Personal Attribute Prediction from Conversations

Yinan Liu, Hu Chen, Wei Shen∗
TKLNDST, College of Computer Science, Nankai University, Tianjin, 300350, China
{liuyn,2120210473}@mail.nankai.edu.cn, shenwei@nankai.edu.cn

ABSTRACT
Personal knowledge bases (PKBs) are critical to many applications, such as Web-based chatbots and personalized recommendation. Conversations containing rich personal knowledge can be regarded as a main source to populate the PKB. Given a user, a user attribute, and user utterances from a conversational system, we aim to predict the personal attribute value for the user, which is helpful for the enrichment of PKBs. However, there are three issues existing in previous studies: (1) manually labeled utterances are required for model training; (2) personal attribute knowledge embedded in both utterances and external resources is underutilized; (3) the performance on predicting some difficult personal attributes is unsatisfactory. In this paper, we propose a framework DSCGN based on the pre-trained language model with a noise-robust loss function to predict personal attributes from conversations without requiring any labeled utterances. We yield two categories of supervision, i.e., document-level supervision via a distant supervision strategy and contextualized word-level supervision via a label guessing method, by mining the personal attribute knowledge embedded in both unlabeled utterances and external resources to fine-tune the language model. Extensive experiments over two real-world data sets (i.e., a profession data set and a hobby data set) show our framework obtains the best performance compared with all the twelve baselines in terms of nDCG and MRR.

1 INTRODUCTION

Personal knowledge bases (PKBs) are critical to many applications, such as Web-based chatbots and personalized recommendation. Conversational systems are becoming very important as many personal assistants emerge (e.g., Siri and Cortana). Understanding user utterances plays an essential role in holding meaningful conversations with users. In reality, the PKB is beneficial for understanding user utterances better. Meantime, conversation data can also be regarded as a main source to populate a PKB by drawing personal attribute knowledge (e.g., hobbies, professions, and visited cities) from the users’ conversations in many platforms (e.g., social media and dialogues). To enrich the PKB using personal knowledge from conversations, predicting personal attributes from conversations is an important task that needs to be solved urgently.

1 INTRODUCTION

在一个重要的任务中，需要解决几个人口统计学问题: (1) 培训数据中的人口统计学问题的建设; (2) 人口统计学知识嵌入在双方语词和外部资源中（例如，Wikipedia）是不充分利用的; (3) 在预测一些难以解决的个人属性时性能不理想。在本篇纸张中，我们提出一个基于非监督的框架 DSCGN，它不需要任何标注的语句。我们定义两种监督：(1) 文档级监督: 通过假设语句来预测个人属性; (2) 句子级监督: 通过为文档中的每个实体提供一个属性标签来预测个人属性。我们在此基础上进行实验，结果表明我们的框架在两个真实世界的数据集上（即，一个专业数据集和一个爱好数据集）优于所有12个基线方法，在nDCG和MRR上取得最佳性能。
we could mine the attribute knowledge embedded in Wikipedia to help predict personal attributes from conversations. To derive high-quality labeled Wikipedia data, for each personal attribute value, we select its corresponding Wiki-page by string matching based on the surface forms of the attribute value (e.g., Dentist) and the Wiki-page title (e.g., Wiki: Dentist). To obtain more labeled Wikipedia data, we construct Wiki-category by identifying at least one relevant category for each attribute value and including all leaf pages under the category like [16]. In this way, the document-level supervision $S_{doc} = \{(d_i, A_{d_i})\}$ could be derived automatically, where $d_i$ denotes a Web document from external resources (e.g., the Wiki-page or a page in Wiki-category) and $A_{d_i}$ denotes the corresponding attribute value of $d_i$.

2.2 Label Guessing Method

When provided an attribute value, humans can understand its semantics based on general knowledge by associating it with some attribute-indicative words. To create contextualized word-level attribute value supervision (contextualized word-level supervision) automatically, inspired by [9], we propose a label guessing method to find attribute-indicative words via the PLM pre-trained with the masked language model (MLM) task on large-scale text corpora.

First, for the $i$-th occurrence $w_i$ of the word $w$ in unlabeled utterances and external resources (e.g., Wiki-page and Wiki-category) provided by the distant supervision strategy, we feed its contextualized embedding $h_{w_i}$ produced by PLM encoder to the MLM head, which will output a probability distribution over the entire vocabulary $V$, indicating the likelihood of each word $v \in V$ appearing at this position.

