A Fine-grained Chinese Software Privacy Policy Dataset for Sequence Labeling and Regulation Compliant Identification

Kaifa Zhao[1], **Le Yu**[1], **Shiyao Zhou**[1], **Jing Li**[1], **Xiapu Luo**[1],
Yat Fei Aemon Chiu[2], **Yutong Liu**[2]
[1]Department of Computing, The Hong Kong Polytechnic University, HKSAR, China
[2]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, HKSAR, China
[1]kaifa.zhao@connect.polyu.hk, lele08.yu@polyu.edu.hk, shiyao.zhou@connect.polyu.hk
[1]{jing-amelia.li, daniel.xiapu.luo}@polyu.edu.hk
[2]{yat-fei-dylan.zhao,yitang.liu}@connect.polyu.hk

Abstract

Privacy protection raises great attention on both legal levels and user awareness. To protect user privacy, countries enact laws and regulations requiring software privacy policies to regulate their behavior. However, privacy policies are written in natural languages with many legal terms and software jargon that prevent users from understanding and even reading them. It is desirable to use NLP techniques to analyze privacy policies for helping users understand them. Furthermore, existing datasets ignore law requirements and are limited to English. In this paper, we construct the first Chinese privacy policy dataset, namely `CA4P-483`, to facilitate the sequence labeling tasks and regulation compliance identification between privacy policies and software. Our dataset includes 483 Chinese Android application privacy policies, over 11K sentences, and 52K fine-grained annotations. We evaluate families of robust and representative baseline models on our dataset. Based on baseline performance, we provide findings and potential research directions on our dataset. Finally, we investigate the potential applications of `CA4P-483`[1] combing regulation requirements and program analysis.

1 Introduction

A privacy policy is a legal document written in natural language that discloses how and why a controller collects, shares, uses, and stores user data (GDPR, 2016; PISS, 2020; NISSTC, 2020). Privacy policies help users understand whether their privacy will be abused and decide whether to use the product. However, privacy policies are tedious, making it hard for users to read and understand them (Staff, 2011). Natural language processing techniques achieve big success on understanding document semantics (Yang et al., 2021; Wen et al., 2021; Ding et al., 2020). Thus, it is necessary to apply natural language processing to analyze the privacy policies (Yu et al., 2016; Andow et al., 2020; Yu et al., 2015; Fan et al., 2020) and help users be aware of apps' privacy access behavior (Zhou et al., 2021).

Chinese **s**oftware **p**rivacy **p**olicy **p**rocessing (`CSP`^{3}) task is a sequence labeling problem that recognizes privacy-related components in the sentences. `CSP`^{3} has two main unique features. First, privacy policies contain an amount of information inside (Yu et al., 2018), such as how the app stores user data, and how to contact app developer. In our dataset, we concentrate on data access-related sentences as the sentences directly related to user privacy. Second, privacy policies are written in a legally binding professional language and contain software jargon. Thus, it requires strong background (Zhou et al., 2022a,b) to understand the statements inside. Both characteristics prevent users from understanding the privacy policies. A well-annotated dataset can facilitate building automatic privacy policy analysis tool and further help users protect their privacy.

Although privacy policy datasets have been proposed recently (Wilson et al., 2016; Zimmeck et al., 2019), labels in existing datasets are coarse-grained (i.e., sentence-level annotations (Wilson et al., 2016)) and limited to few privacy practices (Zimmeck et al., 2019). Besides, existing datasets only include English privacy policies, which limits the application of these datasets in regions with other languages. We construct a fine-grained Chinese dataset for software privacy policy analysis.

In this work, we focus on Android application privacy policies as Android possesses the largest share of mobile operating systems (statcounter, 2022) and a large number of Android privacy data leaks have been revealed (Shrivastava and Kumar, 2021; Sivan et al., 2019). Unlike previous work (Wilson et al., 2016; Zimmeck et al., 2019), we deal with the problem using sequence label-

[1]Our dataset and code are publicly available in https://github.com/zacharykzhao/CA4P-483

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10266 - 10277
December 7-11, 2022 ©2022 Association for Computational Linguistics

ing methods, and pay special attention to Chinese privacy policies. The motivations come from the following four aspects:

First, worldwide regulation departments enact laws (NISSTC, 2020; PISS, 2020; GDPR, 2016; CCPA, 2016; CLPRC, 2016) to regulate the software's behaviors and protect users' privacy. The laws require the software to clarify how and why they need to access user data. Analyzing privacy policies can help users understand how app process their data and identify whether apps comply with laws. Second, for sequence labeling tasks, CSP^3 aims to identify how and why the software collects, shares, and manages users' data according to regulations. CSP^3 can be abstracted as identifying components in the privacy policy documents, such as data type and the purpose of using user data. NLP techniques can help automatically analyze privacy policies. Third, existing privacy policy analysis research is limited to English and totally omits other languages. With over 98.38 billion app downloads (Statista, 2022) and privacy-related regulations enacted in China, it is necessary and urgent to research CSP^3. Last but not least, recent research in other communities, such as software engineering (Yu et al., 2016; Nema et al., 2022) and cyber security (Andow et al., 2020, 2019), demonstrates requirements for analyzing privacy policies to help the analyst identify whether the apps' behavior is consistent with privacy policies.

