Prompt Learning for News Recommendation

Zizhuo Zhang
School of Electronic Information and Communications,
Huazhong University of Science and Technology
Wuhan, China
zhangzizhuo@hust.edu.cn

Bang Wang
School of Electronic Information and Communications,
Huazhong University of Science and Technology
Wuhan, China
wangbang@hust.edu.cn

ABSTRACT

Some recent *news recommendation* (NR) methods introduce a Pre-trained Language Model (PLM) to encode news representation by following the vanilla pre-train and fine-tune paradigm with carefully-designed recommendation-specific neural networks and objective functions. Due to the inconsistent task objective with that of PLM, we argue that their modeling paradigm has not well exploited the abundant semantic information and linguistic knowledge embedded in the pre-training process. Recently, the pre-train, prompt, and predict paradigm, called *prompt learning*, has achieved many successes in natural language processing domain. In this paper, we make the first trial of this new paradigm to develop a *Prompt Learning for News Recommendation* (Prompt4NR) framework, which transforms the task of predicting whether a user would click a candidate news as a cloze-style mask-prediction task. Specifically, we design a series of prompt templates, including discrete, continuous, and hybrid templates, and construct their corresponding answer spaces to examine the proposed Prompt4NR framework. Furthermore, we use the prompt ensembling to integrate predictions from multiple prompt templates. Extensive experiments on the MIND dataset validate the effectiveness of our Prompt4NR with a set of new benchmark results.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Language models*; • **Computing methodologies → Natural language processing**.

KEYWORDS

Prompt Learning, News Recommendation, Pre-trained Language Model

ACM Reference Format:
Zizhuo Zhang and Bang Wang. 2023. Prompt Learning for News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591752

1 INTRODUCTION

Nowadays, online news platforms such as Google News has become a vital portal for people to efficient acquire daily information [4]. News recommendation (NR), as a filtering tool to alleviate the information overload problem [17, 24], can effectively help users to find their mostly interested news articles among the huge amount of news [25, 46].

Most existing neural NR methods have mainly focused on designing various ingenious neural networks to encode news' and users' representations [1, 11, 13, 30, 33, 37, 38, 42, 45, 56, 57, 60]. We summarize the common approaches in these models by Figure 1(a), where the core modules are news encoder, user encoder and similarity measure. In the literature, these modules have been implemented via different neural networks. For example, Wu et al. [45] leverage a multi-head self-attention network as news encoder and user encoder. Wang et al. [37] propose to learn a kind of hierarchical multi-level news representation via stacked dilated convolutions for fine-grained matching. In these neural models, the static word embeddings (e.g., Word2Vec [23] and Glove [27]) are mostly adopted as initializations in model training, which mainly focuses on mining in-domain information in a NR dataset, yet ignoring the abundant semantic and linguistic information from real-world large-scale corpus.

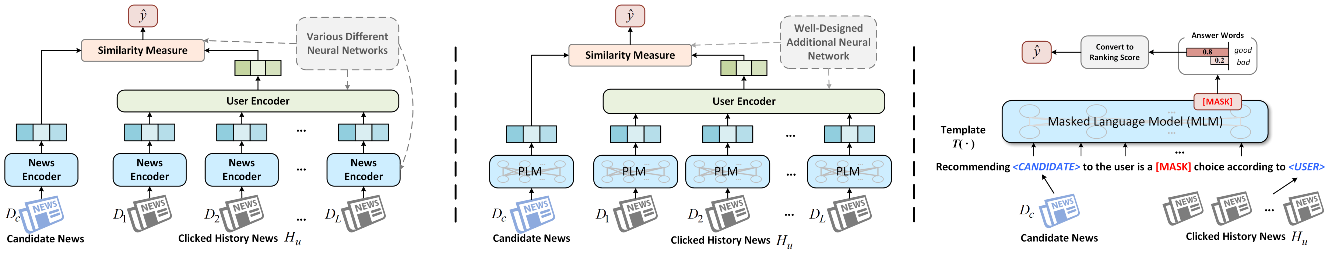
Some recent methods take one step further to introduce a *pre-trained language model* (PLM) for learning news representations [2, 9, 48, 54, 55, 58]. We summarize their common approaches by Figure 1(b), where the vanilla *pre-train and fine-tune* paradigm [15] is used to adapt the downstream NR task. In this paradigm, a PLM is only used as news encoder, yet another neural network is designed for encoding users. A NR-specific objective function is used to train the whole model. Although these methods show promising performance improvements, they have not well exploited the abundant encyclopedia-like knowledge in large-scale PLMs due to the inconsistency between the downstream NR objective and the PLM training objective.

Recently, a novel *pre-train, prompt and predict* paradigm named *prompt learning* has exhibited remarkable successes in many applications in the *natural language processing* (NLP) domain [3, 53]. The basic of this new paradigm is to reformulate a downstream task into the PLM training task via designing task-related *prompt template* and *answer words space* [6, 15]. For its promising potentials, we are also interested to examine the applicability and effectiveness of the prompt learning paradigm in the NR domain. To the best of our knowledge, this paper makes the first trial and proposes a *Prompt Learning for News Recommendation* (Prompt4NR) framework.

Figure 1(c) illustrates the design philosophy of our Prompt4NR framework, where we propose to convert a NR task into a cloze-style mask-prediction task. In particular, given the click history

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591752



(a) Common neural NR model framework

(b) Common PLM-based NR model framework with pre-train and fine-tune paradigm

(c) Idea of Prompt4NR framework with prompt learning paradigm

Figure 1: The differences of neural NR models, pre-train and fine-tune PLM-based models, and our Prompt4NR framework.

of a user $H_u = \{D_1, D_2, ..., D_L\}$ and a candidate news D_c, we first convert H_u into a sentence, denoted as <USER> so as to encode the user interests from history. We also convert D_c into a sentence, denoted as <CANDIDATE> for a candidate news. Then we design a *prompt template* T(<USER>, <CANDIDATE>) to concatenate the two sentences into another sentence with a [MASK] token. After passing a Masked Language Model (MLM), an answer space corresponding to the template is designed to predict the [MASK], which is then converted to a ranking score to determine whether this candidate news should be recommended to the user.

In this paper, we present a series of investigations on how the core design issues of Prompt4NR would impact on the recommendation performance. They include the design issues of (1) Template: What kind of templates is more suitable for integrating news data and user behaviors in the NR domain? (2) Verbalizer: How to map the recommendation labels to answer words for [MASK] prediction? (3) Ensemble: Can we integrate the advantages of different templates to boost the performance? For prompt template, we design three types of templates, including discrete, continuous, and hybrid templates, from the considerations of four mostly interested aspects between <USER> and <CANDIDATE> for the NR task, including the *semantic relevance*, *user emotion*, *user action* and *recommendation utility*. For verbalizer, we construct a binary answer space with two answer words with opposite senses according to the template, which corresponds to the real label of whether the user clicks the candidate news. For ensemble, we use multi-prompt ensembling to make a decision fusion of predictions from different templates. We design extensive experiments on the wide-used MIND dataset [52]. A set of new benchmark results validates the effectiveness of our Prompt4NR framework in terms of better recommendation performance over the state-of-the-art competitors.

2 RELATED WORK

In this section, we review the existing related work on news recommendation and prompt learning.

