beautiful place. httpurl **(others)**

......

Figure 1: Twitter user U with a self-introduction on the top, followed by the previous tweets (user history). U exhibits a mixture of personal interests in Delaware, invertebrates, paleontology, museum, and others.

1 Introduction

The irresistible popularity of social media results in an explosive number of users, creating and broadcasting massive amounts of content every day. Although it exhibits rich resources for users to build connections and share content, the sheer quantities of users might hinder one from finding those they want to follow (Matikainen, 2015). To enable users to quickly know each other, many social platforms encourage a user to write a *self-introduction*, a sentence to overview their personal interests.

A self-introduction is part of a self-described profile, which may else include locations, selfies, user tags, and so forth, and is crucial in online user interactions (McCay-Peet and Quan-Haase, 2016). Previous findings (Hutto et al., 2013) indicate users tend to follow those displaying self-introductions because a well-written self-introduction will brief others about a user's interests and facilitate them to initialize connections. It would benefit users in making like-minded friends and gaining popularity; whereas not all users are skillful in writing a good self-introduction. We are thus interested in how NLP may help and study **self-introduction generation**, a new application to learn user interests from their historical tweets (henceforth **user history**) and brief them in a self-introduction.

Despite substantial efforts made in profiling users, most existing work (Li et al., 2014; Farseev et al., 2015; Farnadi et al., 2018; Chen et al., 2019b) focuses on *extracting* keywords from user history and producing *tag-level user attributes* (e.g., interests, ages, and personality), which may later charac-

[*] Corresponding author

[1] Our code and dataset are released at `https://github.com/cpaaax/UTGED`.

Findings of the Association for Computational Linguistics: ACL 2023, pages 11387–11402
July 9-14, 2023 ©2023 Association for Computational Linguistics

terize personalization and recommendation (Wang et al., 2019a; Liang et al., 2022). However, tag-level attributes profile a user through a fragmented view, while human readers may find it difficult to read. On the contrary, we automate the writing of a *sentence-level* self-introduction via *language generation*, providing a more natural and easy-to-understand way to warm up social interactions. It consequently will enable a better socializing experience and user engagement in social media.

To practically train NLP models with capabilities in self-introduction writing, we collect a large-scale Twitter dataset with 170K public users. Each user presents a self-introduction (manually written by themselves) and previous tweets in their history, corresponding to a total of 10.2M tweets.

For methodology design, we take advantage of cutting-edge practices using pre-trained encoder-decoder for language understanding and generation. However, in real-world practice, users may post numerous tweets exhibiting *lengthy content, noisy writings, and diverse interests*; these may challenge existing encoder-decoder models in capturing salient personal interests and reflecting them in the brief self-introduction writing.

To illustrate this challenge, Figure 1 shows the self-introduction of a Twitter user U and some sampled tweets from U's user history. U exhibits a mixture of interests varying in *Delaware*, *invertebrates*, *paleontology*, *museum*, and others, scatteredly indicated in multiple noisy tweets. It presents a concrete challenge for models to digest the fragmented information, distill the introduction-worthy points, and condense them into a concise, coherent, and engaging self-introduction for further interactions. Moreover, existing NLP models are ineffective in encoding very long documents (Cao and Wang, 2022), whereas popular users may post numerous tweets, resulting in a lengthy history to encode.

Consequently, we propose a novel unified topic-guided encoder-decoder (UTGED) framework for self-introduction generation. First, a neural topic model (Srivastava and Sutton, 2017) clusters words by statistics to learn a mixture of latent topics in characterizing user interests underlying their lengthy history. Then, we inject the latent topics into a BART-based encoder and decoder (Lewis et al., 2020); the encoder employs topic distributions as continuous prompts (Lester et al., 2021; Liu et al., 2021; Li and Liang, 2021) to guide capturing personal interest mixture, and the decoder adopts topic words to control the writing for personalized self-introduction.

In experimental results, the comparison in both automatic and human evaluation show that UTGED outperforms state-of-the-art encoder-decoder models without topic guidance; and ablation studies indicate the individual contribution from topic-guided encoder and decoder. Then, we conduct parameter analyses on topic number and topic prompt length; they are followed by the study on model performance given users varying in historical tweet number, where UTGED consistently performs better. Finally, a case study and an error analysis interpret UTGED's superiority and limitations.

To the best of our knowledge, we present the first NLP study on self-introduction writing from user tweeting history, where we build the first dataset for its empirical studies and show the benefits from latent topics to the state-of-the-art encoder-decoder paradigm. Below are details of our contributions.

- We present a new application to capture personal interests from a user's tweeting history and generate their self-introductions accordingly.
- We approach the application with a novel UTGED (unified topic-guided encoder-decoder) framework, which explores latent topics to represent users' personal interests and to jointly guide user encoding and self-introduction decoding.
- We construct a large-scale Twitter dataset for self-introduction study and extensive experimental results on it show the superiority of UTGED practically and the benefits of latent topics on the task.

2 Related Work

Our work relates to user profiling (by task formulation) and topic modeling (by methodology).

User Profiling. This task aims to characterize user attributes to reflect a personal view. Most previous work focuses on modeling a user's tweeting history (Li et al., 2014) and social network interactions (Qian et al., 2019; Wang et al., 2019a; Chen et al., 2019b; Wang et al., 2021; Wei et al., 2022) to predict user attribute tags (e.g., ages and interests). However, most existing work focuses on classifying user profiles into fragmented and limited tags. Different from them, we study sentence-level self-introduction and explore how NLP handles such personalized generation, which initializes the potential to profile a user via self-introduction writing.

Datasets	**Data Source**	**Source-Target Pair Number**			**Token Number**	
		Train	Valid	Test	Src. len.	Tgt. len.
NYT (Sandhau, 2008)	News	44,382	5,523	6,495	1183.2	110.8
PubMed(Cohan et al., 2018)	Scientific Paper	83,233	4,946	5,025	444.0	209.5
Reddit (Kim et al., 2019)	Social Media	41,675	645	645	482.2	28.0
WikiHow (Koupaee and Wang, 2018)	Knowledge Base	168,126	6,000	6,000	580.8	62.6
Ours (*users' self-introductions*)	Social Media	140,956	17,619	17,624	1581.3	20.0

Table 1: Statistical comparison of our social media self-introduction dataset with other popular summarization datasets. Src. means source (input), Tgt. refers to target (output), and len. stands for average length (word number).

