Deep Learning based Privacy Information Identification approach for Unstructured Text

Yichen Ning¹,²,a, Na Wang¹,²,*, Aodi Liu¹,², Xuehui du¹,²

¹Information Engineering University, Zhengzhou
²He'nan Province Key Laboratory of Information Security
*aemail: Ning_yi_chen@163.com
*email: twftina_w@126.com

Abstract. Data sharing sometimes brings the privacy disclosure risk. Anonymization methods such as k-anonymity, l-diversity prevent privacy disclosure, but such methods are suitable for structured text. There are a lot of unstructured texts in people's lives (such as social network texts, clinical texts), and identifying and structuring the private information (PI) of unstructured texts is a problem. Based on this, we propose a deep learning-based unstructured text PI identification approach, which can extract PI in unstructured text, associate the PI with the corresponding subject, and organize it into structured data, to support follow-up anonymization. This approach is divided into two tasks: PI identification and PI association. We respectively propose a sequence labeling model based on the RoBERTa-BiLSTM-CRF hybrid neural network and a PI association method based on the RoBERTa-HCR hybrid neural network to identify PI and organize it into structured data. The experimental results show that, compared with the benchmark model, RoBERTa-BiLSTM-CRF has better performance; compared with the current Chinese coreference resolution model, the average F1-score value of RoBERTa-HCR is increased by 6%.

1. Introduction
In the current era of highly developed data science, data is deeply coupled with people's lives. On the one hand, we enjoy the convenience brought by applications such as recommendation algorithms, map apps, and health codes. On the other hand, we suffer from personal privacy disclosure caused by data abuse and illegal sharing. For example, on November 9, 2020, there was a case of a person infected with the new crown virus in Shanghai. In less than a day, the detailed information and family information of this person was exposed on the Weibo. This is a serious privacy breach. When data is exchanged and shared, PI should be properly protected.

In order to promote the safe and reasonable use of data, the European Union proposed the General Data Protection Regulation [1] in 2016. It stipulates two legal ways to deal with private information: obtaining the consent of the data owner, or anonymizing the personal PI in the data. Generally speaking, because anonymization will cause data distortion and reduce data availability, obtaining the consent of the data owner should be the preferred way. However, in many cases, especially when using third-party data, obtaining the consent of the data owner will be a rather complicated process. As a result, anonymization is a maneuverable privacy preserving method. Anonymization methods such as k-anonymity, l-diversity, t-closeness preserve personal privacy while ensuring certain availability for structured data. In practice, a large amount of data is based on unstructured texts, which contain PI.
Therefore, in order to achieve privacy preserving, before using k-anonymity, l-diversity and other methods for anonymization, we need to first extract PI from the unstructured text, and associate PI with the subject to construct the structured personal privacy data. However, the current word granularity methods can only extract PI in the text, but do not associate it with the subject.

Based on this, we propose deep learning based unstructured text privacy identification (DLUI) approach. This approach can extract PI from unstructured text, associate PI with the corresponding subject, and then organize it into structured data, which provides effective support for follow-up anonymization. DLUI decomposes the problem of PI identification into two problems: PI extraction and PI association. Aiming at the problem of PI extraction, we propose a sequence labeling model based on the hybrid neural network model RoBERTa-BiLSTM-CRF, which implements the extraction of six types of PI of name, location, time, organization, product, and company in unstructured text. Compared with the benchmark algorithm, the F1-score value of this model is increased by 1.27%. Aiming at the problem of PI association, we propose a PI association method based on RoBERTa-HCR. The method first uses the RoBERTa-HCR coreference resolution model to implement the link between the pronoun and the subject in the unstructured text, and then work with the privacy link method based on semantic role labeling (SRL) to complete the PI association task. Compared with the current Chinese coreference resolution model, average F1 of RoBERTa-HCR increased by 6%.

2. Related work

PI identification method can be divided into two categories:

2.1. Word granularity

In order to identify a certain category of PI such as location, name, academia has proposed some privacy identification methods based on sequence labelling model. In order to use a huge amount of clinical text data without revealing the PI of doctors and patients, Zhe et al. [2] proposed a cascaded approach based on Conditional Random Field (CRF). This method uses regular expression matching and filtering content that may contain PI, then constructs a dataset through manual labelling and enter context, history of present Illness and past medical history feature into CRF network to identify PI. This method needs to manually extract text features, and only extract PI, without establishing an association between PI and the subject. Mehta et al. proposed a CRF-based PI identification and anonymization method [3], which increases the number of input features to improves the performance of the sequence labelling model. This method first identifies PI based on the CRF, and then uses scalable k-anonymity, but it does not take into account that there will be multiple subjects in the text. If it is impossible to determine which PI belongs to which subject, it will cause difficulties in the subsequent use of k-anonymity.

