Low-Resource Personal Attribute Prediction from Conversations

Yinan Liu[1], Hu Chen[1], Wei Shen[1]*, Jiaoyan Chen[2]

[1] TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China
[2] Department of Computer Science, The University of Manchester
{liuyn,2120210473}@mail.nankai.edu.cn, shenwei@nankai.edu.cn, jiaoyan.chen@manchester.ac.uk

Abstract

Personal knowledge bases (PKBs) are crucial for a broad range of applications such as personalized recommendation and Web-based chatbots. A critical challenge to build PKBs is extracting personal attribute knowledge from users' conversation data. Given some users of a conversational system, a personal attribute and these users' utterances, our goal is to predict the ranking of the given personal attribute values for each user. Previous studies often rely on a relative number of resources such as labeled utterances and external data, yet the attribute knowledge embedded in unlabeled utterances is underutilized and their performance of predicting some difficult personal attributes is still unsatisfactory. In addition, it is found that some text classification methods could be employed to resolve this task directly. However, they also perform not well over those difficult personal attributes. In this paper, we propose a novel framework PEARL to predict personal attributes from conversations by leveraging the abundant personal attribute knowledge from utterances under a low-resource setting in which no labeled utterances or external data are utilized. PEARL combines the biterm semantic information with the word co-occurrence information seamlessly via employing the updated prior attribute knowledge to refine the biterm topic model's Gibbs sampling process in an iterative manner. The extensive experimental results show that PEARL outperforms all the baseline methods not only on the task of personal attribute prediction from conversations over two data sets, but also on the more general weakly supervised text classification task over one data set.

Introduction

Personal knowledge bases (PKBs) (Balog and Kenter 2019; Yen, Huang, and Chen 2019)-structured information about entities personally related to the users, their attributes, and the relations between them-are popular nowadays. They can supply plentiful personal background knowledge for a broad range of downstream applications like Web-based chatbots (Ghazvininejad et al. 2018), personalized recommendation (Balog, Radlinski, and Arakelyan 2019; Luo et al. 2019), and personalized search (Lu et al. 2020). A potential resource for building such PKBs is the personal attribute knowledge (e.g., hobbies and medical conditions) extracted from users' conversation data on a lot of platforms such as social media. To draw the personal attribute knowledge embedded in conversations, personal attribute prediction from conversations becomes an increasingly important task.

Given multiple users of a conversational system, a personal attribute, and utterances of these users, the task of personal attribute prediction from conversations aims to predict the ranking of the given personal attribute values for each user. It is noted that this task focuses on the case that the personal attribute values are not explicitly mentioned in utterances, and the given attribute values are ranked based on the underlying semantics of utterances, which is different from the common information extraction task. For example, we could rank the attribute values scientist and teacher high with regard to the profession attribute when the user mentions the words "research", "lab", "teaching" and "educator" in user utterances. However, this task is challenging due to the following aspects: (1) compared with formal documents, user utterances are often short, noisy, colloquial and have diverse topics, and the textual cues in utterances are too implicit to seize; (2) the construction of training data via manually annotating user utterances is time-consuming and labor-intensive; (3) for the personal attribute with too many attribute values (e.g., profession), its several attribute values (e.g., student and teacher) may be related and difficult to distinguish as they often co-occur with the same words in utterances.

Recently, some neural network based models have been explored for this task. These models resort to labeled utterances (Tigunova et al. 2019), external data (e.g., Wikipedia and Web pages) (Liu, Chen, and Shen 2022) or both (Tigunova et al. 2020) as resources of training data. However, there exist three issues in these previous works: (1) they rely on many resources of training data but these resources are not always available and expensive to fetch, which limits their adaptability to new domains or new data; (2) the attribute knowledge embedded in the unlabeled utterances is underutilized; (3) their performance over some difficult personal attributes (e.g., profession and hobby) is still unsatisfactory. Additionally, it is found that text classification methods (Mekala and Shang 2020; Wang, Mekala, and Shang 2021; Zhang et al. 2021) which are adept in mining the textual cues could be used to address this task directly. Unfortunately, they also fail to achieve good performance when

*Wei Shen is the corresponding author.

predicting those difficult personal attributes, which has been verified by our experiments.

Intuitively, the personal attribute knowledge involved in unlabeled utterances is abundant. For example, if the words "law", "legal", "court", and "constitution" frequently co-occur with each other in different utterances, there is a high probability that users who mention these words have the same profession (i.e., lawyer). This kind of word co-occurrence information is beneficial for predicting personal attributes. Moreover, it is found that different word pairs (biterms) constructed from utterances as well as two words belonging to the same biterm may be related to an attribute value in different degrees at the semantic level. This kind of biterm semantic information is attribute-oriented and is also vital to our task. Consequently, how to integrate these two categories of information becomes very crucial.

To tackle the above issues, we propose a novel framework PEARL to **P**redict p**E**rsonal **A**ttributes from conve**R**sations by leveraging the abundant personal attribute knowledge from utterances in a **L**ow-resource setting (without requiring any labeled utterances or external data). Our proposed framework PEARL is composed of a biterm semantic acquisition (BSA) module and an attribute knowledge integration (AKI) module. To capture the biterm semantic information, the BSA module derives the biterm set by searching words with high semantic relevance to the attribute value from utterances and yields the attribute-oriented biterm representation for each biterm based on the pre-trained language model (PLM). To integrate the biterm semantic information with the word co-occurrence information, the AKI module leverages the biterm-attribute value similarity score derived from the BSA module as the prior attribute knowledge to guide the biterm topic model (BTM)'s (Yan et al. 2013) Gibbs sampling process. To further promote the prediction performance, the AKI module can update the prior attribute knowledge based on the new sampling result and refine the Gibbs sampling process guided by the updated prior attribute knowledge in an iterative manner. After multiple iterations in such a way, PEARL can predict probable attribute values for each user by utilizing the final sampling result. Additionally, it is worth mentioning that our proposed framework can not only solve the personal attribute prediction task but also adapt to the more general weakly supervised text classification task.