Topic Modeling. Topic models are popular unsupervised learning methods to explore corpus-level word co-occurrence statistics and represent latent topics via clustering topic-related words. Recent work mostly adopts neural architectures based on Variational Autoencoder (VAE) (Kingma and Welling, 2014), enabling easy joint work with other neural modules (Srivastava and Sutton, 2017).

Latent topics have shown beneficial to many NLP writing applications, such as the language generation for dialogue summaries (Zhang et al., 2021), dialogue responses (Zhao et al., 2017, 2018; Chan et al., 2021; Wang et al., 2022), poetries (Chen et al., 2019a; Yi et al., 2020), social media keyphrases (Wang et al., 2019b), quotations (Wang et al., 2020), and stories (Hu et al., 2022). Most existing methods focus on exploiting topics in decoding and injecting latent topic vectors (topic mixture) to assist generation. In contrast to the above work's scenarios, our application requires digesting much more lengthy and noisy inputs with scattered keypoints; thus, we leverage topics more finely and enable its joint guidance in encoding (by feeding in the topic mixture as topic prompts) and decoding (using topic words to control word-by-word generation).

Inspired by the success of pre-trained language models (PLMs), some efforts have been made to incorporate PLMs into VAE to conduct topic modeling (Li et al., 2020; Gupta et al., 2021; Meng et al., 2022). However, PLMs might be suboptimal in modeling user history (formed by numerous noisy tweets), because PLMs tend to be limited in encoding very long documents (Cao and Wang, 2022). Here, we model latent topics by word statistics, allowing better potential to encode long input.

3 Twitter Self-Introduction Dataset

To set up empirical studies for social media self-introduction, we build a large-scale Twitter dataset.

Data Collection. Following Nguyen et al. (2020), we first downloaded the general Twitter streams from September 2018 to September 2019. Then, we extracted the user ids therein and removed the duplicated ones. Next, we gathered users' tweeting history and self-introductions via Twitter API[2] and filtered out inactive users with less than 30 published tweets. For users with over 100 published tweets, only the latest 100 ones were kept. At last, we maintained the tweet text in English and removed irrelevant fields, e.g., images and videos.

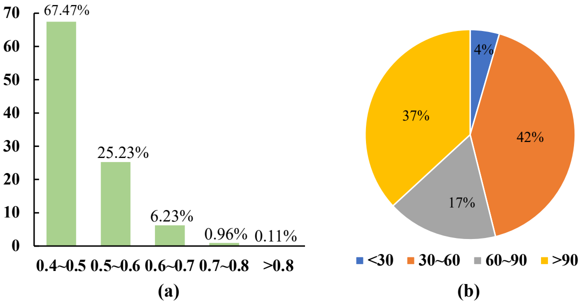

Figure 2: Analysis of the distribution over (a) average similarity of user history tweets (capped at top 30) to self-introduction and (b) tweet number in user history.

Data Pre-processing. First, we removed non-English self-introductions and those too short (<7 tokens) or too long (>30 tokens). Second, we employed SimCSE (Gao et al., 2021) (an advanced model for semantic matching) to measure the text similarity between a user's self-introduction and their tweeting history. Then, for training quality concern, we removed users with self-introductions that exhibit less than 0.4 similarity score [3] on average to the top-30 tweets in history.[4] Third, for the remaining 176,199 unique user samples, each corresponds to a pair of user history (source) and self-introduction (target). For model evaluation, we randomly split the user samples into training (80%), validation (10%), and test (10%) sets.

[2] `https://developer.twitter.com`

[3] The details of the tweets with less than 0.4 similarity score are shown in Appendix A.2.

[4] Users may have less than 30 tweets kept in history (e.g., some tweets without English text were excluded). For these users, we considered all their maintained tweets.

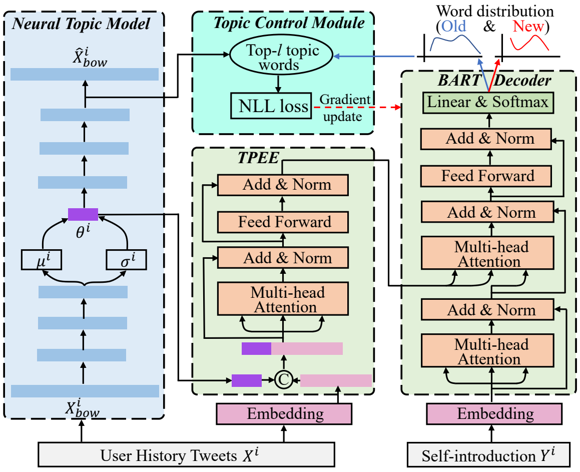

Figure 3: The overview of our UTGED (Unified Topic-Guided Encoder-Decoder) framework, The left module shows a neural topic model (NTM) representing user interests with latent topics. The topic mixtures help the encoder explore user history (middle) and topic words guide the decoder in self-introduction generation (right).

Data Analysis. Encoder-decoder models are widely used in summarization tasks (§2). We then discuss the difference of our task through an empirical lens. The statistics of our dataset and other popular summarization datasets are compared in Table 1. We observe that each of our data sample exhibits a longer source text and a shorter target text compared to other datasets. It indicates the challenge of our self-introduction task, where directly using summarization models may be ineffective.

To further analyze the challenges, Figure 2(a) displays the distribution of SimCSE-measured source-target similarity (averaged over top-30 tweets in user history). It implies that very few tweets are semantically similar to their authors' self-introductions, making it insufficient to simply "copy" from history tweets. We then analyze and show the tweet number distribution in user history in Figure 2(b). It is noticed that 37% users posted over 90 history tweets, scattering interest points in numerous tweets and hindering models in capturing the essential ones to write a self-introduction.

4 Our UTGED Framework

Here we describe our UTGED (unified topic-guided encoder-decoder) framework. Its overview is in Figure 3: latent topics guide the PLMs to encode user history and decode self-introductions.

The data is formulated as source-target pairs $\{X^i, Y^i\}_{i=1}^N$, where $X^i = \{x_1^i, x_2^i, ..., x_m^i\}$ indicates user history with m tweets published by user u^i, Y^i represents the user-written description, and N is the number of pairs. In our task, for user u^i, models are fed in their user history tweets X^i and trained to generate their self-introduction Y^i.