$$p(v|h_{w_i}) = \text{Softmax}(W_2\varphi(W_1h_{w_i} + b)),$$

where $\varphi(\cdot)$ is the activation function; $W_1 \in \mathbb{R}^{h \times h}$, $W_2 \in \mathbb{R}^{|V| \times h}$, and $b \in \mathbb{R}^h$ are learnable parameters pre-trained with the MLM objective of PLM. Intuitively, words that are interchangeable most of the time are likely to have similar meanings. We retain top-K (sorted by $p(v|h_{w_i})$) similar words as the candidate word set $C_{w_i}$ for the word occurrence $w_i$. Thus, we can generate $K$ valid replacement words for each occurrence of the attribute value. Subsequently, for an attribute value $a$, we form the attribute value vocabulary $V_a$ using the top-$M$ words ranked by frequency in all candidate word sets of its occurrences ($C_{a_1}, C_{a_2}, \ldots, C_{a_n}$), discarding stopwords and words that appear in many attribute value vocabularies. Then, we define the semantic similarity $\text{Sim}_{w2a}(w_i, a)$ between word $w_i$ and attribute value $a$ as follows.

$$\text{Sim}_{w2a}(w_i, a) = \frac{|C_{w_i} \cap V_a|}{|V_a|}.$$

If this similarity is larger than a threshold $\beta$, $w_i$ is regarded as an attribute-indicative word with regard to the attribute value $a$. Thereby, from unlabeled utterances and external resources, we obtain contextualized word-level supervision $S_{word} = \{(w_i, A_{w_i})\}$, where $w_i$ denotes an attribute-indicative word and $A_{w_i}$ denotes the corresponding attribute value of $w_i$.

2.3 Confidence based PLM Fine-tuning

A PLM can be fine-tuned by leveraging task-specific training data so that it can achieve a better result when performing that task. In
this paper, to leverage the personal attribute knowledge embedded in both unlabeled utterances and external resources for predicting personal attributes from conversations, we try to fine-tune the PLM based on the document-level supervision $S_{doc}$ yielded by the distant supervision strategy (Section 2.1) and contextualized word-level supervision $S_{word}$ derived by the label guessing method (Section 2.2).

Specifically, considering the balance between model convergence and noise-robustness, and in order to give more attention to the true attribute-indicative words and relieve the effect of the mislabeled words, we propose a confidence based loss function $L_{CGCE}$ based on the generalized cross entropy (GCE) loss [19] by incorporating the confidence based weight function $\xi(t)$ which is calculated by the semantic similarity $Sim_{w2w}(\cdot)$ as follows:

$$L_{CGCE} = - \sum_{t \in S_{doc} \cup S_{word}} \xi(t) \frac{1 - \text{Softmax}(W_{A_t} h_t + b_{A_t})^q}{q}$$

$$\xi(t) = \begin{cases} Sim_{w2w}(t, A_t) & \text{if } t \text{ refers to an attribute-indicative word from } S_{word}, \\ 1 & \text{if } t \text{ refers to a Web document from } S_{doc}. \end{cases}$$

where $W_{A_t} \in \mathbb{R}^{K \times h}$ and $b_{A_t} \in \mathbb{R}^K$ are learnable parameters of the linear layer ($K$ is the number of attribute values); $0 < q < 1$ is a hyperparameter; If $t$ refers to an attribute-indicative word from $S_{word}$, $h_t$ denotes its contextualized embedding, otherwise if $t$ refers to a Web document from $S_{doc}$, $h_t$ denotes its CLS embedding.

After fine-tuning, given a user and a user attribute, we combine all utterances of a user as a sentence and give this sentence as the input of the fine-tuned PLM to obtain the probabilities of different attribute values of this attribute.

3 EXPERIMENTAL STUDY

3.1 Experimental Setting

Data Sets. Two real-world data sets (i.e., a profession data set and a hobby data set) are used in this experiment. They are annotated and publicly provided by [16], and extracted from publicly-available Reddit submissions and comments between 2006 and 2018. They both consist of about 6000 users, with a maximum of 500 and an average of 23 users per attribute value. The number of attribute values for profession (hobby) is 71 (149) respectively. Specially, for each Reddit user, all posts containing explicit personal assertions used for labeling have been removed.

Evaluation Measures. We adopt the same evaluation metrics MRR (Mean Reciprocal Rank) and nDCG (normalized Discounted Cumulative Gain) as previous works [15, 16]. The nDCG score is calculated by averaging all users. Considering that a user may have many attribute values, we follow the assumption that there is only one correct attribute value for each user. Thus, MRR can be calculated independently for each attribute value before averaging.