In response to the problem that different people have different definitions of PI, Sánchez et al. proposed a flexible user-definable sensitive word identification method [4]. This method calculates the similarity between user-defined sensitive words and span in the text. The information content(IC) and the Point-wise Mutual Information(PMI) in the search engine determine whether the span is PI. The method of Hassan et al. [5] inputs the user-defined sensitive words and the text span into the word2vec neural network model, calculates the cosine similarity of the output result, and extracts PI based on the similarity. However, this method can’t consider long distance context information because of word2vec. Both of these methods have the disadvantages of not being able to effectively identify PI of similar address and not being able to identify a specific type of PI.

2.2. Sentence, document granularity

Neerbekey et al.[6] found that a sentence without sensitive words may still contain PI, and proposed a sentence granularity PI identification method based on the RNN neural network model, which uses the grammar tree features and the RNN network to calculate whether a sentence is rephrase of sensitive sentences to determine whether the sentence contains PI.

The text granularity PI identification method is mainly used to label the sensitivity level of text. Tesfay et al.[7] proposed a method which extracts text tf-idf features, uses a binary classification model
to determine whether the text is sensitive text, and then uses multi-classifier to classify sensitive text into different categories. Xu et al. further proposed a method based on Text-CNN [8], which converts the problem of PI identification into a text classification problem, and uses the Text-CNN neural network model to achieve text sensitivity level identification.

We focus on the problem of identifying PI at word granularity. The research has found that current word granularity PI identification method is based solely on sequence labelling model and cannot associate the subject with its PI. It is hard to directly use anonymization methods such as k-anonymity, l-diversity in the future.

3. The proposed approach
According to the definition of PI in the EU's General Data Protection Regulation (EU-GDPR): "personal data means any information relating to an identified or identifiable natural person ('data subject')" [1], we define the following types of information as PI: The subject’s name, ID number, telephone number, and location, time, organization, product, and company related to the subject.

DLUI includes two parts: PI extraction and PI association (shown in Figure 1). In the PI extraction, we propose RoBERTa-BiLSTM-CRF sequence labeling model to extract six kinds of PI: name, location, time, organization, product, and company. ID number and telephone number is extracted by pattern matching method based on regular expression because of its structural specificity. In the PI association, we propose a PI association method based on RoBERTa-HCR. Firstly, the RoBERTa-HCR model is used to link pronouns and subjects in unstructured text. Secondly, SRL based privacy link method is used to establish the relationship between the extracted PI and the corresponding subjects, and finally organize them into structured data.

Figure1. The framework of DLUI

3.1. RoBERTa-BiLSTM-CRF sequence labelling model
We transform the six PI extraction problem into a sequence labeling problem, and proposes a sequence labeling model based on the RoBERTa-BiLSTM-CRF hybrid neural network model. The model consists of the RoBERTa layer, BiLSTM layer and CRF layer. First, obtains the embedding of the characters in
the sentence, and then input these features to the BiLSTM layer to extract more location features, and finally passes them to the CRF network to learn the differences between different characters. Label dependency, and finally get the label corresponding to each character, the model structure is shown in Figure 3.

3.1.1 preprocess
Before the text is input into the model, the text is pre-processed first, including the following steps:

1) Text cleaning. The Chinese text used in this article contains mixed Chinese and English symbols, noise symbols such as: \t, \n, etc. This article uses python's regex module to replace English punctuation with Chinese, and replace noise symbols with spaces.

2) BERT pre-processing. Add [CLS][SEP] symbols at the beginning and end of each sentence, calculate the position_embedding value for each character of each sentence, add padding for multi-batch training, and use attention_mask to eliminate the influence of these padding.

3) Change training data format. Label the sequence according to the label type. We use the "BIOS" method of labeling: "B" represents the beginning of the label, "I" represents the middle part of the label, "O" represents the text outside the label, and "S" marks the label with a length of one. Example show in Figure2.

```
刘/B-PER 冠/I-PER 威/I-PER 在/O 向/B-ORG 里/I-ORG 巴/I-ORG 巴/I-ORG 公/O 司/O 上/O 班/O。
```

3.1.2 model structure
1) RoBERTa layer
The RoBERTa layer is the input layer of the model. This layer encodes the characters in the document into vector form and inputs it to the BiLSTM layer. In this layer, we use the improved model RoBERTa-wwm-ext of the pre-training model BERT\cite{9} proposed by Google. This model combines the advantages of BERT-wwm\cite{10} and RoBERTa\cite{11}: in the pre-training stage, a whole world mask (wwm) strategy is used and RoBERTa training parameters are combined to obtain a more stable model. Because some researchers found that not using the Next Sentence Prediction (NSP) task in BERT pre-training will achieve better results in many downstream tasks, the NSP task is cancelled\cite{12}.