4.1 Neural Topic Model

To explore users' interests hidden in their numerous and noisy tweets, we employ a neural topic model (NTM) (Srivastava and Sutton, 2017) to learn latent topics (word clusters). NTM is based on VAE with an encoder and a decoder to reconstruct the input.

For word statistic modeling, the history tweets in X^i are first processed to a one-hot vector $X^i_{bow} \in \mathbb{R}^{V_{bow}}$ in Bag-of-words (BoW), where V_{bow} indicates NTM's vocabulary size. Then, similar to VAE, NTM encoder transforms BoW vector X^i_{bow} into a K-dimensional latent topic variable $z^i \in \mathbb{R}^K$. Conditioned on z^i, NTM decoder produces $\hat{X}^i_{bow}$ to reconstruct X^i_{bow}. Here presents more details.

NTM Encoder. Given the BoW vector X^i_{bow}, NTM encoder attempts to learn the mean μ^i and standard deviation σ^i based on the assumption that words in X^i exhibit a Gaussian prior distribution. Its mean and standard deviation, μ^i and σ^i will be encoded by the following formula and later be utilized to compute the latent topic vector z^i:

$$\mu^i = f_\mu(f_b(X^i_{bow})); \log\sigma^i = f_\sigma(f_b(X^i_{bow})) \qquad (1)$$

where $f_*(\cdot)$ indicates a single layer perceptron performing the linear transformation of input vectors.

NTM Decoder. We then reconstruct the BoW in X^i based on the NTM-encoded μ^i and σ^i. We hypothesize that a corpus may exist K latent topics, each reflecting a certain user interest and represented by word distribution over the vocabulary V_{bow}. Besides, user history X^i is represented as a topic mixture θ^i to reflect u^i's interest combination over K topics. The procedure is as follows:

- Draw latent topic vector $z^i \sim \mathcal{N}(\mu^i, \sigma^i)$
- Topic mixture $\theta^i = \text{softmax}(f_\theta(z^i))$
- For each word $w \in X^i$:

  Draw $w \sim \text{softmax}(f_\phi(\theta^i))$

where f_θ and f_ϕ are a single layer perceptron. The weight matrix of f_ϕ indicates topic-word distributions (ϕ_1,ϕ_2,...,ϕ_K).

The learned latent topics for X^i will later guide the BART-based self-introduction generation (to be discussed in §4.2). The topic mixture θ^i will be injected into the BART encoder for capturing salient interests and the top-l words $A^i = \{a_1^i, a_2^i, ..., a_l^i\}$ with highest topic-word probability in ϕ_c (c indexes the major topic suggested by θ^i) will go for controlling the writing process of the BART decoder.

4.2 Topic-Guided Generation Model

We have discussed how to model a user u^i's latent interests with NTM and the learned latent topics (θ^i and topic words A^i) will then guide a BART-based encoder-decoder model to generate u^i's self-introduction, Y^i. In the following, we first present how we select tweets (for fitting overly long user history into a transformer encoder), followed by our topic-guided design for encoding and decoding.

Tweet Selection. Recall from §3 that user history tends to be very long (Table 1 shows it has 1581.3 tokens on average). However, BART encoder limits its input length. To fit in the input, we go through the following steps to shortlist representative tweets from a user u^i's lengthy tweeting history, X^i.

First, we measure how well a tweet x^i_u can represent X^i via averaging its similarity to all others:

$$s^i_u = \frac{1}{|m|} \sum_{x^i_v \in X^i} Sim\left(x^i_u, x^i_v\right) \tag{2}$$

where $Sim(x^i_u, x^i_v)$ represents the $x^i_u - x^i_v$ SimCSE-measured cosine similarity. Then, we maintain a shortlist R_i to hold X^i's representative tweets, which is empty at the beginning and iteratively added with x^i_h obtaining the highest similarity score (Eq. 2). To mitigate redundancy in R^i, once x^i_h is put in R_i, it is removed from X^i, and so are other tweets in X^i whose cosine similarity to x^i_h is over a threshold λ (i.e., 0.8). For easy reading, we summarize the above steps in Algorithm 1.

After that, we further rank the shortlisted tweets in R^i based on their overall similarity in X^i (Eq. 2). The top ones are maintained and concatenated chronologically to form a word sequence $R^i = \{w^i_1, w^i_2, ..., w^i_M\}$ (M denotes the word number).

Topic Prompt Enhanced Encoder (TPEE). We then discuss how we encode R^i (selected user history tweets) in guidance of the topic mixture θ^i (featuring latent user interests). The encoding adopts the BART encoder and is trained with θ^i-based prompt fine-tuning (thereby named as TPEE, short for topic prompt enhanced encoder).

We first obtain the topic prompt as follows:

$$B^i = \text{MLP}(\theta^i) \tag{3}$$

where MLP is a feedforward neural network. Following Li and Liang (2021), $B^i \in \mathbb{R}^{d \times L}$ is split into L vectors $[b^i_1, b^i_2, ..., b^i_L]$. L indicates the topic prompt length and each vector $b^i_j \in \mathbb{R}^d$.

To inject the guidance from topic prompts $\{b^i_1, b^i_2, ..., b^i_L\}$ (carrying latent topic features), we put them side by side with the embeddings of words $\{w^i_1, w^i_2, ..., w^i_M\}$ (reflecting word semantics of R^i). Then, a BART encoder $\mathcal{E}$ represents user u^i's salient interests H^i_E in its last layer:

$$H^i_E = \mathcal{E}(\left[b^i_1; b^i_2; ...; b^i_L; e^i_1; e^i_2; ...; e^i_M\right]) \tag{4}$$

where $e^i_j \in \mathbb{R}^d$ is the BART-encoded word embedding of w^i_j and [;] is the concatenation operation.

Topic Words Enhanced Decoder (TWED). Recall in §4.1, NTM generates l topic words (A^i) to depict a user u^i's major latent interests. To further reflect such interests in the produced self-introduction, we employ A^i to control a BART decoder $\mathcal{D}$ in its word-by-word generation process through the topic control module.