Baseline Models. No-keyword + BM25 [16] utilizes a user’s full utterances and a Web document from $S_{doc}$ w.r.t. an attribute value as the input of BM25 [12]. BM25 is a strong unsupervised retrieval model to calculate the similarity between two documents. RAKE (TextRank) + BM25 (KNRM) [16] uses keywords extracted from a user’s full utterances by RAKE (TextRank) and a Web document from $S_{doc}$ w.r.t. an attribute value as the input of BM25 (KNRM). RAKE [13] and TextRank [10] are two SOTA unsupervised keyword extraction methods. KNRM [17] is an efficient neural retrieval model, which considers semantic similarity via term embeddings. BERT IR [16] splits both user utterances and a Web document from $S_{doc}$ into 256-token chunks respectively to fit the input size of BERT, and predicts the relevance between them based on a binary cross-entropy loss. HAMavg (HAMCNN, HAMCNN-attn, and HAMattn) [15] can predict the score of different attribute values by training the neural network based on $S_{doc}$-HAMCNN-attn and HAMattn both utilize attention mechanisms within and across documents from $S_{doc}$-HAMavg and HAMattn both adopt two stacked fully connected layers, while HAMCNN and HAMCNN-attn both adopt text classification CNN. CHARMM$_{BM25}$ (CHARMM$_{KNRM}$) [16] uses keywords extracted from a user’s full utterances by BERT and a Web document from $S_{doc}$ as the input of BM25 (KNRM).

Setting Details. We utilize the BERT base-uncased model as the PLM. The training batch size and the learning rate are set to 4 and 1e-5 respectively. The parameters $K$, $M$, and $\beta$ of the label guessing method are set to 50, 100, and 0.4 respectively. The hyperparameter $q$ is set to 0.4. It is noted that the adopted lists of personal attribute values and the result for each baseline except HAMavg, HAMattn, HAMCNN, and HAMCNN-attn are taken from [16] directly. To represent utterance embedding, the baselines [15] use word2vec embeddings pre-trained on the Google News corpus via Mindspore Framework1. Source code used in this paper is publicly available2.

3.2 Experimental Results

Besides external resources, the baselines BERT IR, CHARMM$_{BM25}$, and CHARMM$_{KNRM}$ require extra labeled utterances for model training, and perform ten fold cross-validation on the profession (hobby) data set under the zero-shot setting in which the attribute values in the training and test data are disjoint. Our method DSCGN and other baselines do not require labeled utterances as input and perform on the profession (hobby) data set directly, which is a more

### Table 1: Performance on predicting personal attributes from conversations using the external resource (Wiki-page or Wiki-category).

| Labeled utterances | Method               | Wiki-page | Wiki-category |
|--------------------|----------------------|-----------|--------------|
|                    | MRR  | nDCG  | MRR  | nDCG  | MRR  | nDCG  | MRR  | nDCG  |
| no                 | No-keyword + BM25   | 15.19  | 16.42 | 17.33 | 13.33 | 14.39 | 14.37 |
|                   | RAKE + BM25         | 16.33  | 17.42 | 19.39 | 14.37 |
|                   | RAKE + KNRM         | 16.33  | 12.32 | 13.34 | 12.31 |
|                   | TextRank + BM25     | 21.31  | 21.46 | 26.45 | 20.42 |
|                   | TextRank + KNRM     | 21.31  | 15.36 | 18.36 | 16.36 |
|                   | HAMavg              | 66.06  | 66.05 | 66.07 | 66.07 |
|                   | HAMattn             | 66.07  | 66.04 | 66.07 | 66.07 |
|                   | HAMCNN              | 20.18  | 22.14 | 27.34 | 17.27 |
|                   | HAMCNN-attn         | 21.28  | 13.39 | 25.31 | 16.25 |
|                   | HAMCNN-KNRM         | 43.57  | 59.29 | 44.60 | 29.49 |
| yes                | BERT IR             | 30.40  | 22.43 | 28.44 | 18.42 |
|                   | CHARMM$_{BM25}$     | 29.46  | 24.47 | 28.47 | 21.43 |
|                   | CHARMM$_{KNRM}$     | 27.44  | 22.44 | 35.55 | 27.49 |

1https://www.mindspore.cn/en
2https://github.com/CodingPerson/DSCGN
Table 2: Ablation study.

| Ablations                                      | profession | hobby |
|-----------------------------------------------|------------|-------|
|                                                | MRR | JDCG | MRR | JDCG |
| DSCGN                                         | 0.44 | 0.60 | 0.29 | 0.50 |
| DSCGN w/o confidence based weight function (ξ(·)) | 0.42 | 0.58 | 0.27 | 0.48 |
| DSCGN w/o label guessing method               | 0.38 | 0.54 | 0.27 | 0.48 |
| DSCGN w/o distant supervision strategy        | 0.06 | 0.23 | 0.04 | 0.20 |

difficult setting, since it pushes “zero-shot” to the extreme — no labeled utterances for any attribute values are given.