2.1 News Recommendation

Various neural networks have been widely developed to encode news and user representations for the NR task [1, 7, 10, 11, 13, 16, 26, 29–33, 37–39, 41–45, 47, 49–51, 56, 57, 59, 60]. For example, Wu et al. [45] propose the NRMS to adopt the multi-head attention network as news and user encoders. Wang et al. [37] propose the FIM to make the fine-grained matching via stacked dilated convolutions. Qi et al. [33] propose the HieRec that represents each user as a hierarchical interest tree based on topics and subtopics of news. Li et al. [13] propose the MINER to capture a user's multiple interests representations. Though effective, these shallow neural networks lack the ability to understand the deep linguistic and semantic information in news texts. Also their models learn knowledge only from supervised signals in the NR task, ignoring the potential benefits from real-world large-scale corpora.

Recently, with great success of PLM in the NLP domain, some methods have incorporated the PLM for the NR task and have achieved substantial improvements [2, 9, 48, 54, 55, 58]. For example, Wu et al. [48] propose an intuitionistic PLM-based NR framework, which directly instantiates the news encoder in existing neural NR models using a PLM. Bi et al. [2] use the multi-task learning framework to enhance the capability of BERT to encode multi-field information of news such as category and entity. These above methods all follow the vanilla pre-train and fine-tune paradigm with a NR-specific objective. Despite achieving some performance improvements, such paradigm may be suboptimal for exploiting knowledge learned from the pre-training process for the NR task, due to inconsistent objective with that of PLM [15].

2.2 Prompt Learning

Prompt learning is a novel learning strategy upon a PLM, which converts a downstream task into the form of [MASK] prediction using the PLM by adding template to the input texts. How to design prompt template is the core component of prompt learning. Generally, the types of templates can be categorized as discrete, continuous and hybrid templates [15]. Discrete templates consist of existing natural words, relying heavily on human experiences. For example, Petroni et al. [28] manually design template to probe knowledge in LMs. Schick et al. [34] propose a semi-supervised training procedure PET to reformulate specific tasks as cloze-style tasks. Although discrete template has succeeded in many tasks, handcrafted templates may be with costs and not globally optimal. To overcome such issues, continuous and hybrid templates introduce learnable prompt tokens to automatically search the templates, such as the AutoPrompt [35], Prefix Tuning [14], P-Tuning [19], P-Tuningv2 [18] and etc. In this paper, we focus on how to exploit the prompt learning for the NR task. We propose and experiment a set of discrete, continuous and hybrid templates to investigate different design considerations.

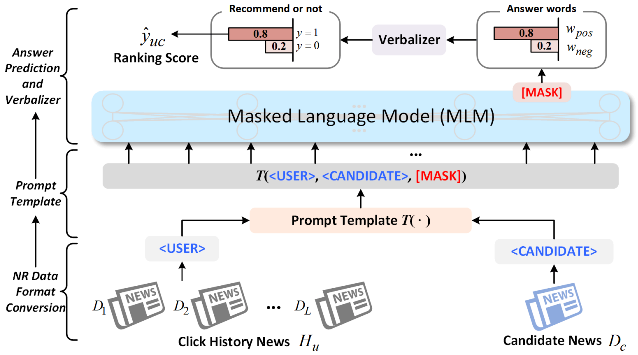


Figure 2: Illustration of the Prompt4NR Framework.

3 THE PROMPT4NR FRAMEWORK

We first present the problem statement of NR and next introduce our Prompt4NR framework. We then explain the training strategy and multi-prompt ensembling.

3.1 The NR Task and Prompt Overview

Denote the $\mathcal{U}$ and $\mathcal{D}$ as the user set and news set, respectively. Each news $D \in \mathcal{D}$ is mainly equipped with its title $Title = \{w_1, w_2, ..., w_M\}$ that is a sequence of words, where M is the number of words. Given a user u's click history $H_u = \{D_1, D_2, ..., D_L\}$ containing L clicked news and a candidate news D_c, the news recommendation (NR) task aims at predicting a ranking score $\hat{y}_{uc}$ to estimate the probability that the user u would click the candidate news D_c. The candidate news with the highest score will be recommended to the user.

Fig. 2 illustrates our Prompt4NR framework, which contains three main modules: (1) NR data format conversion (Section 3.2); (2) Prompt template (Section 3.3); (3) Answer prediction and verbalizer (Section 3.4). We explain their details as follows:

3.2 NR Data Format Conversion

Given a click history H_u and candidate news D_c, we convert them into a natural language sentence to adapt to the subsequent prompt learning paradigm, denoted as <USER> and <CANDIDATE>, respectively. For the <USER>, we concatenate the news titles of a user history H_u, where a *virtual token* [NCLS] is added at the beginning of each title to segment each clicked news. For the <CANDIDATE>, we adopt the title of the candidate news D_c. We formally denote them by:

$$\text{<USER>} \leftarrow \text{[NCLS]}\ Title_1\ ...\ \text{[NCLS]}\ Title_L$$
$$\text{<CANDIDATE>} \leftarrow Title_c$$

where $\{Title_1, ..., Title_L\}$ correspond to news titles in $H_u = \{D_1, D_2, ..., D_L\}$. Intuitively, the <USER> can be regarded as a summary of the user u's area of interest, and the <CANDIDATE> condenses the core textual semantics of a candidate news. Both of them serve as the input textual data for subsequent prompt templates.

3.3 Prompt Template

As the core component of Prompt4NR, prompt template $T(\cdot)$ wraps the input data (<USER>, <CANDIDATE>) to convert the NR task as a cloze-style task to predict the [MASK]:

$$x_{prompt} = T(\text{<USER>}, \text{<CANDIDATE>}, \text{[MASK]}) \tag{1}$$

As the first work exploiting prompt learning for the NR task, we are interested to know what kind of prompt templates would most benefit recommendation performance. To this end, we design three types of templates, including discrete, continuous and hybrid templates, from different perspectives of how to capture the matching signals between a user and a candidate news. Table 1 summarizes our designed prompt templates. We next introduce the design ideas behind these templates.

3.3.1 ***Discrete Templates.*** As the most common type of template engineering in prompt learning, discrete templates form the input data via human-interpretable natural language, which requires some prior experiential knowledge. We design four discrete templates from four different considerations, where each corresponds to one way to measure the matching signals between a user's interest and a candidate news. That is, we explore which style of [MASK] cloze as the similarity measurement is suitable for the NR task.

- **Semantic Relevance:** This is to examine whether the related news contents are the core motivation for a user to read a news. In other words, whether a user is with a kind of persistent interests to some particular topics and contents. To this end, we convert the NR task to determine the relevance between <CANDIDATE> and <USER>, and the answer words are chosen as *"related"* and *"unrelated"*.
- **User Emotion:** This is to investigate whether users' emotional reactions to news would be the most impacting factor. In other words, a user chooses to read a news, as if the news could mostly satisfy a user emotional needs. We use the emotion words *"interesting"* and *"boring"* as the answers to estimate the user emotional reaction to the <CANDIDATE>.
- **User action:** This is to study whether a MLM can directly serve as a click predictor. In other words, interest guides action and action reflects interest. After telling the MLM the <USER> and <CANDIDATE>, we let the MLM directly predict whether the user would click on the news, and the answer words are *"yes"* and *"no"*.
- **Recommendation utility:** This is to explore whether a MLM can itself make a judgement about the potential merits and demerits of recommending a candidate news, that is, the utility of making such a recommendation. To this end, we prompts the MLM with an utilization question, and the answer words are *"good"* and *"bad"* as the recommendation utility predictions.