Our contributions can be summarized as follows:

- We are the first to address personal attribute prediction from conversations under a low-resource setting which does not resort to any labeled utterances or external data.
- We propose a novel framework PEARL which fuses the biterm semantic information and the word co-occurrence information together via leveraging the updated prior attribute knowledge to refine the BTM's Gibbs sampling process in an iterative manner.
- Extensive experimental studies have been conducted for the task of personal attribute prediction from conversations over two data sets, and the task of weakly supervised text classification over one data set. The experimental results show that our framework surpasses all the baseline methods on both tasks.

Preliminary

Task Definition

A user ID of a conversational system is denoted by i ($1 \leq i \leq n$), and the utterances posted by this user can be concatenated as one utterance denoted by u_i. A personal attribute is denoted by c, and $\{c_1, c_2, ..., c_g\}$ is the set of personal attribute values of c. Formally, given user IDs $1, 2, ..., n$, a personal attribute c, and the corresponding utterances of these users, the task of low-resource personal attribute prediction from conversations is to predict the ranking of the personal attribute values (i.e., $c_1, c_2, ..., c_g$) for each user according to these unlabeled utterances without the need of any other resources (e.g., labeled utterances or external data).

Biterm Topic Model

Conventional topic models (e.g., PLSA (Hofmann 1999) and LDA (Blei, Ng, and Jordan 2003)) can reveal the latent topics by implicitly capturing the document-level word co-occurrence patterns, while biterm topic model (BTM) can learn the topics better by modeling the generation of word co-occurrence patterns directly. We introduce BTM here in brief. Given a biterm set B, which is constructed based on a document collection by extracting any two distinct words in a document as a biterm, the generation process of B in BTM can be described as follows:

- Draw a topic distribution $\theta \sim \text{Dir}(\alpha)$
- For each topic z
  (a) Draw a topic-word distribution $\psi_z \sim \text{Dir}(\beta)$
- For each biterm $b \in B$
  (a) Draw a topic $z \sim \text{Multi}(\theta)$
  (b) Draw two words to generate b: w_j,$w_k \sim \text{Multi}(\psi_z)$

where α and β are the Dirichlet priors. The parameters of BTM (i.e., ψ and θ) can be approximately inferred by the Gibbs sampling process. To perform Gibbs sampling, the key step is to calculate the conditional distribution for each biterm b as follows:

$$P(z|\mathbf{z}_{\neg b}, B, \alpha, \beta) \propto (n_z + \alpha) \cdot \frac{(n_{w_j|z} + \beta)(n_{w_k|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2} \quad (1)$$

where $\mathbf{z}_{\neg b}$ denotes the topic assignments for all biterms except b; $n_{w|z}$ denotes how many times the word w is assigned to the topic z; n_z denotes how many biterms are assigned to the topic z; M denotes the number of words in the document collection. Subsequently, the parameters ψ and θ can be calculated as follows:

$$\psi_{w|z} = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta} \quad (2)$$

$$\theta_z = \frac{n_z + \alpha}{|B| + J\alpha} \quad (3)$$

where J is the number of topics. Finally, the topic proportions of a document d can be inferred based on ψ and θ as follows:

$$P(z|d) = \sum_b P(z|b)P(b|d) \quad (4)$$

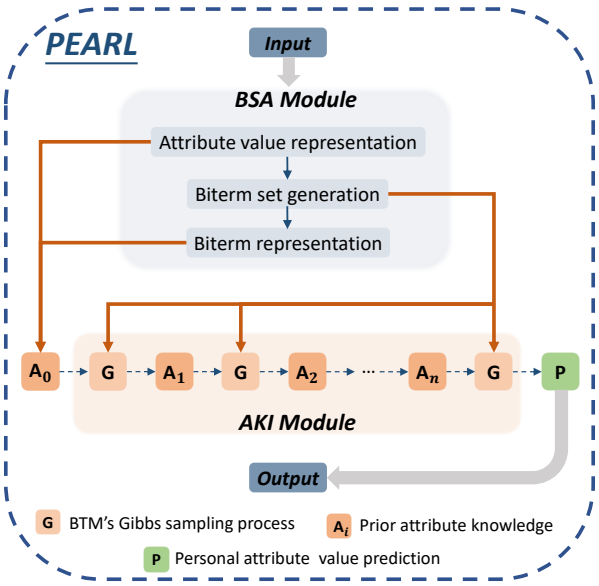


Figure 1: Overview of our framework PEARL.

$$P(z|b) = \frac{P(z)P(w_j|z)P(w_k|z)}{\sum_z P(z)P(w_j|z)P(w_k|z)} = \frac{\theta_z \psi_{w_j|z} \psi_{w_k|z}}{\sum_z \theta_z \psi_{w_j|z} \psi_{w_k|z}} \quad (5)$$

where $P(b|d)$ denotes the relative frequency of b in d. Interested readers please refer to BTM (Yan et al. 2013) for more details.

The Framework PEARL

The overall framework of our proposed PEARL is shown in Figure 1. We begin with the introduction of the BSA module and thereafter describe the AKI module.

Biterm Semantic Acquisition (BSA) Module

The BSA module consists of three parts: (1) attribute value representation (i.e., construct an attribute value representation for each attribute value); (2) biterm set generation (i.e., generate a biterm set for each utterance based on the attribute value representation); (3) biterm representation (i.e., yield an attribute-oriented biterm representation for each biterm based on the biterm set). We elaborate them as follows.

Attribute Value Representation. To understand the semantics of personal attribute values, we first define the static representation s_w for a word w by averaging contextualized representations of its all occurrences in utterances as follows:

$$s_w = \frac{\sum_{u_{i,j}=w} r_{i,j}}{N_w} \quad (6)$$

where $u_{i,j}$ denotes the j-th token of utterance u_i; $r_{i,j}$ denotes the contextualized token representation of $u_{i,j}$ based on a PLM; N_w denotes how many times w appears in all utterances. Thus, for an attribute value c_q which contains only one word, its static representation can be denoted by s_{c_q}. Specially, if c_q contains several words, we average the static representations of all the single words in c_q as s_{c_q} for simplicity.