In this work, we make the following efforts to advance CSP^3:

First, we construct a novel large-scale human-annotated **C**hinese **A**ndroid a**pp**lication **p**rivacy **p**olicy dataset, namely CA4P-483. Specifically, we manually visit the software markets, such as Google Play (Google, 2022a) and AppGallery (Huawei, 2022a), check the provided privacy policy website, and download the Chinese version if available. We finally collect 483 documents. To determine the labels in the privacy policy analysis scenario, we read through Chinese privacy-related regulations and summarize seven components (§2.2). We annotate all occurrences of components in 11,565 sentences from 483 documents. Unlike paragraph-level annotations in existing privacy policy datasets (Wilson et al., 2016), CA4P-483 annotates character-level corpus.

Second, based on CA4P-483, we summarize families of representative baselines for Chinese sequence labeling. In detail, we first evaluate the performance of several classic sequence labeling models on our dataset, including Conditional Random Forest (CRF) (Kudo, 2005), Hidden Markov Model (HMM) (Morwal et al., 2012), BiLSTM (Graves and Schmidhuber, 2005), BiLSTM-CRF (Lample et al., 2016), and BERT-BiLSTM-CRF (Devlin et al., 2018). Recent work shows lattice knowledge improves the performance of Chinese sequence labeling tasks. We involve lexicon-based models, such as Lattice-LSTM (Zhang and Yang, 2018).

Third, we investigate potential applications of CA4P-483. Combining law knowledge, we first identify whether the privacy policy violates regulation requirements based on CA4P-483. We also identify whether the app behaves consistently with privacy policy statements combing software analysis (Zhao et al., 2021; Zhou et al., 2020).

The contributions of this work are three-fold:

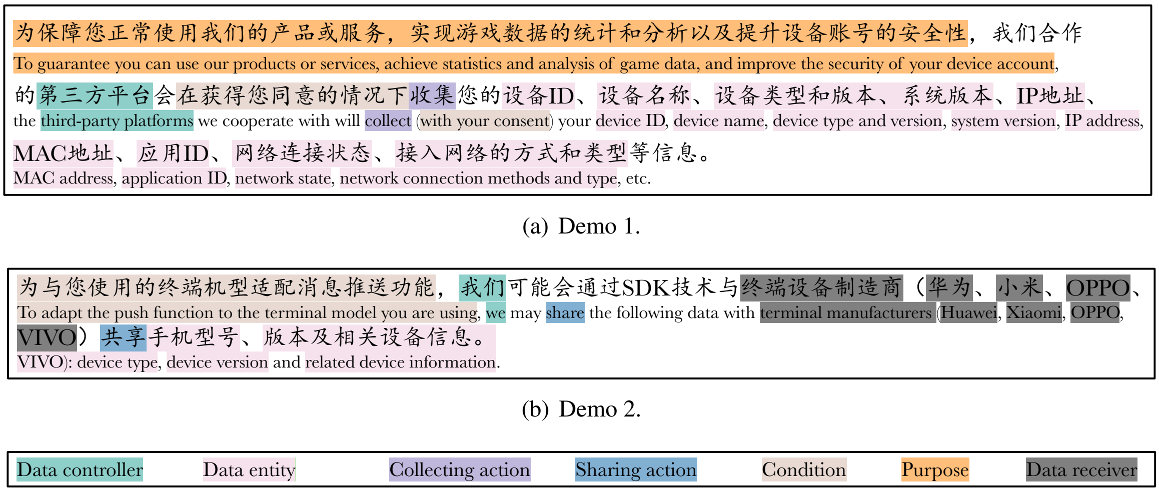
- To the best of our knowledge, we construct the first Chinese privacy policy dataset, namely CA4P-483, integrating abundant fine-grained annotations.
- We experimentally evaluate and analyze the results of different families of sequence labeling baseline models on our dataset. We also summarize difficulties in our dataset, and provide findings and further research topics on our dataset.
- We investigate potential applications of CA4P-483 to regulate privacy policies with law knowledge and program analysis technologies.

2 Dataset Construction

2.1 Dataset collection

We manually collect the Chinese privacy policies from Android application markets. According to application market requirements (Huawei, 2022b; Google, 2022b), developers **must** provide privacy policies to claim their user data access behavior and to ensure apps will not violate laws or regulations. Since privacy policies are publicly available for users to understand the apps' access of personal data, three authors of this paper manually access the most popular apps in markets and visit their privacy policy websites provided at the moment (January 2021). We use *html2text* (Alir3z4, 2011) to extract context. Finally, we use *tagtog* (Cejuela et al., 2014) for document annotation.