The above templates, though seemingly with only a few differences on the template sentences and answer words, are expected to exploit the semantics and linguistics knowledge embedded in a large encyclopedia-like PLM through predefined natural sentences and preselected answer words. On the one hand, we note that this is the core philosophy of the prompt learning paradigm, that is, predicting the probability of an answer word from the PLM vocabulary, as if the task-specific input sentences have been inserted into the large corpus for training the PLM. On the other hand, such manually designed templates, though with well-designed natural sentences, are obviously not exhaustive for all possible cases. As such, we may use some virtual tokens to search a few more cases in a template, that is, a continuous template.

Table 1: Prompt templates designed in this paper, including discrete, continuous and hybrid templates.

Types	Perspectives	Templates T(<USER>, <CANDIDATE>, [MASK])	Answer Words
Discrete Template	Relevance	<CANDIDATE> is [MASK] to <USER>	related/unrelated
	Emotion	The user feels [MASK] about <CANDIDATE> according to his area of interest <USER>	interesting/boring
	Action	User: <USER> [SEP] News: <CANDIDATE> [SEP] Does the user click the news? [MASK]	yes/no
	Utility	Recommending <CANDIDATE> to the user is a [MASK] choice according to <USER>	good/bad
Continuous Template	Relevance	$[Q_1]...[Q_{n_2}]$ <CANDIDATE> $[M_1]...[M_{n_3}]$ [MASK] $[P_1]...[P_{n_1}]$ <USER>	related/unrelated
	Emotion	$[M_1]...[M_{n_3}]$ [MASK] $[Q_1]...[Q_{n_2}]$ <CANDIDATE> $[P_1]...[P_{n_1}]$ <USER>	interesting/boring
	Action	$[P_1]...[P_{n_1}]$ <USER> [SEP] $[Q_1]...[Q_{n_2}]$ <CANDIDATE> [SEP] $[M_1]...[M_{n_3}]$ [MASK]	yes/no
	Utility	$[Q_1]...[Q_{n_2}]$ <CANDIDATE> $[M_1]...[M_{n_3}]$ [MASK] $[P_1]...[P_{n_1}]$ <USER>	good/bad
Hybrid Template	Relevance	$[P_1]...[P_{n_1}]$ <USER> [SEP] $[Q_1]...[Q_{n_2}]$ <CANDIDATE> [SEP] This news is [MASK] to the user's area of interest	related/unrelated
	Emotion	$[P_1]...[P_{n_1}]$ <USER> [SEP] $[Q_1]...[Q_{n_2}]$ <CANDIDATE> [SEP] The user feels [MASK] about the news	interesting/boring
	Action	$[P_1]...[P_{n_1}]$ <USER> [SEP] $[Q_1]...[Q_{n_2}]$ <CANDIDATE> [SEP] Does the user click the news? [MASK]	yes/no
	Utility	$[P_1]...[P_{n_1}]$ <USER> [SEP] $[Q_1]...[Q_{n_2}]$ <CANDIDATE> [SEP] Recommending the news to the user is a [MASK] choice	good/bad

3.3.2 ***Continuous Template.*** Table 1 presents our design of four continuous templates, each corresponding to one discrete template. We add some virtual learnable tokens in front of the <USER>, <CANDIDATE> and [MASK], respectively, denoted as $[P_1]...[P_{n_1}]$, $[Q_1]...[Q_{n_2}]$ and $[M_1]...[M_{n_3}]$, where n_1, n_2, n_3 are numbers of virtual tokens. As for the answer words and token position setting, we refer to previous discrete templates. Although continuous templates provide the model more freedom, the embeddings of these virtual tokens are randomly initialized, which may introduce some ambiguities, leading to under-utilization of the PLM knowledge. We further design a kind of hybrid templates, trying to combine the advantages of both discrete and continuous templates.

3.3.3 ***Hybrid Template.*** In a hybrid template, we preserve those virtual tokens $[P_i]$ and $[Q_j]$ in front of <USER> and <CANDIDATE>, with the aim of automatically searching for the appropriate formats to present these information to PLM. We replace those virtual tokens $[M_k]$ with a natural language with the [MASK] token that is used for answer prediction. As presented in Table 1, we still design four representative natural sentences each corresponding to one of our design considerations. A hybrid template is hence composed of a continuous template, a [SEP] token and a natural sentence. Compared with those continuous templates, we argue that such hybrid templates can enjoy the virtual tokens for more choices by the continuous templates, yet still bearing natural sentences for guiding answer directions by the discrete templates.

3.4 Answer Prediction And Verbalizer

Given a click history H_u and a candidate news D_c, they correspond to a real label $y \in \{0, 1\}$ reflecting whether the user clicks the candidate news ($y = 1$) or not ($y = 0$). We design a *verbalizer* $v(\cdot)$ to map the labels to two answer words in the PLM vocabulary $\mathcal{W}$ as follows:

$$v(y) = \begin{cases} w_{pos}, & y = 1, \\ w_{neg}, & y = 0, \end{cases} \tag{2}$$

where $\mathcal{W}_a = \{w_{pos}, w_{neg}\} \subset \mathcal{W}$ is the answer word space, which can be different according to the used prompt templates. The NR task is converted into a cloze-style task that the pre-trained MLM $\mathcal{M}$ (e.g. BERT [5]) predicts the probability of answer words to be the [MASK]:

$$P(y|H_u, D_c) = P_{\mathcal{M}}\left([MASK] = v(y)|x_{prompt}\right), \tag{3}$$

where $P_{\mathcal{M}}\left([MASK] = v(y = 1)|x_{prompt}\right)$ can be regarded as the confidence of whether to recommend the current candidate news. We use it as the ranking score to form a recommendation list. We note that this paper considers a simple answer space construction with two PLM vocabulary words. We will investigate more complicated answer space construction with more vocabulary words and even virtual answers in our future work.

3.5 Training

Compared with the pre-train and fine-tune paradigm to train additional task-specific neural models, our Prompt4NR model only has parameters of PLM to tune. We adopt the cross entropy loss function to train our model:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log P_i + (1 - y_i)\log(1 - P_i)\right], \tag{4}$$

where y_i and P_i are the gold label and predicted probability of the i-th training instance, respectively. We use the AdamW [22] optimizer with L2 regularization for model training.

3.6 Multi-Prompt Ensembling

Different templates may have different advantages, as they each focus on a particular design consideration, resulting in their different utilizations of the linguistic and semantic knowledge in a PLM. Multi-prompt ensembling fuses the predictions of individual prompts to boost final decision [12, 15]. As we have no prior knowledge about which template is better, we simply sum the probability of the positive answer word in each prompt as the final ranking score:

$$\hat{y} = \sum_{e \in \mathcal{E}} P_e, \tag{5}$$

where P_e is the template e's output probability for w_{pos}, $\mathcal{E}$ is the set of fused templates. In this paper, we consider two kinds of multi-prompt ensembling. One is to fuse predictions from the same

Table 2: Dataset statistics.

MIND	# Users	# News	# Clicks	# Impressions
	94,057	65,238	347,727	230,117

type of templates, where $\mathcal{E} = \{Relevance, Emotion, Action, Utility\}$. That is, a discrete ensembling is to fuse the four discrete templates' predictions. The other is to fuse predictions from different types of templates, named *cross-type ensembling*.

4 EXPERIMENT SETTINGS

In this section, we introduce our experimental settings, including dataset, parameter settings, evaluation metrics and baselines.