To enhance the semantic understanding, inspired by recent weakly supervised text classification methods (Wang, Mekala, and Shang 2021; Mekala and Shang 2020), we propose to construct a word list for each attribute value to store some attribute-related words. For the attribute value c_q, we utilize its surface form to initialize its word list L_{c_q}, and subsequently discover its next attribute-related word owning the highest similarity score in utterances to expand its current word list iteratively. This similarity score can be calculated by the cosine similarity between the static representation of a word and the static representation of an attribute value. If the selected word is in the current word list of an attribute value, we will ignore it and continue to select another word for this iteration until the selected word does not occur in any attribute value's current word list. We will stop the iteration for the attribute value c_q if the overlap between the current word list L_{c_q} and the top $|L_{c_q}|$ similar words generated by the current attribute value representation is below a threshold η, which is a dynamic mechanism. We define the current attribute value representation v_{c_q} for the attribute value c_q as a weighted average representation based on the current word list L_{c_q} of c_q as follows:

$$v_{c_q} = \sum_{i=1}^{|L_{c_q}|} f(s_{L_{c_q,i}}, s_{c_q}) \cdot s_{L_{c_q,i}} \quad (7)$$

where $L_{c_q,i}$ denotes the i-th word in L_{c_q} ($1 \le i \le |L_{c_q}|$) and $s_{L_{c_q,i}}$ denotes the static representation of $L_{c_q,i}$ calculated by Formula 6; $f(s_{L_{c_q,i}}, s_{c_q})$ is the normalized cosine similarity score between $s_{L_{c_q,i}}$ and s_{c_q}. It is also worth mentioning that the final lengths of the word lists for different attribute values could be different as they are determined by the dynamic mechanism.

Biterm Set Generation. Intuitively, different words in an utterance have various levels of importance for predicting personal attributes. Specially, those words that are highly relevant to the attribute value are valuable for our task, so we need to find them out from each utterance to compose the biterm set. Inspired by the previous work (Xie, Girshick, and Farhadi 2016), to estimate the word importance, firstly we calculate the weight $\pi_{i,j}$ for each word $u_{i,j}$ based on the Student t-distribution (Van der Maaten and Hinton 2008) via using the obtained attribute value representation as follows:

$$\pi_{i,j} = \max_{c_q} Sim(u_{i,j}, c_q) \quad (8)$$

$$Sim(u_{i,j}, c_q) = \frac{(1 + \|r_{i,j} - v_{c_q}\|^2/\lambda)^{-\frac{\lambda+1}{2}}}{\sum_{c'_q}(1 + \|r_{i,j} - v_{c'_q}\|^2/\lambda)^{-\frac{\lambda+1}{2}}} \quad (9)$$

where $Sim(u_{i,j}, c_q)$ is the similarity between the word $u_{i,j}$ and the attribute value c_q; λ is the degree of freedom for the Student t-distribution. Subsequently, we aims to select top K words with high weights from each utterance u_i as the keywords, i.e., $u_{i,top_1}, u_{i,top_2}, \ldots, u_{i,top_h}, \ldots$, and u_{i,top_K}, where $1 \le h \le K$ and top_h is the ID of the token with the h-th highest weight in u_i. Thus we can construct a biterm set $\mathfrak{B}_i$ for u_i via the selected K keywords of u_i as follows:

$$\mathfrak{B}_i = \{b_{i,h,l} = (u_{i,top_h}, u_{i,top_l}) | 1 \le h, l \le K\} \quad (10)$$

where $b_{i,h,l}$ denotes a biterm composed of the words u_{i,top_h} and u_{i,top_l}. By concatenating the biterm sets $\mathfrak{B}_1, \mathfrak{B}_2, ..., \mathfrak{B}_n$ we can obtain the final biterm set $\mathfrak{B}$ over all utterances for our task.

Biterm Representation. Considering that the two words belonging to the same biterm have different relevance degrees to the attribute value at the semantic level, we propose to construct the attribute-oriented biterm representation for each biterm to capture the biterm semantic information. After obtaining the biterm set $\mathfrak{B}$, for each biterm $b_{i,h,l}$ in $\mathfrak{B}$, we define its representation $v_{b_{i,h,l}}$ as follows:

$$v_{b_{i,h,l}} = \pi_{i,top_h} r_{i,top_h} + \pi_{i,top_l} r_{i,top_l} \tag{11}$$

It is noted that for the utterance u_i, the weights of its selected K keywords should be normalized in advance.

Attribute Knowledge Integration (AKI) Module

Although biterm topic model (BTM) exploits the word co-occurrence information successfully, the biterm semantic information is ignored in BTM. To fuse these two categories of information, we try to inject the biterm semantic information into BTM via this AKI module. If we run BTM directly on the biterm set $\mathfrak{B}$ generated by the BSA module, i.e., iteratively calculate the conditional distribution $P(z|\mathbf{z}_{\neg b}, B, \alpha, \beta)$ for each biterm (Formula 1) and update the counting variables (i.e., n_z and $n_{w|z}$), each utterance can be assigned a topic by calculating the highest value of $P(z|u_i)$ (Formula 4) over all the topics. However, even if the number of topics is set to be the same as the number of attribute values, the correspondence between the topic and the attribute value is still lacking, which makes it impossible to be applied to our task of personal attribute prediction.