Next, we annotate CA4P-483 based on the law requirements. Specifically, we analyze Chinese privacy-related laws and regulations (NISSTC,



为保障您正常使用我们的产品或服务，实现游戏数据的统计和分析以及提升设备账号的安全性，我们合作
To guarantee you can use our products or services, achieve statistics and analysis of game data, and improve the security of your device account,
的第三方平台会在获得您同意的情况下收集您的设备ID、设备名称、设备类型和版本、系统版本、IP地址、
the third-party platforms we cooperate with will collect (with your consent) your device ID, device name, device type and version, system version, IP address,
MAC地址、应用ID、网络连接状态、接入网络的方式和类型等信息。
MAC address, application ID, network state, network connection methods and type, etc.

(a) Demo 1.

为与您使用的终端机型适配消息推送功能，我们可能会通过SDK技术与终端设备制造商（华为、小米、OPPO、
To adapt the push function to the terminal model you are using, we may share the following data with terminal manufacturers (Huawei, Xiaomi, OPPO,
VIVO）共享手机型号、版本及相关设备信息。
VIVO): device type, device version and related device information.

(b) Demo 2.

Data controller	Data entity	Collecting action	Sharing action	Condition	Purpose	Data receiver

(c) Annotation legend.

Figure 1: Annotation demos from CA4P-483. We translate the statements into English for illustration.

2020; PISS, 2020; of China et al., 2019; Committee, 2022), and find requirements for apps' privacy process behavior. For example, GB/T41391-2022 Article 4.n) claims that "*developers should expressly state the purpose of applying or collecting information to the subject of personal information.*" Finally, we summarize seven types of labels related to requirements for apps' access to user data.

2.2 Fine-grained annotations

For each privacy policy, we concentrate on the sentences that describe the data process behavior. After locating the sentences, we annotate seven components, i.e., the controller, data entity, collection, sharing, condition, purpose, and receiver.

Data controller. According to regulation requirements, the data controller is the party that determines the purpose and means of personal data processing. A data controller could be the app (first party) or the third party. As is shown in Fig.1, data controllers are "*third-party platforms*" in Fig.1(a) while that is "*we*" in Fig.1(b). Thus, we annotate data controllers according to sentence semantics, i.e., who is responsible for processing the data.

Data entity. Data entities are any information that can identify or reflect the activities of a natural person (PISS, 2020). Recent research (Cai and Chen, 2012; Shokri et al., 2017) demonstrates the probability of combining various information to infer and even locate a specific person. Thus, we annotate all data nouns or noun phrases that are requested in privacy policies, including sensitive information, such as device id, and normal information, such as device type.

Collection. Collection actions are verbs that describe how controllers access data, such as gather (收集), obtain (获取).

Sharing. Sharing actions are verbs that indicate whether the data controller will distribute data to others. Although both Sharing and Collection describe how the party access user data, we difference them according to the requirements of laws on the action, such as Article 5 and 9.2 in (PISS, 2020).

Condition. Condition describes the situation where the data controller will access personal data. Laws require data controllers to inform users under what conditions their data will be processed. For example, bank apps may require the users' identification information when activating bank account.

Purpose. Purpose should claim why the data controller processes user data. Laws enact specific requirements for user data access. For example, PISS Article 4.d) requires controllers to clearly state purpose of processing data. Purpose can also help the users understand why the app collects their data and further determine whether to give the consent as is shown in Fig.1(a).

Data receiver. Data receiver describes the parties that receive user data. Laws not only ask apps to clarify who will get shared data (PISS, 2020) but also restrict the data receivers' behavior (NISSTC, 2020), such as why processing user data.

2.3 Human annotation process

Our privacy policy annotation consists of two phases: coarse-grained annotation and fine-grained annotation. Coarse-grained annotation labels privacy policies at paragraph level following previous work (Wilson et al., 2016). Fine-grained annotation labels our defined components at the word level based on coarse-grained annotation.

For the first phase, three authors of this paper, who have researched privacy policies and software engineering for over eight and three years, label ten privacy policies for reference and record a video instruction to guide annotators. Then, we hire thirty undergraduates in our university to annotate the dataset. The three instructors train each annotator for at least four hours to be familiar with the dataset and requirements. Students are asked to annotate 1000 Android apps' privacy policies in Chinese and each privacy policy should be analyzed for at least 30 minutes to ensure quality. Each privacy policy is allocated to at least four annotators. Finally, three instructors inspect each annotation.