4.1 Dataset

We conduct experiments on the public real-world NR benchmark dataset MIND[1] [52], where users' behaviors are recorded by impressions. An impression log records the clicked and non-clicked news that are displayed to a user at a specific time and his historical news click behaviors before this impression. MIND collects the user behavior logs from October 12 to November 22, 2019 from the Microsoft News platform. Following previous work [2, 33], user data in the first four weeks is used to construct users' history, user data in penultimate week is used as training set and user data in the last week as testing set, we extract 5% impressions from training set to form the validating set. Table 2 summarizes the dataset statistics.

4.2 Parameter Settings

In our experiments, we employ the base BERT (12 layers, bert-base-uncased) [5] implemented by HuggingFace[2] transformers [40] as the PLM to experiment our Prompt4NR. We adopt the AdamW optimizer with learning rate lr=2e-5 to train the model. We run the model with 8 NVIDIA RTX-A5000 GPUs in distributed data parallel and the batch size on each of them is set as 16, which is equal to batch size 128 on a single GPU. We apply negative sampling with ratio 4 the same as other baselines. Following previous work [37, 45, 48], we adopt a user's most recent 50 clicked news as the user's click history. For each news in history, we set the maximum length of the title to be 10 words; for each candidate news, we set the maximum length of the title to be 20 words. Besides, we use the average AUC, MRR, NDCG@5 and NDCG@10 over all impressions as the evaluation metrics, which are widely used in the NR studies [33, 37, 48]. All hyperparameters are adjusted on the validating set. We have released the source code at https://github.com/resistzzz/Prompt4NR.

4.3 Baselines

We compare our methods with following competitors, which can be categorized into neural NR models and BERT-enhanced models:

Neural NR models: Various neural networks have been specially designed for the NR task, including: (1) *NPA* [42] uses the personalized attention network on words and history clicked news for news and user representation learning. (2) *LSTUR* [1] captures a user's short-term and long-term interests via a GRU network and the user's embedding respectively. (3) *NRMS* [45] uses the multi-head self-attention to learn news and user representations. (4) *FIM* [37] adopts a hierarchical dilated convolution network to encode news representation from multiple grained views.

Table 3: The overall comparison of performance results.

Dataset		MIND			
Baselines		AUC	MRR	NDCG@5	NDCG@10
Neural Methods	NPA	64.65	30.01	33.14	39.47
	LSTUR	65.87	30.78	33.95	40.15
	NRMS	65.63	30.96	34.13	40.52
	FIM	65.34	30.64	33.61	40.16
PLM-based Methods	BERT-NPA	67.56	31.94	35.34	41.73
	BERT-LSTUR	68.28	32.58	35.99	42.32
	BERT-NRMS	68.60	32.97	36.55	42.78
Prompt4NR (BERT)		AUC	MRR	NDCG@5	NDCG@10
Discrete Template	Relevance	68.77	33.42	37.20	43.36
	Emotion	68.77	33.29	37.12	43.19
	Action	68.76	33.22	37.02	43.26
	Utility	68.94	33.62	37.47	43.61
	Ensembling	**69.34**	**33.76**	**37.71**	**43.80**
Continuous Template	Relevance	69.25	33.72	37.75	43.79
	Emotion	68.76	33.51	37.39	43.47
	Action	68.58	33.37	37.17	43.30
	Utility	69.10	33.96	37.91	43.92
	Ensembling	**69.43**	**34.06**	**38.11**	**44.14**
Hybrid Template	Relevance	68.47	33.26	37.20	43.24
	Emotion	68.59	33.26	37.19	43.29
	Action	**69.37**	**34.02**	**37.96**	**44.00**
	Utility	68.79	33.45	37.35	43.49
	Ensembling	69.22	33.78	37.77	43.87
Cross-Type Ensembling		**69.64**	**34.26**	**38.30**	**44.33**

[1] Boldface with blue indicates the best results in the whole table. In each type of template, only boldface indicates the best results, the second best is underlined.
[2] The cross-type ensembling fuses the best performings in each type for decision, i.e., $\mathcal{E} = \{Discrete\text{-}Utility, Continuous\text{-}Utility, Hybrid\text{-}Action\}$.
[3] The improvements are significant ($p < 0.01$) as validated by student's t-test.

BERT-enhanced models: Wu et al. [48] adopt the BERT as the news encoder on several of the above methods. We reproduce three baselines here for comparison, denoted as *BERT-NPA*, *BERT-LSTUR* and *BERT-NRMS*, for which we only replace their news encoder with BERT and keep the other neural parts (e.g., user encoder) unchanged.

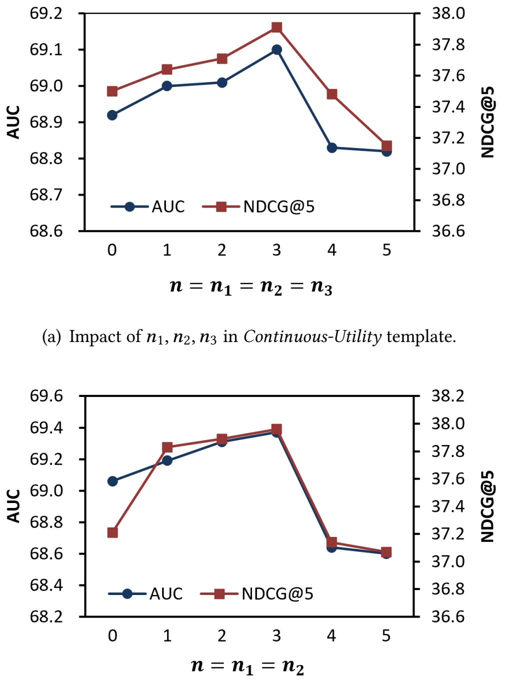
5 EXPERIMENT RESULTS

5.1 Main Experiment Results

Table 3 summarizes the comparison of our Prompt4NR with these state-of-the-art baselines. The boldface with blue indicates the best results in the whole table. In each type of template, only boldface indicates the best result in its kind, while the second best is underlined. From these results, we have following observations:

Firstly, the models using BERT as news encoder are consistently better than those neural models using shallow neural networks as news encoder, e.g., BERT-NRMS outperforms NRMS. This indicates that the pre-trained BERT brings benefits to news encoding, which

[1] https://msnews.github.io/
[2] https://huggingface.co/



(a) Impact of n_1, n_2, n_3 in *Continuous-Utility* template.

(b) Impact of n_1, n_2 in *Hybrid-Action* template.

Figure 3: Impact of the number of virtual tokens in the continuous and hybrid templates.

Table 4: Performance of using different PLMs.

Discrete-Utility	AUC	MRR	NDCG@5	NDCG@10
BERT	68.94	33.62	37.47	43.61
RoBerta	**69.20**	**34.00**	**37.77**	**43.96**
DeBerta	69.01	33.98	37.67	43.74

is attributed to the rich semantic and linguistic knowledge learned by BERT during the pre-training process.

Secondly, our Prompt4NR methods based on prompt-learning paradigm, either discrete, continuous or hybrid templates, almost outperform those BERT-enhanced models based on the pre-train, fine-tune paradigm. Especially, the metrics of MRR, NDCG@5 and NDCG@10 all surpass 33.00, 37.00 and 43.00, respectively, and AUC reults on some templates surpass 69.00. This reflects the superiority of the prompt learning paradigm for developing knowledge embedded in pre-trained BERT to support the NR task.