To remedy this issue, we propose an iterative biterm semantics (BS) based Gibbs sampling process in the AKI module. It is able to build the corresponding relationship between the attribute value and the topic via leveraging the prior attribute knowledge derived from the BSA module to guide the BTM's Gibbs sampling process. First, we associate each topic z with an individual attribute value c_q, and initialize states for the Markov chain randomly like BTM. Next, inspired by (Yang et al. 2020), we define the conditional distribution $P'(c_q|\mathbf{c}_{\neg b_{i,h,l}}, \mathfrak{B}, \alpha, \beta)$ for each biterm $b_{i,h,l}$ in the biterm set $\mathfrak{B}$ via combining the biterm-attribute value similarity score $\Omega(b_{i,h,l}, c_q)$ with the conditional distribution $P(c_q|\mathbf{c}_{\neg b_{i,h,l}}, \mathfrak{B}, \alpha, \beta)$ (Formula 1) as follows:

$$\begin{aligned} P'(c_q|\mathbf{c}_{\neg b_{i,h,l}}, \mathfrak{B}, \alpha, \beta) &= \Omega(b_{i,h,l}, c_q) \\ &\cdot P(c_q|\mathbf{c}_{\neg b_{i,h,l}}, \mathfrak{B}, \alpha, \beta) \end{aligned} \tag{12}$$

where $\mathbf{c}_{\neg b_{i,h,l}}$ denotes the the attribute value assignments for all biterms of $\mathfrak{B}$ except $b_{i,h,l}$. The biterm-attribute value similarity score could encode the prior attribute knowledge involved in utterances well, and we initialize it using the value of $cosine(v_{b_{i,h,l}}, v_{c_q})$, where $v_{b_{i,h,l}}$ and v_{c_q} denote the representations of the biterm $b_{i,h,l}$ and the attribute value c_q respectively, which are yielded by the BSA module. Following the original Gibbs sampling process, we iteratively calculate the conditional distribution for each biterm by Formula 12 and update the counting variables n_{c_q}, $n_{u_{i,top_h}|c_q}$, and $n_{u_{i,top_l}|c_q}$. Finally, the parameters ψ and θ of BTM can be calculated by Formulas 2 and 3 respectively. In this way, the AKI module enables BTM to induce attribute value-aware topics in the inference stage to integrate the biterm semantic information and the word co-occurrence information preliminarily. To further fuse these two categories of information, we propose a simple iteration operation by exploiting superior prior attribute knowledge to refine the Gibbs sampling process. Specifically, we update $\Omega(b_{i,h,l}, c_q)$ according to the parameters ψ in an iterative manner, which are the output of the Gibbs sampling process for each iteration as follows:

$$\Omega(b_{i,h,l}, c_q) = \psi_{u_{i,top_h}|c_q} \cdot \psi_{u_{i,top_l}|c_q} \tag{13}$$

It is noted that this iteration operation can promote the performance of our framework PEARL successfully, which has been verified in our experiments. In practice, it is found that after dozens of iterations, the performance is stable. This iterative procedure is summarized in Algorithm 1, which can automatically learn the parameters of our framework PEARL without requiring any training data. Note that in Algorithm 1, E is the number of iterations for the proposed iteration operation, while T is the number of iterations for the Gibbs sampling process.

After performing the iterative BS based Gibbs sampling process, the ranking of the attribute values can be obtained based on the learned parameters ψ and θ via calculating the probability score $P(c_q|u_i)$ (Formula 4) for each attribute value c_q.

Algorithm 1: Iterative BS based Gibbs sampling process

Input: The biterm set $\mathfrak{B}$, the biterm representation $v_{b_{i,h,l}}$ of the biterm $b_{i,h,l} \in \mathfrak{B}$, the attribute value representations v_{c_1}, v_{c_2}, ..., v_{c_g}, hyperparameters α and β, and the number of iterations E and T.
Output: The parameters ψ and θ.

```
Initialize attribute value assignments randomly for all the biterms in 𝔅
Initialize the biterm-attribute value similarity score Ω(b_{i,h,l}, c_q) by the value of cosine(v_{b_{i,h,l}}, v_{c_q})
for e = 1 to E do
    for t = 1 to T do
        for b_{i,h,l} ∈ 𝔅 do
            Draw an attribute value for b_{i,h,l} from the conditional distribution P'(c_q|c_{¬b_{i,h,l}}, 𝔅, α, β) by Formula 12
            Update n_{c_q}, n_{u_{i,top_h}|c_q}, and n_{u_{i,top_l}|c_q}
        end for
    end for
    Calculate ψ and θ by Formulas 2 and 3 respectively
    Update Ω(b_{i,h,l}, c_q) by Formula 13
end for
```

Experiments

Experimental Setting

Data Sets. For the task of personal attribute prediction from conversations, we perform experiments over two public data

Method type	*Labeled utterances*	*External data*		*Method*	*Profession*		*Hobby*	
		Wiki-page	*Wiki-category*		*MRR*	*nDCG*	*MRR*	*nDCG*
Personal attribute prediction methods	yes	✓	x	BERT IR	0.30	0.45	0.22	0.43
		x	✓		0.28	0.44	0.18	0.42
		✓	x	$\text{CHARM}_{\text{KNRM}}$	0.27	0.44	0.22	0.44
		x	✓		0.35	0.55	0.27	0.49
		✓	x	$\text{CHARM}_{\text{BM25}}$	0.29	0.46	0.24	0.47
		x	✓		0.28	0.47	0.21	0.43
	no	✓	x	No-keyword + BM25	0.15	0.32	0.16	0.42
		x	✓		0.17	0.37	0.13	0.35
		✓	x	RAKE + BM25	0.16	0.33	0.17	0.42
		x	✓		0.19	0.39	0.14	0.37
		✓	x	RAKE + KNRM	0.16	0.33	0.12	0.32
		x	✓		0.13	0.34	0.12	0.31
		✓	x	TextRank + BM25	0.21	0.39	0.21	0.46
		x	✓		0.26	0.45	0.20	0.42
		✓	x	TextRank + KNRM	0.21	0.38	0.15	0.36
		x	✓		0.18	0.36	0.16	0.36
		✓	x	HAM_{avg}	0.06	0.07	0.06	0.05
		x	✓		0.06	0.07	0.03	0.02
		✓	x	$\text{HAM}_{\text{2attn}}$	0.06	0.07	0.04	0.05
		x	✓		0.06	0.07	0.06	0.07
		✓	x	HAM_{CNN}	0.20	0.18	0.22	0.14
		x	✓		0.27	0.34	0.17	0.27
		✓	x	$\text{HAM}_{\text{CNN-attn}}$	0.21	0.28	0.13	0.10
		x	✓		0.25	0.31	0.16	0.25
		✓	x	DSCGN	0.43	0.57	0.29	0.50
		x	✓		0.44	0.60	0.29	0.49
Weakly supervised text classification methods	no	x	x	SeedBTM	0.33	0.55	0.17	0.42
		x	x	ConWea	0.07	0.26	0.04	0.21
		x	x	LOTClass	0.07	0.26	0.04	0.2
		x	x	X-Class	0.34	0.57	0.23	0.47
		x	x	ClassKG	0.07	0.24	0.04	0.19
Our method	no	x	x	PEARL	**0.49**	**0.64**	**0.31**	**0.54**