For the second phase, we select two undergraduates, who coarse-grained annotate the documents with high precision, to conduct the fine-grained annotation. Specifically, we select 483 documents that are well coarse-grained annotated after inspection. Instructors first annotate ten documents to lead undergraduates to annotate. The annotators also keep discussing with instructors once the role of components in sentences are unclear. Each annotator is required to label each privacy policy for at least 30 minutes to guarantee the dataset quality.

Finally, the instructors analyze the annotations and use Fleiss' Kappa metrics (Cohen, 1960; Wilson et al., 2016) to evaluate the agreements. Table 1 shows the average Kappa value (77.20%) satisfies the substantial agreement, i.e., Kaapa value lies in 0.61-0.80, and four components achieve almost perfect agreement (0.81-1.00). The *Condition*, which only gets moderate agreement, is caused by the overlap between labels (details in Appendix 9.3).

2.4 Dataset statistics and comparison

We conduct statistical analysis and show the results in Table 1. CA4P-483 is split into training, development, and test set. Table 1 also gives details of the number of different labels in each set. Table 1 shows that the average length of condition and purpose is much longer than other corpora as the two types are generally in the form of clauses.

# doc					483	
# sentences					11,565	
# sentences with ann					3,385	
Avg sentences len					79.06	
Type	Num	Train	Dev	Test	Avg len	Kappa
Data	21,241	18,925	2,521	2,331	4.68	85.39%
Collect	5,134	4,133	576	528	2.03	73.78%
Share	4,976	3,989	533	505	2.10	84.87%
Controller	8,424	6,085	815	782	2.49	82.22%
Condition	4,917	5,477	716	713	14.41	50.07%
Receiver	3,202	2,776	360	350	4.29	89.88%
Purpose	4,683	6,442	860	867	19.24	74.18%
Total	52,577	47,827	6,381	6,076		

Table 1: The statistics of CA4P-483. Here, "Avg" denotes *average*, "ann" denotes *annotation*, "len" denotes *length*, "#" denotes *the number of*.

We compare CA4P-483 with related datasets in Table 2. We first compare our corpus with Chinese sequence labeling datasets, such as MSRA (Zhang et al., 2006), OntoNotes (Weischedel et al., 2011), Weibo (Peng and Dredze, 2016), PeopleDiary (Zhang and Chen, 2017), Resume (Zhang and Yang, 2018), CLUENER2020 (Xu et al., 2020), and CNERTA (Sui et al., 2021). We also involve widely used English sequence labeling datasets, namely Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018). We also consider privacy policy datasets, namely Online Privacy Policies (OPP-115) (Wilson et al., 2016) and Android app privacy policies (APP-350) (Zimmeck et al., 2019).

We first compare the size and classes in different datasets. Table 2 shows that CA4P-483 contains abundant semantics, i.e., CA4P-483 has seven annotation classes that are larger than most other datasets (seven out of nine). For privacy policy-related datasets, the comparison is conducted with the number of documents as one privacy policy corresponds to one app. OPP-115 annotates at the sentence level, and APP-350 only annotates data controller, data entities, and modifiers. Since APP-350 specifies data entities into 16 categories, APP-350 exhibits more number of classes than CA4P-483. To summarize, CA4P-483 is the first and largest Chinese Android privacy policy dataset with abundant semantic labels.

3 Task and Experiment Setup

3.1 Task description

CSP[3] figures out <u>who</u> <u>collects</u> or <u>shares</u> what kind of <u>data</u> to <u>whom</u>, under which kind of <u>condition</u>, and <u>for what</u>. The underlined words correspond to each type of annotations. As CSP[3] concentrates

Dataset	# Train	# Dev	# Test	Size	Language	# Class
MSRA	41,728	4,636	4,365	50K	Chinese	3
PeopleDairy	20,864	2,318	4,636	23k	Chinese	3
Weibo	1,350	270	270	2k	Chinese	4
Resume	3,821	463	477	2k	Chinese	8
CLUENER2020	10,748	1,343	1,345	13K	Chinese	10
CNERTA	34,102	4,440	4,445	42,987	Chinese	3
Twitter-2015	6,176	1,546	5,078	12,784	English	4
Twitter-2017	4,290	1,432	1,459	7,181	English	4
CA4P-483	14,678	2,059	1,842	18,579	Chinese	7
Dataset	# Train doc	# Dev doc	# Test doc	Size	Language	# Class
OPP-115	75 doc	/	40 doc	115 doc	English	12
APP-350	188 doc	62 doc	100 doc	350 doc	English	18
CA4P-483	386 doc	48 doc	49 doc	483 doc	Chinese	7

Table 2: A comparison between CA4P-483 and other popular sequence labeling datasets. # denotes *"number"*. *"doc"* denotes *"documents"*.

on data access-related sentences, we first locate the sentences based on data collection and sharing words (Andow et al., 2020; Yu et al., 2016). We summarize the word list based on laws, app market requirements and previous works (Yu et al., 2016; Andow et al., 2019, 2020) (detailed in Appendix 9.1). Given the sentences $C = c_1, c_2, ..., c_n$ and its labels $L = l_1, l_2, ..., l_n$, where c_i denotes the i-th Chinese characters and l_i denotes the c_i's label, the task is to identify sequence labels.