Thirdly, three different types of templates achieve comparable performances, and the order of their bests is "*Discrete-Utility* < *Continuous-Utility* < *Hybrid-Action*", in which *Hybrid-Action* is the best in all of the one-single kind templates. Furthermore, the performance gaps within the four discrete templates are relatively smaller than those within the continuous and hybrid templates. This is because of those learnable virtual tokens in continuous and hybrid templates that bring more potentials yet with more variations.

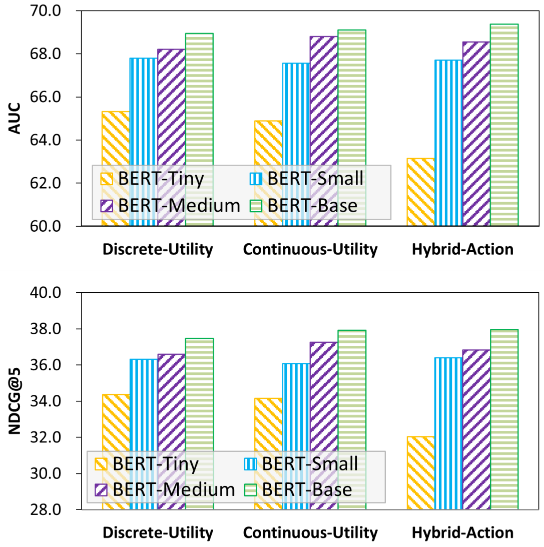


Figure 4: Impact of the BERT scale.

Finally, performing multi-prompt ensembling obviously improves the recommendation performance than using only a one-single prompt on the discrete and continuous templates. This indicates the effectiveness of prompt ensembling to fuse multi-prompts for better decision. On the hybrid templates, the ensembling result is better than the relevance, emotion and utility template, but worse than the action template. This suggests that the prerequisite for prompt ensemble to work is that the performance gap between the individual fused prompts should not be too large. In addition, cross-type ensembling, which fuses the three best in each type, i.e., *Discrete-Utility*, *Continuous-Utility* and *Hybrid-Action*, achieves the best performance. This reflects that prompt ensembling should not be limited by template type, and that cross-type ensembling may be a good choice in some cases.

By the way, we notice that the three bests in each template type are from the *Utility*, *Utility* and *Action* prompt, respectively. In the four considerations of prompt templates, the *Action* and *Utility* prompts are directly related to the recommendation task; While the *Relevance* and *Emotion* prompts are not directly related. The *Relevance* and *Emotion* prompts have transformed the interest matching by inferring the semantic relevance between news texts and predicting users' emotional reactions, respectively. We suspect those templates directly associated with the recommendation properties other than using another kind of underlying inference may be more suitable for the NR task.

5.2 Hyperparameter Analysis

The core hyperparameters of our Prompt4NR are the numbers of virtual tokens in continuous templates (i.e. n_1, n_2, n_3) and hybrid templates (i.e. n_1, n_2). We notice that there are so many combinations of them taking different values, even if their values vary only in a small range. We adopt a coarse hyperparameter tuning strategy, that is, we let n_1, n_2, n_3 take the same value for tuning, denoted as $n = n_1 = n_2 = n_3$.

In particular, we vary n in $\{0, 1, 2, 3, 4, 5\}$, where we notice that when $n = 0$ without any tokens is called NullPrompt [21]. Considering the paper length, we only present the tuning results of the best two, i.e., *Continuous-Utility* and *Hybrid-Action*, the trends of the others are similar. Figure 3 plots the recommendation performance against different values of n. We find that the trends are the same for both continuous and hybrid templates. That is, as n increases, the performance first improves and then decreases. This indicates that only a few of virtual tokens are enough for prompting: Too fewer may lack guidance to organize information to a PLM; While too many may suffer from more ambiguities because randomly initialized virtual tokens lack pre-training knowledge. Furthermore, we notice that even Null Prompt (i.e. $n = 0$), the AUC can reach 68.9, which has surpassed those BERT-enhanced methods. This again reflects the superiority of the prompt learning paradigm on the NR task, with its ability to better exploit the PLM to capture the potential relationships between news texts.

5.3 Impact of PLM

5.3.1 ***Impact of The PLM's Scale.*** We investigate the impact of using different scales of BERT, i.e., BERT-Tiny (2 layers), BERT-Small (4 layers), BERT-Medium (8 layers) and BERT-Base (12 layers). We perform different scales of BERT on the best template in each type. Figure 4 plots the results of AUC and NDCG@5. We observe that using larger PLMs with more parameters and deeper layers usually leads to better recommendation performance. This is not unexpected, because a larger PLM usually has stronger abilities to capture the deep linguistic and semantic information of news texts, and the performance of prompt learning also depends on the PLM's such capability. We reasonably guess the performance can be further improved by using larger BERT (e.g., BERT-Large with 24 layers). However, we believe that we should strike a balance between the recommendation performance and the model scale, as an oversized PLM may not be suitable for real application scenarios.

5.3.2 ***Impact of Different PLMs.*** We replace the BERT with other two popular PLMs, viz., the RoBerta [20] and DeBerta [8], to investigate the impact of using different PLMs. Both the RoBerta and DeBerta are the improved version of the BERT[3]. The RoBerta removes the next sentence prediction and is pre-trained on a larger corpus with larger batch size; The DeBerta introduces a disentangled attention mechanism and an enhanced mask decoder to improve BERT. Table 4 presents the experiment results. We observe that both the RoBerta and DeBerta outperform the BERT, and RoBerta achieves the best results. This suggests that the downstream NR task can benefit from the improvements made in the pre-training process.

5.4 Performance of Few-shot Learning

Some researchers have reported that the prompt learning paradigm has some robustness under few-shot scenarios for some NLP tasks, such as text classification [36], implicit discourse relation recognition [53]. We would also like to examine our Prompt4NR and competitors under few-shot learning scenarios. We adopt the down-sampling to gradually reduce the training set, while keeping the validating set and testing set unchanged.

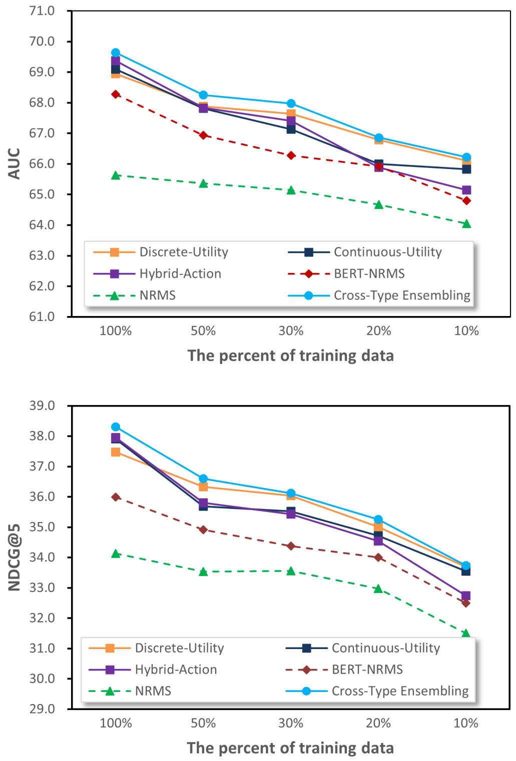


Figure 5: Performance comparison of few-shot learning.