Table 1: Performance on the task of personal attribute prediction from conversations. All the results of the personal attribute prediction baselines are taken from DSCGN (Liu, Chen, and Shen 2022). The performance of all the weakly supervised text classification methods is reproduced via their open-source solutions.

sets: (1) profession data set; (2) hobby data set. These two data sets are extracted from publicly-available Reddit submissions and comments (2006 - 2018), and are annotated and provided by the authors of (Tigunova et al. 2020). All utterances containing explicit personal attribute assertions used for annotation have been removed. The given attribute values for each personal attribute are defined based on Wikipedia lists. The number of attribute values for profession and hobby are 71 and 149 respectively. Both data sets consist of about 6000 users and have a maximum of 500 and an average of 23 users for each personal attribute value.

Evaluation Measures. Following the previous personal attribute prediction studies (Tigunova et al. 2019, 2020; Liu, Chen, and Shen 2022), we adopt the same ranking metrics MRR (Mean Reciprocal Rank) and nDCG (normalized Discounted Cumulative Gain) to evaluate all the methods.

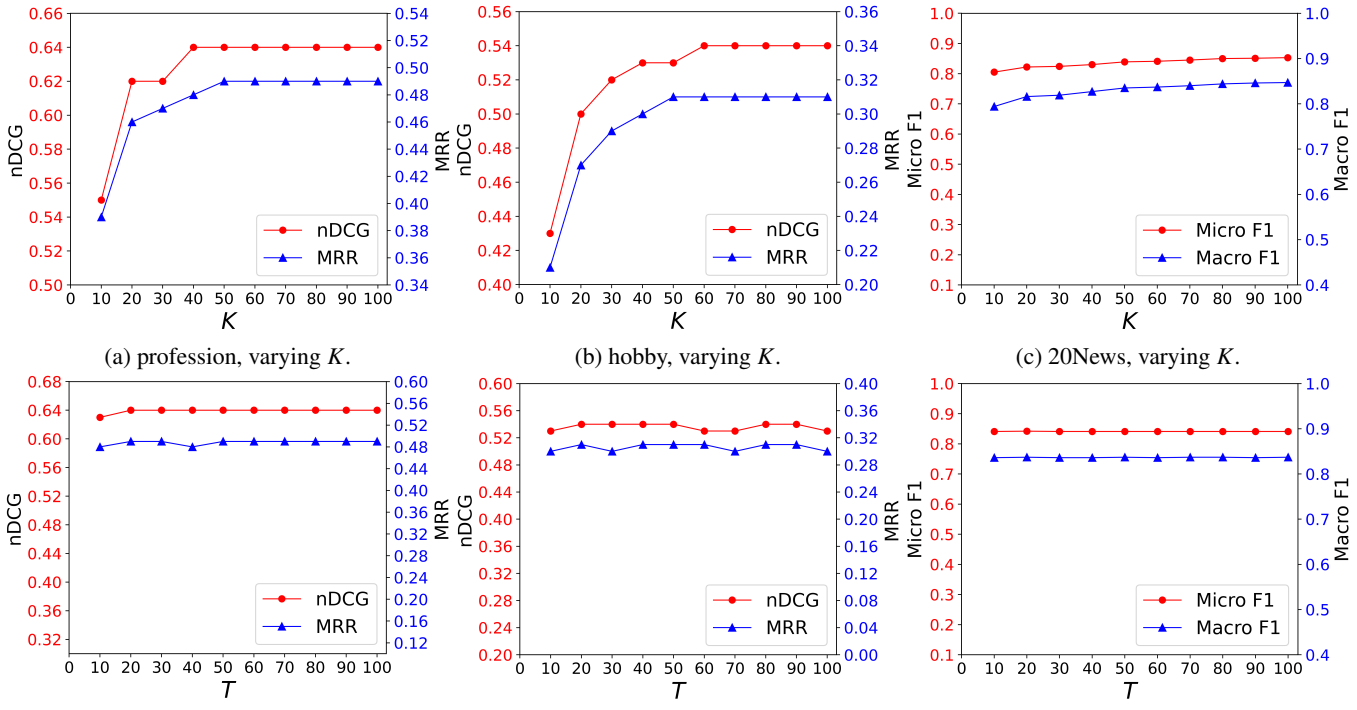
Setting Details. The threshold η, the Dirichlet prior β, the number of keywords K for each utterance, the degree of freedom λ, the numbers of iterations E and T are set to 75%, 0.01, 60, 1, 20 and 50 respectively. The Dirichlet prior α is set to $50/g$, where g is the number of attribute values. BERT base-uncased model is adopted as the PLM. The number of attribute-related words for each attribute value is set to a minimum of 10 and a maximum of 40. The experiments are implemented by MindSpore Framework[1]. The source code and data sets used in this paper are publicly available[2].