3.2 Summarize models

This section introduces baseline methods for sequence labeling task on CA4P-483.

3.2.1 Probabilistic models

Hidden Markov Model (HMM): HMM[2] (Freitag and McCallum, 2000) is one of the most classic probabilistic models and is applied as our baselines.
Condition Random Field (CRF): CRF[3] (Lafferty et al., 2001) aggregates the advantages of HMM and counters the label bias problems.

3.2.2 Neural network models

BiLSTM: BiLSTM[2] (Graves and Schmidhuber, 2005) uses neural network to learn a mapping relation from sentences to labels through the nonlinear transformation in high-dimensional space.
BiLSTM-CRF: BiLSTM-CRF[2] uses BiLSTM as a encoder to map the sentences in to a hingh dimension vector and uses CRF as a decoder.
BERT-BiLSTM-CRF: Since BiLSTM-CRF is still limited to the word vector presentation, BERT-BiLSTM-CRF[4] (Dai et al., 2019) uses BERT as a feature extractor and takes advantage of BiLSTM and CRF for sequence labeling.

3.2.3 Lattice enhanced models

As Chinese words are not naturally separated by space, character-based methods omit the information hidden in word sequences. Thus, lattice-based methods that integrate lattice information are proposed for Chinese sequence labeling and achieve the promised performance.
LatticeLSTM: LatticeLSTM[5] (Zhang and Yang, 2018) takes inputs as the character sequence together with all character subsequences that match the words in a predefined lexicon dictionary.

3.3 Setup and implementation details

We evaluate baselines on an Ubuntu 20.04 server with 5 NVIDIA GeForce 3090 (24 GB memory for each), 512 GB memory, and an Intel Xeon 6226R CPU. Next, we present our implementation details. For HMM, the number of states, i.e., class number in our dataset with the BIO tag, is set as 22, and the number of observations, i.e., the number of different characters, is set as 1756, which is default value[2]. For CRF, we use the default settings in CRF++[3]. For BiLSTM and BiLSTM-CRF, embedding size is 128, learning rate is 0.001, and we train models using 30 epochs with a batch size of 64. For BERT-BiLSTM-CRF[4], we use the Chinese bert-base[6] pretrained model and fine tune it on our training data. The BiLSTM is set with 128 hidden

[2] https://github.com/luopeixiang/named_entity_recognition
[3] http://crfpp.sourceforge.net/
[4] https://github.com/macanv/BERT-BiLSTM-CRF-NER
[5] https://github.com/LeeSureman/Batch_Parallel_LatticeLSTM
[6] https://github.com/google-research/bert

	BiLSTM-CRF			BERT-BiLSTM-CRF			LaticeLSTM			Manual Agreements		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Collect	51.80%	57.50%	54.47%	50.59%	68.89%	58.34%	69.23%	67.05%	65.10%	96.30%	92.07%	94.14%
Condition	81.75%	72.76%	77.00%	31.59%	46.46%	37.61%	72.76%	77.00%	81.75%	93.53%	84.50%	88.79%
Data	77.85%	58.60%	66.44%	51.11%	67.19%	58.06%	58.60%	66.44%	77.85%	96.20%	91.79%	93.94%
Controller	64.10%	61.08%	62.50%	56.53%	63.80%	59.94%	61.08%	62.50%	64.10%	96.96%	90.18%	93.45%
Purpose	70.88%	54.61%	60.64%	40.45%	48.46%	44.09%	54.61%	60.64%	70.88%	95.64%	92.61%	94.10%
Share	68.31%	51.83%	58.88%	59.08%	45.61%	51.48%	51.83%	58.88%	68.31%	96.10%	94.71%	95.40%
Receiver	91.70%	92.68%	92.19%	22.96%	27.84%	25.17%	92.68%	92.19%	91.70%	97.33%	85.00%	90.75%
O	91.70%	92.68%	92.19%	46.22%	57.35%	51.18%	92.57%	92.79%	92.35%	/	/	/
Average	86.94%	86.90%	86.84%	37.54%	49.42%	42.66%	72.27%	72.27%	72.27%	96.01%	90.12%	92.94%

Table 3: Evaluation performance of three types of methods on our dataset. "O" denotes *others*.

	P	R	F1
HMM	77.47%	66.11%	69.63%
CRF	85.52%	86.28%	85.63%
BiLSTM	85.13%	85.99%	85.05%
BiLSTM-CRF	86.94%	86.90%	86.84%
BERT-BiLSTM-CRF	46.22%	57.35%	51.18%
Lattice-LSTM	78.63%	80.75%	79.67%

Table 4: Overall performance of baseline methods on our dataset.