Effectiveness Study. From the results in Table 1, it can be seen that whichever external resource (i.e., Wiki-page or Wiki-category) is used, our proposed method DSCGN obtains the best performance compared with the twelve baselines on both data sets in terms of MRR and nDCG. To be specific, there are nine baselines that just leverage the external resources as distant supervision like us. Nevertheless, our method significantly outperforms them, which shows that our method can leverage the personal attribute knowledge embedded in external resources better. In addition to the external resources, the baselines CHARMBAU25, CHARMMKRM, and BERT IR utilize extra labeled utterances for model training, which puts DSCGN at a disadvantage. However, in spite of this disadvantage, we found that our framework still promotes by at least about 9 (5) percentages compared with the best baseline in terms of MRR (nDCG) over the profession data set, and obtains the best performance in terms of MRR (nDCG) over the hobby data set, indicating the superiority of our proposed framework for the task of predicting personal attributes from conversations.

Ablation Study. To show the different parts of our framework (i.e., confidence based weight function (ξ(·)), label guessing method, and distant supervision strategy) are crucial for our task, we conduct an ablation study by removing each part of the proposed framework DSCGN and testing the remaining parts over the profession (hobby) data set using Wiki-category (Wiki-page) as the external resource. From the experiment results shown in Table 2, compared with these truncated versions of DSCGN, the full version of DSCGN obtains the best performance, which demonstrates each part of our framework is helpful for our task. To be specific, removing the label guessing method or the distant supervision strategy, the performance of DSCGN degrades a lot, indicating the personal attribute knowledge embedded in utterances and external resources are complementary.

4 RELATED WORK

Personal Attribute Prediction from Conversations. Previous works [4, 5] relied on the condition that user attributes are explicitly mentioned in the utterances, so these methods are not suitable for the profession (hobby) data set without personal assertions used in this paper. Recently, HAM [15] proposed some models based on neural network architectures (e.g., stacked fully connected layers and CNNs), to predict a score of different attribute values (e.g., different professions) for a given subject-predicate pair by using average methods or attention mechanisms within and across utterances. CHARM [16] first leveraged the pre-trained LM BERT to extract keywords related to the attribute from user utterances, and then adopted some SOTA ranking models based on information retrieval techniques (i.e., RAKE and KNRM) to match these keywords against documents that indicate possible values of the attribute. Specially, they proposed a reinforce policy gradient method to train the process of keywords extraction.

Personal Attribute Prediction from Social Media. In recent years, numerous methods have been proposed to predict attributes (e.g., age [3, 7], gender [3, 8], and occupational class [11]) of users from social media (e.g., Twitter and Facebook). The vast majority of these works rely on rich meta-data of social media such as hashtag, user profile description, and social network structure. This makes them unsuitable for our task, as conversations do not have these rich meta-data. In addition, there are only three works [2, 3, 11] that predict user latent attributes based solely on user-generated text. However, from the results provided by [15], we found that these three methods perform not well on the task of predicting personal attributes from conversations, so we did not select them as our baselines in this paper. Additionally, some recent works explored the task of named entity location prediction from Twitter [6, 14], but their methods are not applicable for personal attributes.

5 CONCLUSION

To tackle three issues existing in previous studies (i.e., labeled utterances shortage, personal attribute knowledge underutilization, and unsatisfactory performance), we propose a flexible PLM based framework DSCGN without requiring any labeled utterances. Two categories of supervision (i.e., document-level supervision and contextualized word-level supervision) are utilized by DSCGN to fine-tune the language model via a noise-robust loss function. To demonstrate the effectiveness of DSCGN, we conduct extensive experiments over two real-world data sets, which show whichever external resource (i.e., Wiki-page or Wiki-category) is used, our framework obtains the best performance compared with all the twelve baselines in terms of nDCG and MRR.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (No. U1936206), Natural Science Foundation of Tianjin (No. 19JQCNJ00100), YESS by CAST (No. 2019QNRC001), and CAAI-Huawei MindSpore Open Fund.

REFERENCES

[1] Kristian Balog and Tom Kersten. 2019. Personal knowledge graphs: A research agenda. In SIGIR. 217–220.
[2] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GRAM: New Groningen Author-profiling Model—Notebook for PAN. In CEUR Workshop Proceedings, Vol. 1866.
[3] Roy Christopher Bayot and Teresa Gonçalves. 2017. Age and gender classification of tweets using convolutional neural networks. In International Workshop on Machine Learning, Optimization, and Big Data. Springer, 337–348.
[4] Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2007. Extracting social networks and biographical facts from conversational speech transcripts. In ACL. 1040–1047.
[5] Xiang Li, Gokhan Tur, Dirle Hakkani-Tür, and Qi Li. 2014. Personal knowledge graph population from user utterances in conversational understanding. In 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, 224–229.
[6] Yinan Liu, Wei Shen, Zonghai Yao, Jianyong Wang, Zhenglu Yang, and Xiaoqie Yuan. 2021. Named entity location prediction combining twitter and web. IEEE Transactions on Knowledge and Data Engineering 33, 11 (2021), 3618–3633.
[7] Yaguang Liao and Lisa Singh. 2021. Age Inference Using A Hierarchical Attention Neural Network. In CIKM. 3273–3277.