Effectiveness Study

We compare our proposed framework with the following personal attribute prediction methods. BERT IR (Dai and Callan 2019) trains BERT to calculate the relevance between an utterance u_i and a document denoted by d_{ext} from external data (e.g., Wiki-page or Wiki-category) w.r.t. an attribute value based on a binary cross-entropy loss. To fit the input size of BERT, u_i and d_{ext} are both split into 256-token chunks. CHARM (Tigunova et al. 2020) extracts keywords from u_i by a cue detector and retrieves

[1]https://www.mindspore.cn/en

[2]https://github.com/CodingPerson/PEARL



(a) profession, varying K. (b) hobby, varying K. (c) 20News, varying K.

(d) profession, varying T. (e) hobby, varying T. (f) 20News, varying T.

Figure 2: Parameter study over the profession, hobby, and 20News data sets.

the relevant documents from external data by a value ranker. $\text{CHARM}_{\text{BM25}}$ (resp. $\text{CHARM}_{\text{KNRM}}$) adopts BM25 (Robertson et al. 1995) (resp. KNRM (Xiong et al. 2017)) as the value ranker of CHARM. No-keyword + BM25 uses u_i and d_{ext} directly as the input of BM25. RAKE (TextRank) + BM25 (KNRM) utilizes keywords extracted from u_i by the unsupervised keyword extraction approach RAKE (Rose et al. 2010) (TextRank (Mihalcea and Tarau 2004)) and d_{ext} as the input of BM25 (KNRM). HAM (Tigunova et al. 2019) can predict the score of the attribute value for u_i by training the neural network based on the external data. HAM_{avg}, $\text{HAM}_{\text{2attn}}$, HAM_{CNN}, and $\text{HAM}_{\text{CNN-attn}}$ are four different configurations of HAM. HAM_{CNN}/$\text{HAM}_{\text{CNN-attn}}$ (resp. HAM_{avg}/$\text{HAM}_{\text{2attn}}$) adopts a text classification CNN (resp. two stacked fully connected layers). $\text{HAM}_{\text{CNN-attn}}$/$\text{HAM}_{\text{2attn}}$ (resp. HAM_{avg}/HAM_{CNN}) adopts attention mechanisms (resp. average methods) within and across utterances. DSCGN (Liu, Chen, and Shen 2022) fine-tunes BERT over unlabeled utterances and external data to predict an attribute value score list for u_i.

Additionally, we add some recent SOTA weakly supervised text classification methods as baselines, which can be used to train an utterance classifier to predict the probability of attribute values for each utterance u_i. Specifically, ConWea (Mekala and Shang 2020) can utilize user-provided seed words to create a contextualized utterance corpus, which is further leveraged to train an utterance classifier and expand seed words iteratively. SeedBTM (Yang et al. 2020) could utilize user-provided seed words to extend BTM into an utterance classifier based on the word embedding technique. LOTClass (Meng et al. 2020) generates some attribute-indicative words for each attribute value to fine-tune a PLM on a word-level category prediction task, and then does self-training on unlabeled utterances. X-Class (Wang, Mekala, and Shang 2021) can learn attribute-oriented utterance representations by a PLM and use the utterance-attribute value pairs generated by the Gaussian mixture model clustering process to train an utterance classifier. ClassKG (Zhang et al. 2021) can generate pseudo labels for utterances by annotating keyword subgraphs, and train an utterance classifier with the pseudo labels.

Apart from external data (e.g., Wiki-page or Wiki-category), the personal attribute prediction baselines BERT IR, $\text{CHARM}_{\text{BM25}}$, and $\text{CHARM}_{\text{KNRM}}$ require the labeled utterances to train the model and execute ten-fold cross-validation on each data set under a zero-shot setting in which the attribute values in the training set and the testing set are disjoint. Other personal attribute prediction baselines directly perform on the unlabeled utterances and external data without the requirement of labeled utterances, which is a relatively difficult setting, as it pushes "zero-shot" to the extreme – no labeled utterances for any attribute values are provided. Compared with all the personal attribute prediction baselines, our framework PEARL and all the weakly supervised text classification methods perform on the unlabeled utterances only, which is an extremely difficult low-resource setting, as no other resources are provided except for the unlabeled utterances. Additionally, it is noted that some weakly supervised text classification methods require user-provided seed words, for each attribute value, we use

Ablations	*Profession*		*Hobby*	
	MRR	*nDCG*	*MRR*	*nDCG*
PEARL	**0.49**	**0.64**	**0.31**	**0.54**
w/o AKI	0.41	0.60	0.23	0.45
w/o the iteration operation	0.45	0.62	0.29	0.52

Table 2: Ablation study on personal attribute prediction from conversations.

its surface form as its seed word.

From the results in Table 1, it can be seen that although consuming minimal resources, our proposed framework PEARL still yields the best performance compared with eighteen baselines on both data sets. To be specific, there are five weakly supervised text classification baselines with the same low-resource setting as us. Nevertheless, our method significantly surpasses them, which may be due to the fact that our framework can mine the personal attribute knowledge embedded in unlabeled utterances better. Despite that all the personal attribute prediction methods leverage more resources (e.g., labeled utterances or external data) than PEARL, PEARL still promotes by at least 2 (resp. 4) percentages compared with the best personal attribute prediction baseline DSCGN in terms of MRR (resp. nDCG) over both data sets. All the above experimental results validate the superiority of our proposed framework for predicting personal attributes from conversations.

Parameter Study

To investigate the robustness of our framework, we conduct sensitivity analysis to understand the impact of the parameter K (i.e., the number of keywords for each utterance) and T (i.e., the number of iterations for the Gibbs sampling process in Algorithm 1) on our PEARL's performance over the profession and hobby data sets. From the trend plotted in Figure 2a and Figure 2b, we can see that when $K > 50$ (resp. $K > 60$), the performance achieved by PEARL is very stable on the profession (resp. hobby) data set, and is insensitive to the parameter K. From the trend plotted in Figure 2d and Figure 2e, it can be seen that the performance of PEARL is relatively stable on both data sets, and is insensitive to the parameter T.