For easy understanding, we first describe how the original BART decode. At the t-th step, the decoder $\mathcal{D}$ is fed in its previous hidden states $H^i_{D,t}$, the BART encoder's hidden states H^i_E (Eq. 4), and latest generated word Y^i_t, resulting in hidden step o^i_{t+1}. Based on that, the next word is generated following the token distribution p^i_{t+1}. The concrete workflow is shown in the formula as follows:

$$p^i_{t+1} = \text{softmax}(W_e o^i_{t+1}) \tag{5}$$

$$o^i_{t+1}, H^i_{D,t+1} = \mathcal{D}(H^i_E, H^i_{D,t}, Y^i_t) \tag{6}$$

where $H^i_{D,t+1}$ stores all the previous decoder hidden states till step $t+1$, W_e is learnable and to map the latent logit vector o^i_{t+1} to the target vocabulary.

Then, we engage topic words A^i to control the above procedure by the topic control module. Inspired by BoW attribute model (Dathathri et al., 2020), we calculate the following log-likelihood loss to weigh the word generation probability p^i_{t+1} over each topic word $a^i_j \in A^i$:

$$\log p(A^i|Y^i_{t+1}) = \log\left(\sum_j p^i_{t+1}[a^i_j]\right) \tag{7}$$

The gradient from $\log p(A^i|Y^i_{t+1})$ is further involved in updating all decoder layers ($H^i_{D,t}$) in $\mathcal{D}$:

$$\widetilde{H}^i_{D,t} = \Delta H^i_{D,t} + H^i_{D,t} \tag{8}$$

$$\Delta H^i_{D,t} \leftarrow \Delta H^i_{D,t} + \alpha \frac{\nabla_{\Delta H^i_{D,t}} \log p(A^i|H^i_{D,t} + \Delta H^i_{D,t})}{\|\nabla_{\Delta H^i_{D,t}} \log p(A^i|H^i_{D,t} + \Delta H^i_{D,t})\|^\gamma} \tag{9}$$

where $\widetilde{H}^i_{D,t}$ indicates the updated (topic-controlled) decoder's states, $\Delta H^i_{D,t}$ means the gradient update to $H^i_{D,t}$, α is the step size, and γ is the normalization value. Furthermore, we adopt the same topic-controlling strategy to update the encoder's final layer states H^i_E and derive the updated states $\widetilde{H}^i_E$ based on Eq. 8 and 9. With Eq. 5 and 6, we can accordingly obtain the final token distribution $\widetilde{p}^i_{t+1}$ based on the topic-controlled encoder and decoder states $\widetilde{H}^i_E$, $\widetilde{H}^i_{D,t}$, and previous predicted word Y^i_t.

4.3 Joint Training in a Unified Framework

To couple the effects of NTM (described in §4.1) and topic-guided encoder-decoder module for self-introduction generation (henceforth SIG discussed in §4.2), we explore the two modules in a unified framework and jointly train them for better collaborations. The loss function of the unified framework is hence a weighted sum of NTM and SIG:

$$\mathcal{L} = \alpha\mathcal{L}_{NTM} + (1 - \alpha)\mathcal{L}_{SIG} \quad (10)$$

where $\mathcal{L}_{NTM}$ and $\mathcal{L}_{SIG}$ are the loss functions of NTM and SIG. α is the hyper-parameter trading off their effects and is set to 0.01 in our experiments.

For NTM, the learning objective is computed as:

$$\mathcal{L}_{NTM} = D_{KL}(\delta(z)||\rho(z|X)) - \mathbb{E}_{\rho(z|X)}[\delta(X|z)] \quad (11)$$

where $D_{KL}(\cdot)$ indicates the Kullback-Leibler divergence loss and $\mathbb{E}[\cdot]$ is reconstruction loss.[5] For the SIG, it is trained with the cross-entropy loss:

$$\mathcal{L}_{SIG} = -\sum_i \sum_t \log p^i_t \quad (12)$$

In practice, we first train the unified framework with Eq.10 and exclude A^i (topic words output of NTM). Then, during inference, we fix UTGED, employ A^i to control the decoding process and generate the final self-introduction with Eq.7~Eq.9.

5 Experiments and Discussions

5.1 Experimental Setup

Model Settings. We implemented NTM (§4.1) based on (Srivastava and Sutton, 2017) and set its topic number K to 100. Its BoW vocabulary size V_{bow} is set to 10K and hidden size to 200. The input of NTM is the BoW of original user history X^i while the input of SIG is capped at 1,024 tokens based on the shortlisted tweets in R^i (§4.2).[6]

The SIG model is based on the BART and built on 6 encoding layers and 6 decoding layers. We adopted AdamW and SGD to optimize the SIG and NTM, respectively. The learning rate is set to 5×10^{-5} for SIG and 1×10^{-4} for NTM. The topic prompt length L is set to 7. To warm up joint training (Eq.10), we pre-train NTM with Eq.11 for 100 epochs. During joint training, batch size is set to 8 and the maximum epoch to 5. In topic-controlled decoding, α is set to 0.25 and γ to 1.5 (Eq.9). Topic word number l is set to 30. Models are trained on a 24GB NVIDIA RTX3090 GPU.

Evaluation Metrics. For automatic evaluation, we adopt ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L), which are popular metrics in language generation based on output-reference word overlap and originally for summarization tasks (Lin, 2004). We also conduct a human evaluation on a 5 point Likert scale and over three criteria: *fluency* of the generated language, *consistency* of a self-introduction to the user's history, and *informativeness* of it to reflect essential user interests.

Baselines and Comparisons. We adopt extractive and abstractive summarization models in comparison. The former extracts user history tweets as the self-introduction by ranking them with: (1) **BERTExt** (Liu and Lapata, 2019) (based on BERT (Devlin et al., 2019)) (2) **TextRank** (Mihalcea and Tarau, 2004) (unsupervised graph ranking based on similarity) (3) **Consen** (unsupervised ranking with the averaged similarity to others (Eq.2)).

For abstractive competitors, models all follow the encoder-decoder paradigm. We employ **T5** (Raffel et al., 2020), **BART** (Lewis et al., 2020), and **PEGASUSU-X** (Phang et al., 2022), all based on PLMs and are state-of-the-art abstractive summarizers. We also compare to **GSum** (Dou et al., 2021), which employs highlighted sentences, keywords, relations, and retrieved summaries.

In addition, we examine the upper-bound tweet selection (shortlist given reference self-introduction). Here SimCSE first measures the similarity between the reference and each tweet in user history X^i. **Oracle$_\mathbf{E}$** then extracts the tweet with the highest similarity score. For **Oracle$_\mathbf{A}$**, we rank tweets based on the similarity score and the top ones are fed into BART for a generation. Furthermore, to explore the potential of our topic-guided design over **Oracle$_\mathbf{A}$** model, we feed **Oracle$_\mathbf{A}$**'s input to our UTGED and name it **Oracle$_\mathbf{A}$+Topic**.