2) BiLSTM layer

Although the BERT model can get word embedding containing contextual information, the RoBERTa model uses a self-attention-based transformer, which weakens the position information in the calculation (only relying on position embedding). Character order is very important in sequence labeling. Long Short-Term Memory (LSTM) use historical information to effectively extract position information, and prevent the vanishing gradient. Bidirectional Long Short-Term Memory (BiLSTM) is a two-way LSTM that can extract position information more effectively. Therefore, we use Bi-LSTM to enhance the position information features of characters in sentences.

3) CRF layer

The label of the current character is dependent on the label of the surrounding characters. For example, B-PER label can only be followed by I-PER label, and S-PER label cannot be followed by I-PER label. CRF calculates the conditional probability of each character corresponding to a label by calculating a transition matrix representing the transfer score of adjacent labels and an emission matrix representing label score corresponding to each character. Adding the CRF layer after the Bi-LSTM layer make label transfer more accurate.

3.2. RoBERTa-HCR based private information association method

In this section, the identified PI is associated with the corresponding subject, and the association problem can be transformed into a problem of judging the dependency between the subject and PI. There is a SRL tool that can solve. At the same time, because pure SRL will link pronouns with PI, it is necessary to link pronouns with the subject. It’s a coreference resolution problem. This section proposes a RoBERTa-HCR based private information association method. This method first uses the RoBERTa-HCR coreference resolution model to link pronouns with the subject, and second uses the privacy link method based on SRL to associate private information with corresponding subject, and finally organize it into structured data.

3.2.1. model description

In order to resolve the Chinese text coreference resolution task, we propose RoBERTa-HCR model based on the high-order coreference resolution (HCR) \cite{13}. The model also uses the RoBERTa-wwm-ext pre-training model to map text sentences to a more representative space. In the PI association, the distance between the two associated span may be very far, and the antecedent information in BERT is not enough. Therefore, in this case, the use of HCR can explicitly obtain more antecedent information in span representation.

HCR hybrid neural network model key parameters are calculated as follows:

The goal of the task is to learn the n-th iterator’s probability distribution shown in (1).

$$P_n(\hat{y}_i) = \frac{e^{s(g^n_i, g^n_{\hat{y}_i})}}{\sum_{y' \in \gamma(i)} e^{s(g^n_{y'}, g^n_{\hat{y}_j})}}$$ (1)

$\hat{y}_i$ represents the antecedent of the i-th span, $\gamma(i)=\{\varepsilon, l, \ldots, m-1\}$ is the set of all possible antecedents of the i-th span, $m$ is the total number of spans in the text. $\varepsilon$ representing dummy antecedents, when the i-th span has no antecedents or it is not mention, there are only dummy antecedents in the set.
The initial span vector and co-reference score are calculated as (2),(3).

$$s(g^n_i, g^n_j) = s_m(g^n_i) + s_m(g^n_j) + s_a(g^n_i, g^n_j)$$

$$g^0 = [x_{start(i)}, x_{end(i)}, \hat{x}_i, l(i)]$$

$x$ represents the output of the Roberta layer, $l(i)$ encodes the length feature of span, the calculation formula of $\hat{x}_i$ is shown as (4),(5),(6).

$$c_i = FN_i(x_i)$$

$$c_{it} = \frac{\exp(c_i)}{\sum_{t'=start}^{end} \exp(c_{t'})}$$

$$\hat{x}_i = \sum_{t=\text{start}}^{\text{end}} c_{i,t} x_i$$

mention score and score of antecedent $j$ of span $i$ in the $n$-th iteration are shown as (7),(8)

$$s_m(g^n_i) = \omega_m FN_m(g^n_i)$$

$$s_a(g^n_i, g^n_j) = \omega_a FN_a([g^n_i, g^n_j, g^n_i \circ g^n_j, r(i, j)])$$

$r(i, j)$ encodes the length feature of span. Before calculate span vector of next iteration, calculate antecedent probability distribution of this iteration $P_n(y_i)$, and then use attention mechanism and gated sum mechanism to obtain the next iteration span vector $g^{n+1}_i$. The calculate formula are shown as (9),(10),(11).

$$a^n_i = \sum_{y_i \in Y(i)} P_i(y_i) \cdot g^n_{y_i}$$

$$f^n_i = S\left(FN_f \left[g^n_i, a^n_i\right]\right)$$

$$g^{n+1}_i = f^n_i \cdot g^n_i + \left(1 - f^n_i\right) \cdot a^n_i$$

According to the research [13], the second-order is the best.