Ablation Study

To verify the importance of different parts in our framework PEARL, we first remove the AKI module to make the BSA module work alone and perform attribute prediction by Formula 4 directly (i.e., $P(b|d)$ is set to 1). Considering that the AKI module alone cannot predict personal attributes from conversations, which has been discussed in the AKI part, so we cannot provide the effect of PEARL removing the BSA module. In addition, to examine the effectiveness of the iteration operation, we run PEARL by executing the BSA module first and thereafter executing the AKI module without the iteration operation (i.e., E in Algorithm 1 is set to 1).

From the results in Table 2, we can draw the following observations: (1) without the AKI module, the performance of PEARL declines significantly on both data sets, which demonstrates the importance of the AKI module. This may be attributed to the fact that the biterm semantic information and the word co-occurrence information are complementary to some extent, and our framework PEARL can harness this complimentary knowledge effectively for better prediction; (2) without the iteration operation, PEARL performs worse over both data sets, which validates the point that the iteration operation can indeed obtain superior prior attribute knowledge derived from the AKI module to refine the Gibbs sampling process and enhance the prediction performance.

Method	*20News*	
	Micro-F1	*Macro-F1*
SeedBTM	44.9	37.3
ConWea	75.7	73.3
LOTClass	73.8	72.5
X-Class	78.6	77.8
ClassKG	83.8	82.7
PEARL	**84.1**	**83.7**
w/o AKI	68.5	64.3
w/o the iteration operation	82.7	81.6

Table 3: Performance on the task of weakly supervised text classification. The performance of the baselines SeedBTM and ClassKG is reproduced via their open-source solutions. The results of the remaining baselines are taken from X-Class (Wang, Mekala, and Shang 2021).

Experimental Analysis on Weakly Supervised Text Classification Task

To adapt our framework to the weakly supervised text classification task, we replace the attribute value (resp. utterance) with the class (resp. text) and select the class with the highest probability as the predicted label for the text based on the output ranking of PEARL. Following the previous weakly supervised text classification studies (Meng et al. 2020; Wang, Mekala, and Shang 2021), we adopt the same evaluation metrics, i.e., micro-F1 and macro-F1, to evaluate PEARL, PEARL's two truncated versions, and five SOTA weakly supervised text classification methods over the common benchmark data set 20News (Lang 1995), which consists of 17817 texts. Different from the profession (resp. hobby) data set containing 71 (resp. 149) attribute values as labels, 20News only has a fairly small number of five topics as labels (i.e., computer, sports, science, politics, and religion).

From the results in Table 3, it can be seen that (1) our framework PEARL exceeds all weakly supervised text classification methods, exhibiting its superiority and universality for the task of weakly supervised text classification; (2) PEARL outperforms its two truncated versions, which also demonstrates the effectiveness of the AKI module and the iteration operation for the text classification task.

To explore the robustness of our framework on the weakly supervised text classification task, we further conduct sensitivity analysis to understand the impact of the parameters K (i.e., the number of keywords for each utterance) and T (i.e.,

the number of iterations for the Gibbs sampling process in Algorithm 1) on our PEARL's performance over the 20News data set. From the trend plotted in Figure 2c, we can see that when $K > 50$, the performance of PEARL changes little and is insensitive to the parameter K. From the trend plotted in Figure 2f, it can be seen that PEARL performs relatively stable, and is insensitive to the parameter T.

Related Work

Personal Attribute Prediction

The task of personal attribute prediction usually contains two aspects of studies: (1) personal attribute prediction from conversations; (2) personal attribute prediction from social media. We elaborate them as follows.

Previous works extracted the profession attribute from conversations to build a PKB for a user by maximum-entropy classifiers (Jing, Kambhatla, and Roukos 2007) and sequence-tagging CRFs (Li et al. 2014). These methods assumed that users explicitly mentioned the attribute value in their utterances, so they are not applicable for the profession and hobby data sets without personal assertions. Recently, HAM (Tigunova et al. 2019) predicted scores of different attribute values for an utterance based on the stacked fully connected layers or CNNs by utilizing average approaches or attention mechanisms within and across utterances. CHARM (Tigunova et al. 2020) first extracted some keywords from utterances by leveraging BERT, and then matched these keywords against Web documents indicating possible attribute values from external data via some SOTA information retrieval ranking models (i.e., RAKE and KNRM). Specially, the procedure of the keyword extraction was trained by a reinforce policy gradient method. DSCGN (Liu, Chen, and Shen 2022) yielded two categories of supervision, i.e., document-level supervision via a distant supervision strategy and contextualized word-level supervision via a label guessing method from unlabeled utterances and external data, to fine-tune the language model with a noise-robust loss function. Different from all the above methods which consume plenty of resources, our framework does not rely on any labeled utterances or external data.

Additionally, numerous works aim to predict personal attributes (e.g., age (Bayot and Gonçalves 2017; Mac Kim et al. 2017; Liu and Singh 2021), gender (Bayot and Gonçalves 2017; Mac Kim et al. 2017; Vijayaraghavan, Vosoughi, and Roy 2017; Basile et al. 2017), location (Shen, Liu, and Wang 2018; Liu et al. 2021), political preference (Vijayaraghavan, Vosoughi, and Roy 2017; Preoţiuc-Pietro et al. 2017; Xiao et al. 2020), ethnicity (Preoţiuc-Pietro and Ungar 2018), and occupational class (Preoţiuc-Pietro, Lampos, and Aletras 2015)) from social media such as Facebook and Twitter. The results provided by (Tigunova et al. 2019) show that three of these works (Basile et al. 2017; Bayot and Gonçalves 2017; Preoţiuc-Pietro, Lampos, and Aletras 2015) obtain unsatisfactory performance when predicting personal attributes from conversations, so they are not selected as our baselines. The remaining works rely on rich meta-data of social media (e.g., user profile, hashtag, and social network structure) that do not exist in conversation data, so they are unsuitable for our task.