Figure 5 summarizes the performance comparison of few-shot learning, where dashed lines denote two competitors NRMS and BERT-NRMS, and full lines denote three one-single templates and a cross-type ensembling of our Prompt4NR. Not surprisingly, both our Prompt4NR and competitors suffer from performance degradation as training data decreases. We observe that the line of NRMS is always at the bottom; The line of our three one-single templates are higher than that of the BERT-NRMS in most of cases; The line of cross-ensembling is higher than all the others. The results show that the PLM has stronger capability to handle few-shot scenarios than shallow neural networks, viz., the prompt learning is more robust than the vanilla pre-train and fine-tune paradigm. The prompt ensembling strategy can further boost the robustness. Furthermore, it is observed that when decreasing training set, the *Discrete-Utility* is in an upper hand position more often relative to *Continuous-Utility* and *Hybrid-Action*. This may be due to that those virtual tokens in the continuous and hybrid templates are not sufficiently trained as the training data decreases.

5.5 Visualization

To examine what BERT has learned in our Prompt4NR framework, we visualize the attention weights in BERT of a random case in Figure 6. Since BERT has 12-layers and each layer has 12-heads, we visualize the [MASK]'s attention weights of the first layer and

[3] BERT, RoBerta and DeBerta all adopt the base version with 12 layers. Their model name on huggingface are {bert-base-uncased, roberta-base, deberta-base}.

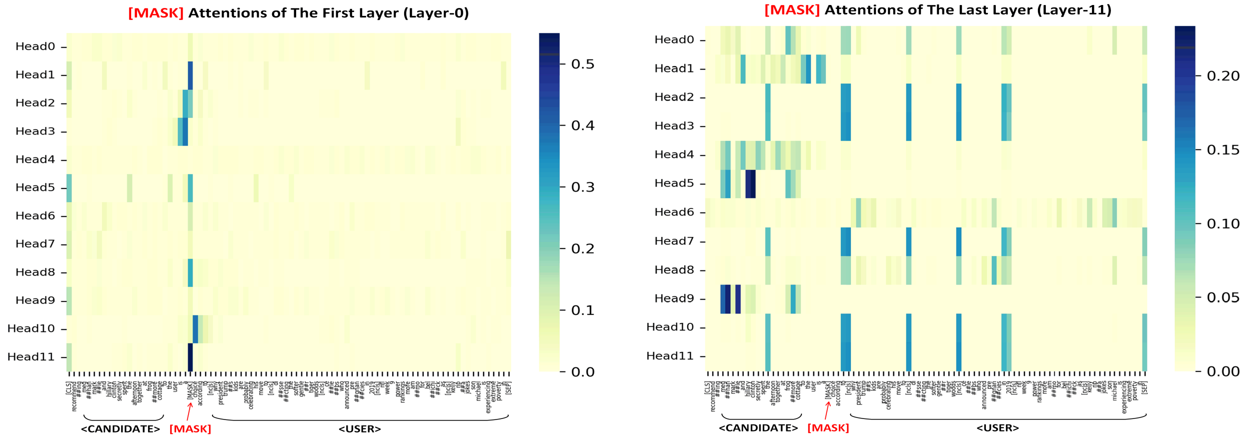


Figure 6: Visualization of the BERT's 12-heads attention weights of the [MASK] token for a sampled instance: The left is the attention weights of the first layer; The right is the last layer. The X-axis is the input sentence, the Y-axis is the 12 heads of BERT, and the BERT has been trained by using the *Discrete-Utility* template.

the last layer to help us understand the learning process. In the first layer, the attentions mainly focus on these tokens in the area around [MASK]. In the last layer, we observe that the focus is different for different heads. Some heads still focus on the area around [MASK], e.g., Head1. Some heads pay more attention on the area of <CANDIDATE>, e.g., Head4, Head5 and Head9. Some heads pay more attention on the area of <USER>, where {Head6, Head8} and {Head2, Head3, Head7, Head10, Head11} focus on different points; Head6 and Head8 have wide attentions on specific tokens in <USER>, we called token-level attention; While Head2, Head3, Head7, Head10 and Head11 focus especially on the areas around of several [NCLS] tokens, we called news-level attention, because the virtual token [NCLS] at the beginning of each historical news may have ability to represent its following news. The attentions of Head0 is dispersed throughout the sentence. In summary, compared to the first layer, the last layer has clearer attention points, and each head makes its contributions. This reflects the fact that the BERT, by being trained in our Prompt4NR framework, can discriminately use available information for the [MASK] prediction and recommendation.

5.6 Case Study

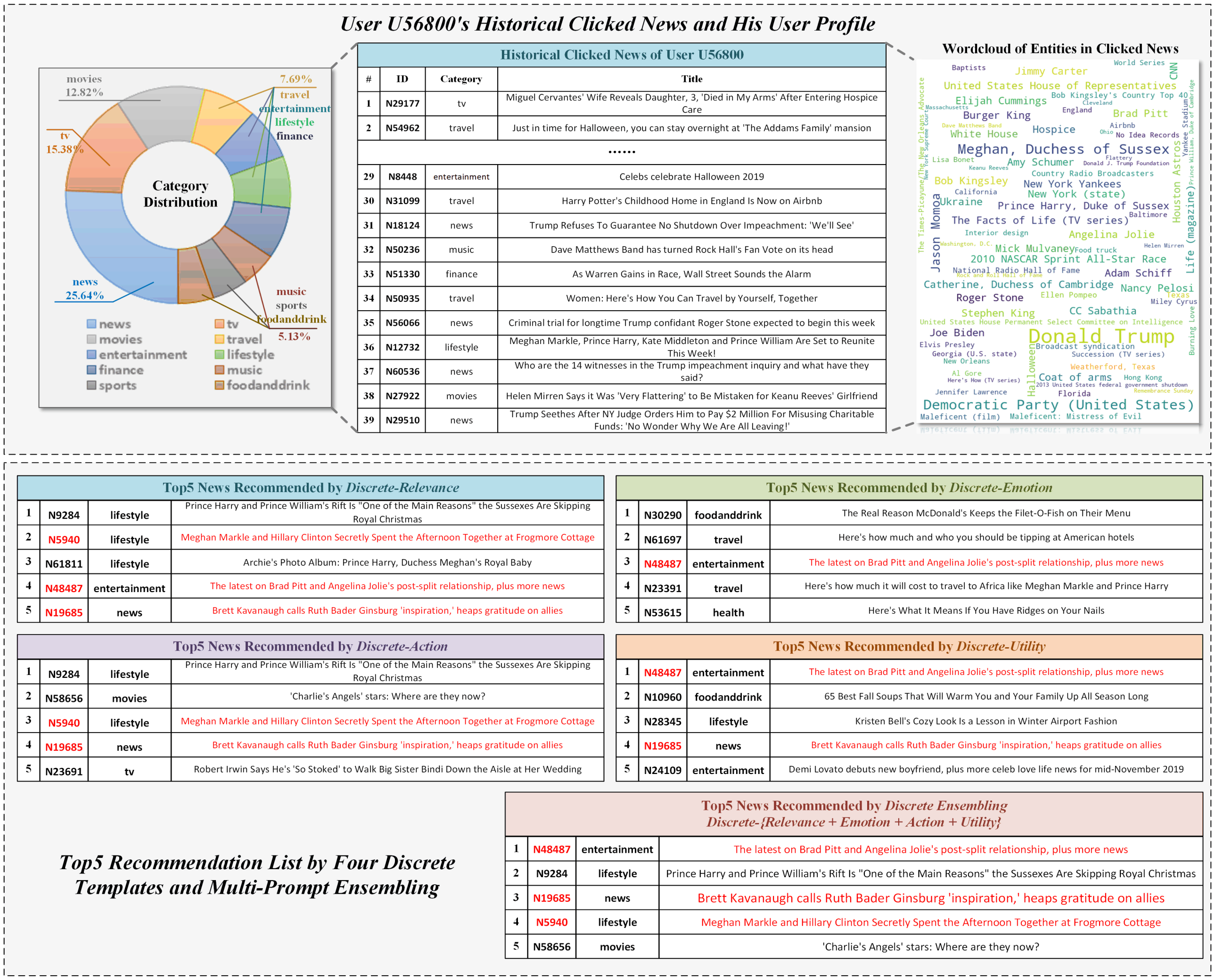
We conduct a case study to further intuitively shed light on the effectiveness of our Prompt4NR. Figure 7 illustrates the top5 recommendation list formed by four one-single *Discrete-*{*Relevance, Emotion, Action, Utility*} template and a *Discrete-Ensembling* in an impression. In this impression, it consists of 39 historical news clicked by the user *U56800* and 60 candidate news for ranking[4]. For more intuitive presentation, we mark the category label of news. Due to space consideration, not all 39 history news are shown in details, but we manually make a coarse user profile, including a category distribution (i.e., pie chart) and an entity frequency distribution (i.e., in the word-cloud, the higher frequency, the larger the word size) of his history news. From this case, we have following observations:

This user has clicked news covering 10 categories. From the word-cloud of the user's clicked entities, words like "*Meghan, Duchess of Sussex, Prince Harry, Duke of Sussex, Duchess of Cambridge*" are frequently in sight, and we can intuitively guess that the royal lifestyle may be one of the user's interests, but not the only one. In other words, this user's areas of interest are diverse. Besides, these candidate news are from multiple fields covering 14 categories (not plot due to space) and involve with numerous different entities. These observations imply that making an accurate recommendation for this user may not be an easy task.

Although not easy, it is observed that our five recommendation lists all contain the ground-truth of user actually clicked news, simply called as *true news*. We notice that each template has its own style of recommendation list. For example, the *Discrete-Relevance* hits more true news, but less diversity. The *Discrete-Emotion* recommends five news belonging to five different categories, even the "health" news never appears in the user's history. The *Discrete-Action* and *Discrete-Utility* both hit two true news: But the *Discrete-Action* tends to the user's interest of royal lifestyle, and *Discrete-Utility* finds another interest point of "entertainment" news. These results prove our motivation that though different templates seemingly with only a few differences, they may pay attention on different knowledge embedded in a PLM. Furthermore, it is observed that the *Discrete-Ensembling* achieves a win-win situation compared with each one-single template. Specifically, the *Discrete-Ensembling* also hits three true news, and their ranking positions are more beneficial: Five recommended news belonging to four categories presents diverse news to the user. This is an indication that the multi-prompt ensembling has the ability to incorporate the advantages of each one-single prompt for better recommendation.

Incidentally, we notice an interesting point. The 20-th historical news *N31748* is "Trailer - Charlie's Angels" about movies. In the top5 news recommended by the *Discrete-Action*, there is a news *N58656* "'Charlie's Angels' stars: Where are they now?", which is very related to *N31748.* Although *N58656* is not the true news, we believe that this is a nice recommendation, which can guide the

[4] These identifiers presented in the case study can be directly matched to the raw dataset, such as *Uxxxx, Nxxxx.*



User U56800's Historical Clicked News and His User Profile

Historical Clicked News of User U56800

#	ID	Category	Title
1	N29177	tv	Miguel Cervantes' Wife Reveals Daughter, 3, 'Died in My Arms' After Entering Hospice Care
2	N54962	travel	Just in time for Halloween, you can stay overnight at 'The Addams Family' mansion
......			
29	N8448	entertainment	Celebs celebrate Halloween 2019
30	N31099	travel	Harry Potter's Childhood Home in England Is Now on Airbnb
31	N18124	news	Trump Refuses To Guarantee No Shutdown Over Impeachment: 'We'll See'
32	N50236	music	Dave Matthews Band has turned Rock Hall's Fan Vote on its head
33	N51330	finance	As Warren Gains in Race, Wall Street Sounds the Alarm
34	N50935	travel	Women: Here's How You Can Travel by Yourself, Together
35	N56066	news	Criminal trial for longtime Trump confidant Roger Stone expected to begin this week
36	N12732	lifestyle	Meghan Markle, Prince Harry, Kate Middleton and Prince William Are Set to Reunite This Week!
37	N60536	news	Who are the 14 witnesses in the Trump impeachment inquiry and what have they said?
38	N27922	movies	Helen Mirren Says it Was 'Very Flattering' to Be Mistaken for Keanu Reeves' Girlfriend
39	N29510	news	Trump Seethes After NY Judge Orders Him to Pay $2 Million For Misusing Charitable Funds: 'No Wonder Why We Are All Leaving!'

Top5 News Recommended by *Discrete-Relevance*

1	N9284	lifestyle	Prince Harry and Prince William's Rift Is "One of the Main Reasons" the Sussexes Are Skipping Royal Christmas
2	N5940	lifestyle	Meghan Markle and Hillary Clinton Secretly Spent the Afternoon Together at Frogmore Cottage
3	N61811	lifestyle	Archie's Photo Album: Prince Harry, Duchess Meghan's Royal Baby
4	N48487	entertainment	The latest on Brad Pitt and Angelina Jolie's post-split relationship, plus more news
5	N19685	news	Brett Kavanaugh calls Ruth Bader Ginsburg 'inspiration,' heaps gratitude on allies

Top5 News Recommended by *Discrete-Emotion*

1	N30290	foodanddrink	The Real Reason McDonald's Keeps the Filet-O-Fish on Their Menu
2	N61697	travel	Here's how much and who you should be tipping at American hotels
3	N48487	entertainment	The latest on Brad Pitt and Angelina Jolie's post-split relationship, plus more news
4	N23391	travel	Here's how much it will cost to travel to Africa like Meghan Markle and Prince Harry
5	N53615	health	Here's What It Means If You Have Ridges on Your Nails

Top5 News Recommended by *Discrete-Action*

1	N9284	lifestyle	Prince Harry and Prince William's Rift Is "One of the Main Reasons" the Sussexes Are Skipping Royal Christmas
2	N58656	movies	'Charlie's Angels' stars: Where are they now?
3	N5940	lifestyle	Meghan Markle and Hillary Clinton Secretly Spent the Afternoon Together at Frogmore Cottage
4	N19685	news	Brett Kavanaugh calls Ruth Bader Ginsburg 'inspiration,' heaps gratitude on allies
5	N23691	tv	Robert Irwin Says He's 'So Stoked' to Walk Big Sister Bindi Down the Aisle at Her Wedding

Top5 News Recommended by *Discrete-Utility*

1	N48487	entertainment	The latest on Brad Pitt and Angelina Jolie's post-split relationship, plus more news
2	N10960	foodanddrink	65 Best Fall Soups That Will Warm You and Your Family Up All Season Long
3	N28345	lifestyle	Kristen Bell's Cozy Look Is a Lesson in Winter Airport Fashion
4	N19685	news	Brett Kavanaugh calls Ruth Bader Ginsburg 'inspiration,' heaps gratitude on allies
5	N24109	entertainment	Demi Lovato debuts new boyfriend, plus more celeb love life news for mid-November 2019

Top5 Recommendation List by Four Discrete Templates and Multi-Prompt Ensembling

Top5 News Recommended by *Discrete Ensembling*
Discrete-{Relevance + Emotion + Action + Utility}

1	N48487	entertainment	The latest on Brad Pitt and Angelina Jolie's post-split relationship, plus more news
2	N9284	lifestyle	Prince Harry and Prince William's Rift Is "One of the Main Reasons" the Sussexes Are Skipping Royal Christmas
3	N19685	news	Brett Kavanaugh calls Ruth Bader Ginsburg 'inspiration,' heaps gratitude on allies
4	N5940	lifestyle	Meghan Markle and Hillary Clinton Secretly Spent the Afternoon Together at Frogmore Cottage
5	N58656	movies	'Charlie's Angels' stars: Where are they now?