3.2.2. SRL based privacy link method

The task of SRL is centered on the predicate of the sentence, studying the relationship between the components of the sentence and the predicate, and describing the relationship between them with semantic roles.

SRL based privacy link method first uses the SRL tool to label sensitive sentences, links the span identified by RoBERTa-HCR marked with Arg0 or Arg1 with the PI appearing in Arg1, Arg0, ArgM-TMP and ArgM-LOC, and then links the spans belonging to the same subject. We use the SRL tool in language technology platform (LTP) developed by the social computing and information retrieval research center of Harbin Institute of technology.

4. Experiment

4.1. dataset and environment

In order to evaluate the performance of the method proposed in this paper, a large number of high-quality Chinese texts annotated with name, location, time, organization, product and company are required. After comparing CLUENER, BOSON, People's Daily 1996 2014, Weibo datasets, etc., we finally chose the BOSON dataset with high data annotation quality and complete entity annotations. In the evaluation of coreference resolution model, we use the Chinese part of ontonotes 5.0 dataset [14].
The hardware and software environment of the experiment is as follows: operating system is ubuntu 18.04LTS, CPU is Intel® Xeon(R) Platinum 8168 CPU @ 2.70GHz × 9, GPU is TITAN V, memory size is 256GB, Pytorch version is 1.7.0, python version is 3.7.

4.2. sequence labelling model performance evaluation
We use precision, recall and f1 score experimental evaluation metric. This experiment selects the BERT-BiLSTM-CRF model as the benchmark model. The experimental hyper parameters are set as follows: the maximum length of the training input text is 128; after the BERT input layer, add a dropout layer with a rate of 0.1; the Bi-LSTM hidden layer dimension is 200 dimensions; a layer of feedforward is added between the BERT layer and the CRF layer Neural network, batch size is set to 24, epoch is set to 60; using adam optimizer training, the learning rate of the BERT layer is set to 3e-05, the learning rate of the Bi-LSTM layer is 1e-03, and the learning rate of the CRF layer is 1e-03. Result is shown in Table 1.

| label type | RoBERTa-BiLSTM-CRF (our method) | BERT-BiLSTM-CRF |
|------------|---------------------------------|-----------------|
| Pre        | Recall                          | F1              | Pre        | Recall                          | F1              |
| company    | 72.48%                          | 71.17%          | 71.82%     | 75.94%                          | 63.96%          | 69.44%          |
| location   | 84.60%                          | 86.35%          | 86.97%     | 81.82%                          | 86.15%          | 83.93%          |
| organization | 82.70%                         | 81.29%          | 81.99%     | 77.85%                          | 76.53%          | 77.19%          |
| name       | 94.18%                          | 94.37%          | 94.27%     | 95.12%                          | 94.16%          | 94.64%          |
| product    | 70.43%                          | 77.04%          | 73.58%     | 71.43%                          | 75.31%          | 73.32%          |
| time       | 82.49%                          | 81.71%          | 82.10%     | 82.08%                          | 82.66%          | 82.37%          |
| overall    | 82.84%                          | 83.52%          | 83.18%     | 81.81%                          | 82.02%          | 81.91%          |

It can be seen that except name and time, our model is better than the benchmark model in other items including the overall performance, and the total f1-score is 1.27 higher %. This is because the RoBERTa model uses longer text training in the pre-training phase, mask whole word and removes the NSP task, so it has stronger representation ability than the ordinary BERT model, and thus obtains a better training effect.

4.3. co-reference resolution model performance evaluation
This experiment compares the existing Chinese coreference resolution model[15] as a benchmark model. In this experiment, MUC, B3, CEAFφ4 and their average F1-score are used as evaluation metric. The hyperparameters of this experiment are set as follows: the maximum input length of the training text is 512, batch size is 6, and the epoch is 100; after 10 epochs, the f1 value does not increase, stop training; use the adam optimizer to optimize, and BERT learning rate is 2e-5, the learning rate of other parameters is 3e-4. Table 2 shows the comparison results between RoBERTa-HCR and benchmark.
It can be seen that the proposed method is superior to the benchmark model method in MUC, B3, ceaf φ 4 and the average F1 score of the three, and the average F1 score is 6% higher. Because this model explicitly increases the antecedent information in span representation, and uses RoBERT which has strong representation ability.