5.2 Main Comparison Results

Table 2 shows the main comparison results. We first observe the inferior results from all extractive models, including Oracle$_E$. It is because of the non-trivial content gap between users' history tweets and their self-introductions (also indicated in Figure 2). Directly extracting tweets from user

[5] We refer readers to more details of NTM in Srivastava and Sutton (2017), which are beyond the scope of this paper.

[6] We also test NTM with BoW on $\{R^i\}_{i=1}^N$ and observe slightly worse results. It is possibly because NTM is based on word statistics and would barely be affected by lengthy input.

Method	R-1	R-2	R-L
Extractive			
BERTExt	11.67	1.92	10.04
TextRank	13.60	2.93	11.66
Consen	14.86	2.90	12.89
Abstractive			
T5	23.93	7.31	20.93
PEGASUS-X	24.10	7.44	21.07
GSum	22.19	5.99	19.27
BART	23.92	7.46	20.91
UTGED (Ours)	**24.99***	**8.05***	**21.84***
Oracle$_E$	20.89	5.94	18.04
Oracle$_A$	28.97	10.23	25.29
Oracle$_A$+Topic	29.36	10.39	25.62

Table 2: Main comparison results. UTGED achieves the best results (highlighted) and the performance gain is significant to all comparison models (indicated by * and measured by paired t-test with p-value<0.05).

Method	R-1	R-2	R-L
BART	23.92	7.46	20.91
BART+S	24.26	7.68	21.17
BART+S+E	24.78	7.95	21.65
BART+S+E+D (UTGED)	24.99	8.05	21.84

Table 3: Ablation study results. S: tweet selection (to shortlist tweets from user history); E: w/ TPEE (topic-guided encoder); D: w/TWED (topic-guided decoder).

history is thus infeasible to depict self-introduction, presenting the need to involve language generation. For this reason, abstractive methods exhibit much better performance than extractive baselines.

Among comparisons in abstractive models, UTGED yields the best ROUGE scores and significantly outperforms the previous state-of-the-art summarization models. It shows the effectiveness in engaging guidance of latent topics from lengthy and noisy user history, which may usefully signal the salient interests for writing a self-introduction.

In addition, by comparing Oracle$_\mathrm{A}$ results and model results, we observe a large margin in between. It suggests the challenge and importance of tweet selection for user history encoding, providing insight into future related work. Moreover, interestingly, Oracle$_\mathrm{A}$+Topic further outperforms Oracle$_\mathrm{A}$, implying topic-guided design would likewise benefit the upper-bound tweet selection scenarios.

Ablation Study. Here we probe into how UTGED's different modules work and show the ablation study results in Table 3. All modules (tweet selection (S), TPEE (E), and TWED (D)) contrite positively because they are all designed to guide models in focusing on essential content reflecting user interests against lengthy input. TPEE may show larger individual gain than the other two, possibly because the topic mixtures directly reflect user interests and are easier for the model to leverage.

Method	Fluency	Informativeness	Consistency
GSum	3.43	2.65	2.28
BART	3.85	3.21	2.89
UTGED	3.66	3.68	3.27

Table 4: Human evaluation results. Cohen's Kappa for all annotator pairs is 0.63 on average (good agreement).

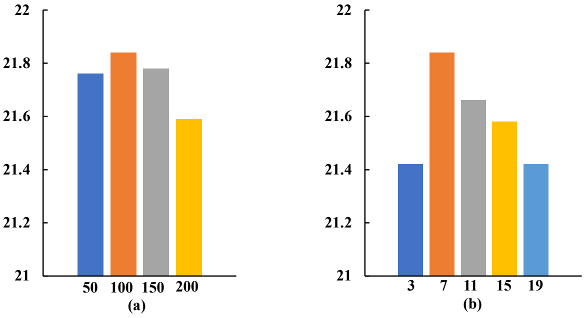

Figure 4: Parameter analysis results. The X-axis shows topic number (a) and prompt length (b); Y-axix shows the R-L score measured on our UTGED's output.

Human Evaluation. To further test how useful our output is to human readers, we randomly select 100 samples from test set and train 3 in-house annotators from NLP background to rate the generated self-introductions. As shown in Table 4, UTGED is superior in informativeness and consistency. It implies latent topics can usefully help capture salient interests from lengthy and noisy user history. However, its fluency is lower than that of BART, indicating that topic words slightly perturb the pre-trained decoder (Dathathri et al., 2020).

5.3 Quantitative Analysis

To better study UTGED, we then quantify the topic number, prompt length, and input tweet number to examine how they affect performance. Here only R-L is shown for better display, and similar trends were observed from R-1 and R-2. For the full results, we refer readers to Appendix A.3.

Varying Topic Number. The first parameter analysis concerns the topic number K (NTM's hyperparameter). As shown in Figure 4(a), the score first increases then decreases with larger K and peaks the results at $K = 100$. We also observe $K = 200$ results in much worse performance than other Ks, probably because modeling too fine-grained topics is likely to overfit NTM in user interest modeling, further hindering self-introduction generation.

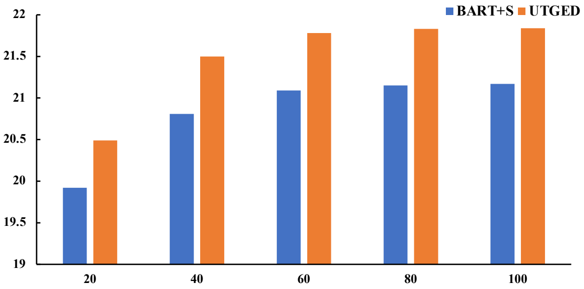

Figure 5: The comparisons between BART+S and UTGED while varing sentence number. X-axis: the values of sentence number; Y-axis: the R-L score.

Varying Prompt length. Likewise, we analyze the effects of prompt length L in Figure 4(b). The best score is observed given L=7, much better than very short or very long prompt length. Longer prompts may allow stronger hints from NTM, helpful to some extent; however, if the hint becomes too strong (given too-long prompt), topic features may overwhelm the encoder in learning specific features for self-introduction writing.