Weakly Supervised Text Classification

In recent years, many methods have been proposed to use the label surface name as the weakly supervised signal to solve the text classification task without requiring any labeled documents or external data, whose setting is consistent with our proposed low-resource setting. Therefore, these methods can be utilized to predict personal attributes by training an utterance classifier based on unlabeled utterances. These weakly supervised text classification approaches (Meng et al. 2020; Mekala and Shang 2020; Wang, Mekala, and Shang 2021; Zhang et al. 2021) usually trained a classifier by using the pseudo labels and refined the model over the unlabeled data via self-training. However, they perform not well on the profession and hobby data sets, which has been verified in our experiments.

Conclusion and Future Work

To predict personal attributes from conversations under a low-resource setting which does not resort to any labeled utterances or external data, we propose a novel framework PEARL that combines the biterm semantic information with the word co-occurrence information seamlessly in an iterative manner. Extensive experiments have demonstrated the effectiveness of our framework PEARL against many SOTA personal attribute prediction methods and weakly supervised text classification methods. In addition, we argue that the profession and hobby data sets can be utilized to measure the efficacy of weakly supervised text classification methods to a certain extent, which can benefit the text classification research community for future studies.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. U1936206, 62272247), YESS by CAST (No. 2019QNRC001), and CAAI-Huawei MindSpore Open Fund.

References

Balog, K.; and Kenter, T. 2019. Personal knowledge graphs: A research agenda. In *SIGIR*, 217–220.

Balog, K.; Radlinski, F.; and Arakelyan, S. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *SIGIR*, 265–274.

Basile, A.; Dwyer, G.; Medvedeva, M.; Rawee, J.; Haagsma, H.; and Nissim, M. 2017. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In *CEUR Workshop Proceedings*, volume 1866.

Bayot, R. K.; and Gonçalves, T. 2017. Age and gender classification of tweets using convolutional neural networks. In *International Workshop on Machine Learning, Optimization, and Big Data*, 337–348. Springer.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Dai, Z.; and Callan, J. 2019. Deeper text understanding for IR with contextual neural language modeling. In *SIGIR*, 985–988.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *AAAI*, 5110–5117.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

Jing, H.; Kambhatla, N.; and Roukos, S. 2007. Extracting social networks and biographical facts from conversational speech transcripts. In *ACL*, 1040–1047.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, 331–339. Elsevier.

Li, X.; Tur, G.; Hakkani-Tür, D.; and Li, Q. 2014. Personal knowledge graph population from user utterances in conversational understanding. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, 224–229. IEEE.

Liu, Y.; Chen, H.; and Shen, W. 2022. Personal Attribute Prediction from Conversations. In *WWW*, 223–227.

Liu, Y.; Shen, W.; Yao, Z.; Wang, J.; Yang, Z.; and Yuan, X. 2021. Named entity location prediction combining twitter and web. *IEEE Transactions on Knowledge and Data Engineering*, 33(11): 3618–3633.

Liu, Y.; and Singh, L. 2021. Age Inference Using A Hierarchical Attention Neural Network. In *CIKM*, 3273–3277.

Lu, S.; Dou, Z.; Xiong, C.; Wang, X.; and Wen, J.-R. 2020. Knowledge Enhanced Personalized Search. In *SIGIR*, 709–718.

Luo, L.; Huang, W.; Zeng, Q.; Nie, Z.; and Sun, X. 2019. Learning personalized end-to-end goal-oriented dialog. In *AAAI*, volume 33, 6794–6801.

Mac Kim, S.; Xu, Q.; Qu, L.; Wan, S.; and Paris, C. 2017. Demographic inference on twitter using recursive neural networks. In *ACL*, 471–477.

Mekala, D.; and Shang, J. 2020. Contextualized weak supervision for text classification. In *ACL*, 323–333.

Meng, Y.; Zhang, Y.; Huang, J.; Xiong, C.; Ji, H.; Zhang, C.; and Han, J. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In *EMNLP*, 9006–9017.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *EMNLP*, 404–411.

Preoţiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *ACL*, 1754–1764.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond binary labels: political ideology prediction of twitter users. In *ACL*, 729–740.

Preoţiuc-Pietro, D.; and Ungar, L. 2018. User-level race and ethnicity predictors from twitter text. In *COLING*, 1534–1545.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.; and Gatford, M. 1995. Okapi at TREC-3. *NIST special publication*, 109–123.

Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1: 1–20.

Shen, W.; Liu, Y.; and Wang, J. 2018. Predicting named entity location using Twitter. In *ICDE*, 161–172.

Tigunova, A.; Yates, A.; Mirza, P.; and Weikum, G. 2019. Listening between the lines: Learning personal attributes from conversations. In *WWW*, 1818–1828.

Tigunova, A.; Yates, A.; Mirza, P.; and Weikum, G. 2020. CHARM: Inferring Personal Attributes from Conversations. In *EMNLP*, 5391–5404.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vijayaraghavan, P.; Vosoughi, S.; and Roy, D. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In *ACL*, 478–483.

Wang, Z.; Mekala, D.; and Shang, J. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *NAACL*, 3043–3053.

Xiao, Z.; Song, W.; Xu, H.; Ren, Z.; and Sun, Y. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *KDD*, 2258–2268.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, 478–487. PMLR.

Xiong, C.; Dai, Z.; Callan, J.; Liu, Z.; and Power, R. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *SIGIR*, 55–64.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *WWW*, 1445–1456.

Yang, Y.; Wang, H.; Zhu, J.; Wu, Y.; Jiang, K.; Guo, W.; and Shi, W. 2020. Dataless short text classification based on biterm topic model and word embeddings. In *IJCAI*, 3969–3975.

Yen, A.-Z.; Huang, H.-H.; and Chen, H.-H. 2019. Personal Knowledge Base Construction from Text-based Lifelogs. In *SIGIR*, 185–194.

Zhang, L.; Ding, J.; Xu, Y.; Liu, Y.; and Zhou, S. 2021. Weakly-supervised Text Classification Based on Keyword Graph. In *EMNLP*, 2803–2813.