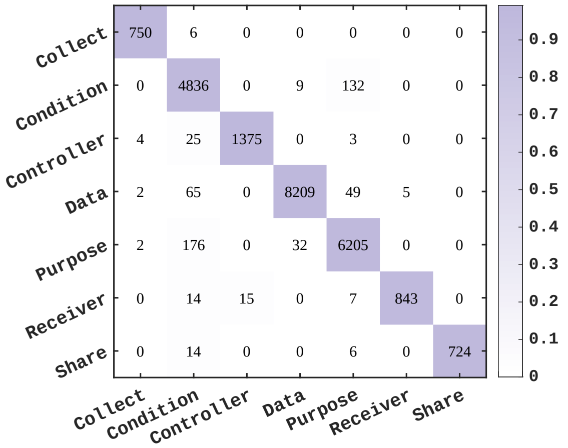
layer and a learning rate of 1e-5. BERT-BiLSTM-CRF model is trained on our dataset with default settings[4] where the batch size is 64, learning rate is $1e^{(} - 5)$, dropout rate is 0.5, gradient clip is 0.5, and early stop strategy is "*stop if no decrease*". For Lattice-LSTM, we use the same lattice provided in (Zhang and Yang, 2018).

4 Evaluation

In this section, we evaluate baseline methods on all 18,579 sentences that are divided into training, development, and testing sets as detailed in Table 2. Following previous research (Wilson et al., 2016; Sui et al., 2021), we apply precision (P), recall (R), and F1-score (F1) to evaluate baselines.

Table 4 shows the overall performance of families of baselines on CA4P-483. Table 4 shows that BiLSTM-CRF achieves the most promising performance, which may benefit from the enhanced presentation ability of bidirectional LSTM and CRF for capturing the context information. Lattice-LSTM performs a strong representation of capturing lattice information, while some clauses in our labels may mislead the model learning the patterns.

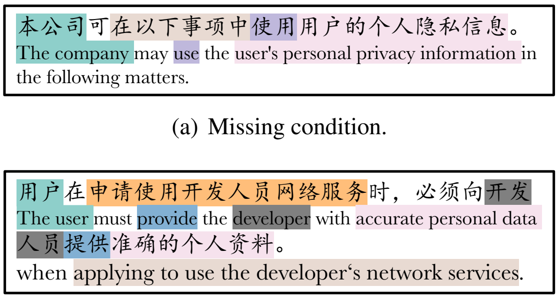
We analyze the identification performance of each component to investigate the challenges and limitations of CA4P-483. Table 3 demonstrates the detailed performance of baselines, i.e., CRF-based models, BERT-based models, and Lattice-based models. Besides, we also compare the performance with *manual agreements* to demonstrate task difficulties. Table 3 demonstrates that BiLSTM-CRF and Lattice-LSTM achieve over 90% performance on *Receiver* because the *Receiver* possesses few overlaps with other labels and is in the format of words. *Collect* and *share* only achieve around 60% precision and F1-score because the two types of entities perform some overlapping, as is shown in Fig.1 and Fig.5. Table 3 shows that BiLSTM-CRF achieves better precision on *Condition* than Lattice-LSTM, which may be caused by the fact that *Condition* and *Purpose* are mainly in the format of attributive clauses rather than words.



	Collect	Condition	Controller	Data	Purpose	Receiver	Share
Collect	750	6	0	0	0	0	0
Condition	0	4836	0	9	132	0	0
Controller	4	25	1375	0	3	0	0
Data	2	65	0	8209	49	5	0
Purpose	2	176	0	32	6205	0	0
Receiver	0	14	15	0	7	843	0
Share	0	14	0	0	6	0	724

Figure 2: Confusion matrix of BiLSTM-CRF results on CA4P-483.

Next, we analyze the confusion matrix of BiLSTM-CRF results that performs the best on CA4P-483. In Fig.2, the depth of background color denotes the proportion of classification, the darker the color the higher the proportion, and the digit denotes the number of classification results. Fig.2 indicates that most of the misclassification samples are related to *Condition*.



本公司可在以下事项中使用用户的个人隐私信息。
The company may use the user's personal privacy information in the following matters.

(a) Missing condition.

用户在申请使用开发人员网络服务时，必须向开发人员提供准确的个人资料。
The user must provide the developer with accurate personal data when applying to use the developer's network services.

(b) Error prediction when controller is user.

Figure 3: The visualization of divergence between ground truth and prediction.