Users w/ Varying Tweet Number. Recall in Figure 2(b), users largely vary tweet number in history (attributed to different active degrees). We then examine how models work given varying tweet numbers in history. BART+S and UTGED are tested, both with tweet selection (S) to allow very long input and Figure 5 shows the results. Both models exhibit growing trends for more active users, benefiting from richer content in their history to infer self-introduction. Comparing the two models, UTGED performs consistently better, showing the gain from NTM's is robust over varying users.

5.4 Qualitative Analysis

Case Study. Figure 6 shows a user sample interested in "teaching" and "reading". It can be indicated by topic words like "student", "book", and "school" produced by NTM. From BART's output, we find its errors in "seesaw specialist" further mislead the model in writing more irrelevant content (e.g., "google certified educator" and "google trainer"). It may be caused by the common exposure bias problem in language generation (Ranzato et al., 2016; Zhang et al., 2019). On the contrary, UTGED's output is on-topic till the end, showing topic guidance may mitigate off-topic writing. [7]

Error Analysis. In the main comparison (Table 2), UTGED performs the best in yet also has a non-trivial gap to Orcation$_A$. Here we probe its limitations and discuss the two major error types in Figure 7. First, the output may contain grammatical mistakes, e.g., "deals, deals", limited by BART'S decoder capability and topic words' effects. It calls for involving grammar checking in decoding. The second error type is propagated from wrong latent topics. As shown in the error case (second row), the user is a provider of "pet"-style clothes, whereas NTM may cluster it with other "pet lover"-users and further mislead the writing process. Future work may explore a better topic modeling method to mitigate the effects of mistakenly clustering.

Source: "someone is proud of her artwork now on display in our library!", "we were excited to hear from to learn more about summer reading!", "second graders are becoming familiar with the intricacies of tinytap on our ipads as we prepare for an assured learning experience on folktales", "our makerspace is on the move!" · · · · · ·
Topic words: life, love, learning, school, writing, book, read, yoga, kids, students, education, quotes, community, children, time · · · · · ·
BART: webersen elementary media specialist, seesaw specialist, google certified educator, google trainer, apple certified educator.
UTGED: elementary library media specialist at webster hill elementary school. i love to connect with my students and help them grow as independent learners.
Target: i proudly teach all pk-5 webster hill students. we learn to think critically, research efficiently, meaningfully integrate technology, and find joy in reading.

Figure 6: A Twitter user sample and the related results. From top to down shows user history (source T^i), major topic words (A^i), BART output, UTGED output, and reference self-introduction (target Y^i). We inspect the topic words helpful for our task and color them in red.

Grammar Error	***G:*** travel with mei is a travel blog with travel tips, deals, deals and more. ***T:*** travel in holiday is a blog that aims to inspire more people that there are more life and adventure to discover in this world.
Topic Error	***G:*** we are a group of pet lovers who love dogs and cats and want to share them with you! ***T:*** we put your pets on your pants! available for adults and kids makes perfect birthday and holiday gifts leggings and tops

Figure 7: Examples of major error types for the generation results of UTGED (G) and the target reference (T).

6 Conclusion

We have presented a new application to generate personalized self-introduction, where a large-scale Twitter dataset is gathered for experiments. A novel unified topic-guided encoder-decoder framework is proposed to leverage latent topics for distilling essential user interests from the numerous noisy tweets a user has posted. Empirical results show our model outperforms advanced PLM-based models, shedding light on the potential of latent topics in helping PLMs digest lengthy and noisy input.

[7] More topic word cases could be found in Appendix A.4 and longer source tweets are shown in Appendix A.5.

Acknowledgements

This paper is substantially supported by the NSFC Young Scientists Fund (No.62006203, 62106105), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/25200821), and the Innovation and Technology Fund (Project No. PRP/047/22FX).

Limitations

First, inference efficiency is one of the main limitations of this work. The BART model takes about 14 minutes to complete the inference on our dataset, while our UTGED needs 92 minutes. The reason for the slow inference is that UTGED requires heavy computation to update the gradient to the encoder's states and decoder's states (as shown in Eq.7~Eq.9). Future work may consider how to advance model efficiency further.

Second, the lack of multimodal content in the published tweets would result in another limitation. The images contained in the published tweets are ignored in this work. However, due to the complicated relationships between images and texts in a multimodal tweet, images might provide complementary content and complete the meanings of the message (Vempala and Preotiuc-Pietro, 2019). Therefore, future studies might explore self-introduction generation using multimodal tweets (images and text) to indicate personal interests.

Ethics Statement

Our paper constructs a large-scale Twitter dataset for a self-introduction generation. The data acquisition procedure follows the standard data collection process regularized by Twitter API. Only the public users and tweets are gathered. The downloaded data is only used for academic research. For our experiments, the data has been anonymized for user privacy protection, e.g., authors' names are removed, @mention and URL links are changed to common tags. Following Twitter's policy for content redistribution, we will only release the anonymized data. Additionally, we will require data requestors to sign a declaration form before obtaining the data, ensuring that the dataset will only be reused for the purpose of research, complying with Twitter's data policy, and not for gathering anything that probably raises ethical issues, such as sensitive personal information. For the human annotations, we recruited the annotators as part-time research assistants with 15 USD/hour payment.

References

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Zhangming Chan, Lemao Liu, Juntao Li, Haisong Zhang, Dongyan Zhao, Shuming Shi, and Rui Yan. 2021. Enhancing the open-domain dialogue evaluation in latent space. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4889–4900. Association for Computational Linguistics.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019a. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4925–4931. ijcai.org.

Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. 2019b. Semi-supervised user profiling with heterogeneous graph attention networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2116–2122. ijcai.org.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4830–4842. Association for Computational Linguistics.

Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 171–179. ACM.

Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting multiple sources for user profile learning: a big data study. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*, pages 235–242. ACM.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Pankaj Gupta, Yatin Chaudhary, and Hinrich Schütze. 2021. Multi-source neural topic modeling in multi-view embedding spaces. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4205–4217. Association for Computational Linguistics.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. 2022. Fuse it more deeply! A variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 697–716. Association for Computational Linguistics.

Clayton J. Hutto, Sarita Yardi, and Eric Gilbert. 2013. A longitudinal study of follow predictors on twitter. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*, pages 821–830. ACM.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks. 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 2519–2531. Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4678–4699. Association for Computational Linguistics.