| Method       | MUC  | B3  | CEAFφ4 | CONLL |
|--------------|------|-----|--------|-------|
| Pre          | Recall | F1  | Pre    | Recall | F1  | Pre    | Recall | F1  | Avg.F1 |
| benchmark    | 74.8  | 64.1| 69.0   | 67.9   | 54.2 | 60.2   | 63.2   | 53.1 | 57.7   | 62.3 |
| Our model    | 76.8  | 71.9| 74.3   | 68.7   | 63.9 | 66.2   | 68.0   | 63.0 | 65.4   | 68.6 |

Table 2. RoBERTa-HCR performance comparison
4.4. sample
In this section, we display the sample of DULI output. The text we use is as follows (translate to English).

This morning, Wang Wei and I went to the legendary No. 5 Yonghua road to play wolf killing. When I got the script of the murderer, I split up. The killing script has two full pages about my killing process. I killed five people, including myself. That's ridiculous. In the afternoon, I went shopping with him in Nanping Street. He bought a Sony earphone of 5000 yuan. I’m envied. In the evening, I went to the new global bar with him. In the middle of the night, my dormitory closed, so I went to Nanjiang Hotel to have a sleep. Today I’m really hi.

The output structured data is shown in Table3. We find it well identified and well structured.

| name        | time          | location            | time, location         | product                     |
|-------------|---------------|---------------------|------------------------|-----------------------------|
| subject 1   | ***           | Nanjiang Hotel      | No. 5 Yonghua road, this morning; Nanping Street, afternoon; global bar, evening | Wolf killing, Sony earphone |
| subject 2   | ***           | ***                 | No. 5 Yonghua road, this morning; Nanping Street, afternoon; global bar, evening | Wolf killing                |

5. Conclusion
We focus on PI identification, and proposes DULI, which organizes unstructured private information into structured form. DULI divides private information identification into two core tasks: private information extraction and private information association. We propose a sequence labeling model based on Roberta-BiLSTM-CRF to solve private information extraction problem firstly. Then, RoBERTa-HCR based private information association method is proposed, which associates the PI with the subject to generate structured data. Experimental results verify the effectiveness of the proposed method.

In future work, we try to propose solutions for the special situation of pronoun missing in some Chinese texts.

Acknowledgment
This work is supported by National Natural Science Foundation of China (No. 61802436)

References
[1] Europe Union, 2016. General Data Protection Regulation. European Union Regulation 2016/679. https://gdpr-info.eu
[2] Jian Z, Guo X, Liu S, et al. (2017) A Cascaded Approach for Chinese Clinical Text De-Identification with Less Annotation Effort. Journal of Biomedical Informatics, 73:76-83.
[3] Mehta B, Rao U P, Gupta R, et al. (2019) Towards privacy preserving unstructured big data publishing. Journal of Intelligent & Fuzzy Systems, 36(4):3471-3482.
[4] Sánchez, David, Batet M. (2017) Toward sensitive document release with privacy guarantees. Engineering Applications of Artificial Intelligence, 59:23-34.
[5] Hassan F, Sanchez D, Soria-Comas J, et al. (2019) Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings. In: 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications. Rotorua, New Zealand. pp. 358-365
[6] Neerbeky J, Assentz I, Dolog P. (2017) TABOO: Detecting Unstructured Sensitive Information Using Recursive Neural Networks. In: 33rd International Conference on Data Engineering (ICDE). San Diego, CA, USA. pp. 1399-1400.
[7] Tesfay W B, Serna J, Rannenberg K. (2019) PrivacyBot: Detecting Privacy Sensitive Information in Unstructured Texts. In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). Granada, Spain. pp. 53-60.
[8] Xu G, Qi C, Yu H, et al. (2019) Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network. In: 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). Guilin, China. pp. 474-479.

[9] Devlin J, Chang M, Lee K, et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: North American chapter of the association for computational linguistics. Minneapolis, Minnesota. pp. 4171-4186.

[10] Cui Y, Che W, et al. (2019) Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101.

[11] Liu Y, Ott M, Goyal N, et al. (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[12] Joshi M, Chen D, Liu Y, et al. (2020) Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8: 64-77.

[13] Lee K, He L, Zettlemoyer L. (2018) Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana. pp. 687-692.

[14] Weischedel R, Palmer M, Marcus M, et al. OntoNotes Release 5.0[J].

[15] Jian F, Kong F. (2020) Coreference Resolution Incorporating Structural Information. COMPUTER SCIENCE,3: 231-236.