To have a deep understanding of divergences between ground truth and predictions, we inspect the misclassifications. We find that the algorithm may fail to identify *Conditions*, which are in the adverbial clause as shown in Fig.3(a) where the highlighting for Chinese is ground truth and highlighting for English is prediction results. Besides, when the data controller is the user, as is shown in Fig.3(b), the algorithms fail to distinguish *Purpose* and *Condition*. More illustrations in Appendix 9.3 also reveal that models need to be well designed to learn deep semantic information, such as distinguishing overlapping among components, and distinguishing *Purpose* in modifiers.

5 Case Study

In this section, we will present cases of potential applications of CA4P-483, such as whether privacy policies comply with regulatory requirements and whether privacy policies is consistent with the apps' functionalities.

Regulation compliance identification. Chinese privacy-related laws (PISS, 2020; NISSTC, 2020; CLPRC, 2016) ask developers to clearly claim purpose conditions for processing user privacy data. We first investigate the distribution of annotations in CA4P-483. Fig.4 sketches the box plot of the frequency of components in each privacy policy. Fig.4 indicates that some privacy policies claim data processing without clarifying the purpose and condition, i.e., the minimum frequency of *Data* is positive while that of *Purpose* is zero. We manually inspect privacy policies. We find that the privacy policies, whose package name is *com.yitong.weather*, claim the app collects users' data while omitting to give the purposes or conditions of data access, which violates regulation requirements. Thus, CA4P-483 can facilitate the research in the area of privacy compliance identification (Andow et al., 2019; Barth et al., 2022).

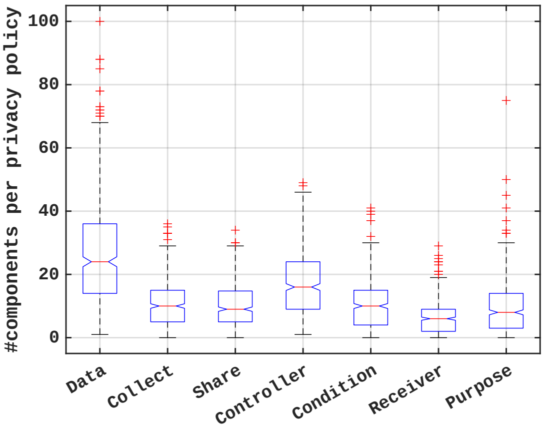


Figure 4: Components distribution of CA4P-483.

App behavior consistency identification. To improve the security of the Android community, researchers design systems (Andow et al., 2020; Yu et al., 2018) to identify the consistency between privacy policies and app behaviors to prevent apps from abusing user data or conducting malicious behavior. One popular method to check the app's behavior is dynamic analysis (Yan and Yin, 2012), i.e., running the app on the device and checking the *log* information. To investigate the application of CA4P-483 in security community, we first identify the privacy policies without *purpose* or *condition* components. Then, we install the app on one smartphone, manually interact with the app and try our best to trigger all possible functions in the app by clicking every visible buttons. We use *logcat* to capture the app's running information. We find that the app (id: *com.chengmi.signin*) requests device storage to use the app's functionalities while no condition-related statements are claimed in its privacy policy. With more intelligent automatic software engineering techniques, CA4P-483 can facilitate the research in this area, and more vulnerabilities in the consistency between app behavior and privacy policy could be investigated.

6 Discussion

In this section, we first discuss difficulties in CA4P-483. Then, we propose potential research topics on CA4P-483. Finally, we discuss limitations of CA4P-483. Besides, we also discuss ethical concerns in Appendix 9.2.

6.1 Dataset difficulties

Based on evaluation results in §4 and related work, we raise the following difficulties: 1) How to distinguish overlaps between components? 2) How to effectively deal with length variation of components? 3) Difficulties in semantic analysis.

Different from traditional sequence labeling tasks, components in our data set may contain other components. One scenario is the Purpose or Conditions maybe used to decorate the data, for example, "*We will collect your login information* (我们会收集您的登录信息)" where the *login* may be understood as the purpose of *information*. Since traditional sequence labeling methods predict one character with one label, it is hard to distinguish components overlaps. One possible solution is using multi-model algorithms (Sui et al., 2021) that demonstrate effectiveness for distinguishing boundaries between entities. Similar to traditional news or social media datasets that use voice or images as additional information, integrating apps' analysis results help distinguish different components.

Second, existing sequence labeling tasks mainly concentrate on entity recognition, while practical applications may require labeling clauses for further analysis. Table 1 shows that average length of components in CA4P-483 varies from 2.03 to 19.24. CSP[3] not only require identifying words but also ask the models to identify the role of clauses.

The semantic analysis of privacy policies is still a difficulty. Laws require apps to clearly clarify how apps collect and share user data. Privacy policies can claim that apps will *share* data with third parties or third parties will *collect* user data. In this way, it becomes essential to understand the context to distinguish the controller and action type. It could be a solution to use multi-model algorithms integrating program analysis to improve the performance; however, identifying the third party and app itself remains a challenge in program analysis.