Jiwei Li, Alan Ritter, and Eduard H. Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 165–174. The Association for Computer Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Shangsong Liang, Yupeng Luo, and Zaiqiao Meng. 2022. Profiling users for question answering communities via flow-based constrained co-embedding model. *ACM Trans. Inf. Syst.*, 40(2):34:1–34:38.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics.

Janne Tapani Matikainen. 2015. Motivations for content generation in social media. *Participations: Journal of Audience and Reception Studies*.

Lori McCay-Peet and Anabel Quan-Haase. 2016. A model of social media engagement: User profiles, gratifications, and experiences. In Heather O'Brien and Paul A. Cairns, editors, *Why Engagement Matters: Cross-Disciplinary Perspectives of User Engagement in Digital Media*, pages 199–217. Springer.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3143–3152. ACM.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *CoRR*, abs/2208.04347.

Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. Graphie: A graph-based framework for information extraction. 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 751–761. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Evan Sandhau. 2008. The new york times annotated corpus. In *Linguistic Data Consortium*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. 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 2830–2840. Association for Computational Linguistics.

Dongjie Wang, Pengyang Wang, Kunpeng Liu, Yuanchun Zhou, Charles E. Hughes, and Yanjie Fu. 2021. Reinforced imitative graph representation learning for mobile user profiling: An adversarial training perspective. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4410–4417. AAAI Press.

Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022. CARE: causality reasoning for empathetic responses by conditional graph generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 729–741. Association for Computational Linguistics.

Lingzhi Wang, Jing Li, Xingshan Zeng, Haisong Zhang, and Kam-Fai Wong. 2020. Continuity of topic, interaction, and query: Learning to quote in online conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6640–6650. Association for Computational Linguistics.

Pengyang Wang, Yanjie Fu, Hui Xiong, and Xiaolin Li. 2019a. Adversarial substructured representation learning for mobile user profiling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 130–138. ACM.

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

Wei Wei, Chao Huang, Lianghao Xia, Yong Xu, Jiashu Zhao, and Dawei Yin. 2022. Contrastive meta learning with behavior multiplicity for recommendation. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1120–1128. ACM.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. 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 9450–9457. AAAI Press.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. 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 4334–4343. Association for Computational Linguistics.

Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2021. Unsupervised abstractive dialogue summarization for tete-a-tetes. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14489–14497. AAAI Press.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *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 1098–1107. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. 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 654–664. Association for Computational Linguistics.

A Appendix

A.1 Tweet Selection Algorithm

Algorithm 1 Selecting representative tweets

```
Require: collected tweets pool for user $u^i$: $X^i = \{x_1^i, x_2^i, ..., x_m^i\}$
Ensure: representative tweets shortlist $R^i$
initial $R^i = \{\}$
repeat
    calculate overall similarity score between
            a tweet and other tweets in $X^i$;
    assume the tweet with the highest score is
            $x_h^i$, remove $x_h^i$ from $X^i$ to $R^i$;
    calculate the similarity score between $x_h^i$
            and remained tweets in $X^i$;
    for the tweets whose similarity score is
            higher than $\lambda$, remove it from $X^i$;
until there are no tweets in $X^i$
```

A.2 Data filtering

Similarity Score	R-1	R-2	R-L
$[0, 0.1)$	8.43	0.91	7.67
$[0.1, 0.2)$	11.49	1.94	10.50
$[0.2, 0.3)$	17.87	4.50	15.03
$[0.3, 0.4)$	24.25	7.52	21.35

Table 5: The results of Oracle$_A$+Topic on low-similarity data samples.

Here we show the original dataset's distribution before filtering: [-1, 0): 9,680; [0, 0.1): 123,560; [0.1, 0.2): 257,759; [0.2, 0.3): 193,847; [0.3, 0.4): 157,478; [0.4, 0.5): 118,881; [0.5, 0.6): 44,455; [0.6, 0.7): 10,977; [0.7, 0.8): 1,691; [0.8, 0.9): 169; [0.9, 1.0]: 26. We observe that the number of user samples first increases and then decreases, indicating that self-introductions are related to the user's historical tweets. Otherwise, the data distribution will tend to exhibit a long tail (based on social media characteristics).

Additionally, we tested a sample of 10,000 users with similarity scores fall in the ranges of [0.3,0.4), [0.2,0.3), [0.1, 0.2), [0,0.1) and results of the best model Oracle$_A$+Topic on low-similarity data samples are shown in Table 5. The results indicate that low-similarity data samples do impact negatively on the training results.

A.3 Full Experimental Results

Varying Topic Number. We show the results from BART+S+E on the left of "/" and those from UTGED on the right.

K	R-1	R-2	R-L
50	24.58/24.85	7.85/7.99	21.49/21.76
100	24.78/24.99	7.95/8.05	21.65/21.84
150	24.77/24.98	7.93/8.01	21.60/21.78
200	24.67/24.71	7.86/7.90	21.52/21.59

Table 6: The effects of topic number K.

Varying Prompt length. We show the results from BART+S+E on the left of "/" and those from UTGED on the right.

L	R-1	R-2	R-L
3	24.36/24.53	7.68/7.78	21.25/21.42
7	24.78/24.99	7.95/8.05	21.65/21.84
11	24.48/24.69	7.90/8.01	21.43/21.66
15	24.53/24.74	7.84/7.89	21.41/21.58
19	24.34/24.48	7.76/7.81	21.30/21.42

Table 7: The effects of prompt length L.

Varying Sentence Number. We show the results from BART+S on the left of "/" and those from UTGED on the right.

SN	R-1	R-2	R-L
20	22.79/23.41	6.90/7.15	19.92/20.49
40	23.82/24.57	7.45/7.83	20.81/21.50
60	24.12/24.90	7.63/8.02	21.09/21.78
80	24.23/24.95	7.70/8.06	21.15/21.82
100	24.26/24.99	7.68/8.05	21.17/21.84

Table 8: The effects of sentence number (SN).