Figure 7: Case study of the top-5 news recommended by the four Discrete-{Relevance, Emotion, Action, Utility} templates and Discrete-Ensembling for an impression of user *U56800*. The recommended news actually clicked by the user is highlighted in red. The top is the historically clicked news and handmade profile of a user *U56800*, including the category distribution and entity word-cloud of his history news. The bottom presents the five recommendation lists.

user to review his potential interest and make a click action. This is a reflection of the Prompt4NR capability to self-discover and serialize latent relations between news.

In addition, it is observed that the four one-single templates are not perfect for thoroughly mining the user's interests. For example, no sports news is recommended here, but there is a sport news *N40094* of "Baker Mayfield quick to condemn teammate Myles Garrett for brawl, attack on Mason Rudolph" actually clicked by the user. This suggests us that our four one-single templates are not enough to cover all recommendation considerations, which motivates us to design more excellent templates in the future work.

6 CONCLUSION

In this paper, we have proposed the Prompt4NR, a prompt learning framework for news recommendation, which transforms the NR task as a cloze-task for the [MASK] prediction task. As the first trial work, we have conducted extensive experiments with a set of various types of prompt templates, including discrete, continuous and hybrid templates. Moreover, we have adopted the multi-prompt ensembling to incorporate advantages from different prompts. Experiment results have validate the superiority of our Prompt4NR over the state-of-the-art competitors.

7 ACKNOWLEDGEMENTS

This work is supported in part by National Natural Science Foundation of China (Grant No: 62172167). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

REFERENCES

[1] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 336–345.

[2] Qiwei Bi, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Hanfang Yang. 2022. MTRec: Multi-Task Learning over BERT for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2663–2669.

[3] Lu Dai, Bang Wang, Wei Xiang, and Yijun Mo. 2022. Bi-Directional Iterative Prompt-Tuning for Event Argument Extraction. *arXiv preprint arXiv:2210.15843* (2022).

[4] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. 271–280.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[7] Shansan Gong and Kenny Q. Zhu. 2022. Positive, Negative and Neutral: Modeling Implicit Feedback in Session-Based News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1185–1195.

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

[9] Qinglin Jia, Jingjie Li, Qi Zhang, Xiuqiang He, and Jieming Zhu. 2021. RMBERT: News recommendation via recurrent reasoning memory network over BERT. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 1773–1777.

[10] Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018. Weave&rec: A word embedding based 3-d convolutional network for news recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1855–1858.

[11] Taeho Kim, Yungi Kim, Yeon-Chang Lee, Won-Yong Shin, and Sang-Wook Kim. 2022. Is It Enough Just Looking at the Title? Leveraging Body Text To Enrich Title Words Towards Accurate News Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4138–4142.

[12] 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*. 3045–3059.

[13] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-Interest Matching Network for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 343–352.

[14] 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 (Volume 1: Long Papers)*. 4582–4597.

[15] 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. *arXiv preprint arXiv:2107.13586* (2021).

[16] Rui Liu, Huilin Peng, Yong Chen, and Dell Zhang. 2020. HyperNews: Simultaneous News Recommendation and Active-Time Prediction via a Double-Task Deep Neural Network.. In *IJCAI*. 3487–3493.

[17] Shenghao Liu, Bang Wang, and Minghua Xu. 2017. Event recommendation based on graph random walking and history preference reranking. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 861–864.

[18] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602* (2021).

[19] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385* (2021).

[20] 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 arXiv:1907.11692* (2019).

[21] Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2824–2835.

[22] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[24] Yijun Mo, Bixi Li, Bang Wang, Laurence T Yang, and Minghua Xu. 2018. Event recommendation in social networks based on reverse random walk and participant scale control. *Future Generation Computer Systems* 79 (2018), 383–395.

[25] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1933–1942.

[26] Keunchan Park, Jisoo Lee, and Jaeho Choi. 2017. Deep neural networks for news recommendations. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2255–2258.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[28] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. 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)*. 2463–2473.

[29] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. PP-Rec: News Recommendation with Personalized User Interest and Time-aware News Popularity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 5457–5467.

[30] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. FUM: Fine-Grained and Fast User Modeling for News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1974–1978.

[31] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. FUM: Fine-grained and Fast User Modeling for News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1974–1978.

[32] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News Recommendation with Candidate-Aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1917–1921.

[33] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 5446–5456.

[34] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 255–269.

[35] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. [n. d.]. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4222–4235.

[36] Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021. TransPrompt: Towards an Automatic Transferable Prompting Framework for Few-shot Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2792–2802.

[37] Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained interest matching for neural news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 836–845.

[38] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*. 1835–1844.

[39] Jingkun Wang, Yipu Chen, Zichun Wang, and Wen Zhao. 2021. Popularity-Enhanced News Recommendation with Multi-View Interest Representation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1949–1958.

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[41] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3863–3869.

[42] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584.

[43] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*. 1154–1159.

[44] Chuhan Wu, Fangzhao Wu, Mingxiao An, Tao Qi, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with heterogeneous user behavior. 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)*. 4874–4883.

[45] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. 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)*. 6389–6394.

[46] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2022. Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems (TOIS)* (2022).

[47] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. User Modeling with Click Preference and Reading Satisfaction for News Recommendation.. In *IJCAI*. 3023–3029.

[48] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[49] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Two Birds with One Stone: Unified Model Learning for Both Recall and Ranking in News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 3474–3480.

[50] Chuhan Wu, Fangzhao Wu, Tao Qi, Qi Liu, Xuan Tian, Jie Li, Wei He, Yongfeng Huang, and Xing Xie. 2022. Feedrec: News feed recommendation with various user feedbacks. In *Proceedings of the ACM Web Conference 2022*. 2088–2097.

[51] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2560–2564.

[52] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.

[53] Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. ConnPrompt: Connective-cloze Prompt Learning for Implicit Discourse Relation Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*. 902–911.

[54] Shitao Xiao, Zheng Liu, Yingxia Shao, Tao Di, Bhuvan Middha, Fangzhao Wu, and Xing Xie. 2022. Training large-scale news recommenders with pretrained language models in the loop. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4215–4225.

[55] Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, Tao Qi, and Qi Liu. 2021. Tiny-NewsRec: Efficient and Effective PLM-based News Recommendation. *arXiv preprint arXiv:2112.00944* (2021).

[56] Hui Zhang, Xu Chen, and Shuai Ma. 2019. Dynamic news recommendation with hierarchical attention network. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1456–1461.

[57] Qi Zhang, Qinglin Jia, Chuyuan Wang, Jingjie Li, Zhaowei Wang, and Xiuqiang He. 2021. Amm: Attentive multi-field matching for news recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1588–1592.

[58] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation.. In *IJCAI*. 3356–3362.

[59] Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2021. Combining explicit entity graph with implicit text information for news recommendation. In *Companion Proceedings of the Web Conference 2021*. 412–416.

[60] Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. Dan: Deep attention neural network for news recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5973–5980.