A.4 Topic Words

Topic idx	Topic words
3	training, golf, health, fitness, yoga, back, today, day, life, healthy, sports, club, monday, time, dealer, great, body, week, gym, fit, workout, run, team, motivation, free, fun, weight, stay, weekend, start
4	live, video, game, check, games, share, ps, play, twitch, gaming, stream, small, pokemon, playstation, gta, broadcast, playing, streaming, nintendo, youtube, switch, pc, xbox, indie, added, streamer, fortnite, gamer, retro, minecraft
11	visit, info, html, training, high, machine, quality, contact, products, uk, india, product, power, glass, air, water, printing, solutions, metal, system, industry, range, manufacturer, steel, equipment, custom, safety, project, services, construction
14	art, artist, comic, tattoo, anime, indie, game, comics, painting, dev, drawing, illustration, cosplay, writing, star, digital, canvas, horror, furry, fantasy, fan, sketch, artwork, inktober, artists, found, den, fanart, poetry, original
27	free, online, win, today, pm, sale, lottery, play, join, code, golf, tickets, offer, betting, click, tips, app, lyft, open, club, link, casino, buy, promo, money, apply, store, deposit, bonus, sports
35	news, india, energy, uk, gold, oil, market, jet, global, industry, international, air, pakistan, africa, solar, forex, aviation, charter, china, trade, cruise, indian, dubai, cargo, power, crypto, mining, world, report, south
54	travel, visit, beach, luxury, book, hotel, world, experience, stay, holiday, tour, bengal, island, beautiful, enjoy, jet, charter, adventure, cruise, yacht, pool, summer, hotels, explore, park, vacation, disney, city, maldives, discover
81	business, shop, local, small, find, online, today, support, day, sale, buy, happy, friday, service, city, biz, make, photo, great, monday, gift, store, details, black, deals, car, weekend, check, services, ca
86	today, pm, team, game, school, day, week, tonight, great, tomorrow, high, night, girls, students, season, state, year, congratulations, support, win, college, play, friday, back, st, boys, student, good, pride, senior
94	music, show, house, rock, radio, dance, album, playing, guitar, song, listen, artist, tickets, metal, night, band, live, festival, tonight, country, friday, hop, dj, single, hip, pop, jazz, party, mix, reggae

Figure 8: Randomly sampled 10 topics with their top-30 topic words.

A.5 Detailed Case Study

Source: "someone is proud of her artwork now on display in our library!", "fifth graders can't wait to read this summer! thanks for reaching out to our kids virtually!", "we were excited to hear from to learn more about summer reading!", "im grateful i spent today in a school with students and teachers talking about story, compassion, and our hearts. thank you!", "it was an incredible day at webster hill with! thank you for sharing your energy, enthusiasm, and love of reading with our students!", "second graders are becoming familiar with the intricacies of tinytap on our ipads as we prepare for an assured learning experience on folktales!", "our makerspace is on the move!", "second graders are taking brief notes using information from pebblego and creating an expert ebook with the book creator app!", "kindergarten friends are browsing for informational texts and previewing the pictures to help them determine the main topic or what it is mostly about. we're practicing some seesaw skills to share our learning, too!", "third graders are becoming independent s of our library! here, they're noticing patters with call numbers to collaboratively organize e books. we want them to be able search for and locate books on any topic or area of interest! well on our way.", "computer science truly connects to all content areas. here, a student is modifying musical notes and tempo to get a keyboard to play a popular song!", "we had an exciting morning at webster hill! it was such a pleasure to welcome and other special guests to a fourth grade library media class on coding.", "more ozobot fun!", "getting to know dot and dash!", "programming ozobot to read color patters!", "pre k has been practicing following specific directions like a robot! we had lots of fun with a red light, green light song!", "after browsing for books, pre k friends engage in some fun centers that encourage cooperation. we're even starting to recognize some letters!", "mrs. bender and i have been spending lots of time making our library extra special for our amazing students! we are so excited to see everyone !", "fifth graders are starting to meet their middle school library media specialist!", "coding with cubetto!", "some research inspired by a true story!", "i was so excited to participate in a virtual author visit with our very own poet lms, jill dailey. amazing.", "this year, i 'm getting to spend some time in classrooms working with students in small groups to apply their knowledge of informational texts. so much fun!", "primary students enjoyed reading neither this week with a message of acceptance. we used our love of the character to then spark some creativity and research! we designed new creatures from two animals using seesaw and then began exploring pebblego for facts.", "browsing for good books!", "supporting our budding early emergent readers with a repetitive text, familiar song, and some fun connections with drawing tools in seesaw!", "first graders can identify common text features and how they help readers!", "fifth graders presented their website evaluations, citing evidence from the text and indicators of a reliable source to explain whether or not to use a site for research!", "in kindergarten, we are making connections to our own lives with the characters and settings in stories!", "second graders are identifying information that is safe to share online and showing us what they know with a seesaw activity!", "first graders are using strategies to recount the most important details in literature. here, we illustrated some of what we thought the author could n 't leave out! we even got to practice with our digital learning platform, seesaw.", "library media lessons take place in the classroom this year!", "we're back! our kindergarten friends learned about seesaw this week and began using drawing, photo, and audio recording tools to complete activities. we are digital learners!", "the men and women's soccer teams shared their love of reading with webster hill!", "officer cogle and mr.k shared a story and an important message of supporting one another for our first ever live, virtual, whole school read aloud using google meet!", "state of connecticut superior court judge and webster hill alumnus! susan quinn cobb shared a story, gave background on her job, and took questions from our students." • • • • • •

Topic words: life, love, learning, school, writing, book, read, yoga, kids, students, education, quotes, community, children, time, reading, learn, math, books, autism, world, chat, quote, story, change, motivation, writers, people, things, english

BART: webersen elementary media specialist, seesaw specialist, google certified educator, google trainer, apple certified educator.

UTGED: elementary library media specialist at webster hill elementary school. i love to connect with my students and help them grow as independent learners.

Target: i proudly teach all pk-5 webster hill students. we learn to think critically, research efficiently, meaningfully integrate technology, and find joy in reading.

Figure 9: A Twitter user sample and the related results. From top to down shows user history (source T^i), topic words (A^i), BART output, UTGED output, and reference self-introduction (target Y^i). The source text consists of 70 tweets, and here we randomly sample half of them to put in the figure for a better display.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section: line 606

☑ A2. Did you discuss any potential risks of your work?
Ethics statement section: line 625

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstraction section: line 001

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4: line 257; line 344; line 379

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will provide the license after publication.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We will provide it after publication.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics statement section: line 625

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We will provide it after publication.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3: line 217

C ☑ Did you run computational experiments?

Section 5: line 443

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Limitation section: line 606

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5: line 425

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We will provide it after publication.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We will provide it after publication.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics Statement Section

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics Statement Section

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We will provide it after publication.