6.2 Further directions

The CA4P-483 enables research in directions of interest to natural language processing, privacy protection, and cyber security (Zhu et al., 2022a,b). We propose some potential research interests for further work below.

Emotional analysis in privacy policies. Existing research (Andow et al., 2019) figures out privacy policies may conflict among contexts. For example, the privacy policy may claim NOT to collect user data in one sentence while claiming to access user data in other sections. Existing methods (Andow et al., 2019, 2020; Yu et al., 2018) use negative words to identify whether conflicts exist in the privacy policy and ignore complications like a double negative. In Chinese privacy policy, negative representations are more complicated (Liu, 2012). Thus, emotional analysis can help analysts better understand the semantics of privacy policies..

Privacy compliance detection. CA4P-483 provides detailed labels for data usage, including the purpose and conditions. It is necessary to investigate the detailed requirements of laws and further identify whether existing privacy policies violation.

Cyber security investigation. Privacy policies ought to reflect the functionalities of apps. Some apps may conceal the malicious behavior in their functionalities and do not claim the behavior in privacy policies. CA4P-483 can help identify the consistency between apps' functionalities and behavior by combing natural language process algorithms and code analysis.

6.3 Limitations

CA4P-483 provides detailed annotations for data access statements in privacy policies. However, analyzing privacy policies using CA4P-483 depends on the performance of locating data access-related sentences. We use data collection and sharing words to locate the sentences. However, some Purpose and Condition claims maybe given as an enumeration format, such as "*we will not share your personal data under the following conditions*". CA4P-483 is limited when capturing information in enumeration format.

Privacy policies possess timeliness. App developers should provide a privacy policies when publishing the apps. When the apps' functionality updates, the privacy policies ought to be updated accordingly. The data set is limited to the timestamp we collected. When combining our dataset with program analysis, this factor should be considered.

7 Related Work

Prior privacy policy datasets are all English and omit other languages. OPP-115 (Wilson et al., 2016) collects 115 English websites' privacy policies and makes annotations at the sentence level. OPP-115 designs labels based on previous works (McDonald and Cranor, 2008; Staff, 2011). APP-350 (Zimmeck et al., 2019) gathers Android apps' privacy policies written in English. APP-350 only conducts limited annotations, including two types of data controllers, namely first party and third party, thirteen types of specific data, and two types of modifiers, i.e., do and do not.

Existing Chinese sequence labeling datasets are generally gathered from News (Zhang et al., 2006; Zhang and Chen, 2017; Sui et al., 2021) and social media (Peng and Dredze, 2016; Weischedel et al., 2011; Zhang and Yang, 2018). The datasets include abundant corpus, but their annotations are limited to location, person name, and organization. Even though CLUENER2020 (Xu et al., 2020) expands the labels, such as game, gvoerment, book, the datasets are still hard to be applied in specific downstream tasks. CNERTA (Sui et al., 2021) includes another media data, i.e., voice data, to improves the sequence labeling performance.

8 Conclusion

This paper introduces the first Chinese Android application privacy policy dataset, CA4P-483. CA4P-483 contains fine-grained annotations based on requirements of privacy-related laws and regulations. The dataset can help promote natural language processing research on practical downstream tasks. We also conduct experimental evaluations of popular baselines on our dataset and propose potential research directions based on the result analysis. We also conduct case studies to investigate potential applications of our dataset and potential application of our dataset to help software engineering and cyber security protect user privacy. In the future, we hope that we can build new models for CA4P-483 to counter the challenges.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This work was partially supported by Hong Kong RGC Project (PolyU15224121), NSFC Young Scientists Fund (62006203), and HKPolyU collaborative research project (ZGBE).

References

Alir3z4. 2011. html2text. https://github.com/Alir3z4/html2text.

Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. Policylint: investigating internal privacy policy contradictions on google play. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 585–602.

Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words: Entity-Sensitive privacy policy and data flow analysis with PoliCheck. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 985–1002. USENIX Association.

Susanne Barth, Dan Ionita, and Pieter Hartel. 2022. Understanding online privacy—a systematic review of privacy visualizations and privacy by design guidelines. *ACM Computing Surveys (CSUR)*, 55(3):1–37.

Liang Cai and Hao Chen. 2012. On the practicality of motion based keystroke inference attack. In *International Conference on Trust and Trustworthy Computing*, pages 273–290. Springer.

CCPA. 2016. California consumer privacy act regulations. https://govt.westlaw.com/calregs.

Juan Miguel Cejuela, Peter McQuilton, Laura Ponting, Steven J Marygold, Raymund Stefancsik, Gillian H Millburn, Burkhard Rost, FlyBase Consortium, et al. 2014. tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. *Database*, 2014.

CLPRC. 2016. Cybersecurity law of the people's republic of china. http://www.gov.cn/xinwen/2016-11/07/content_5129723.htm.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

National Information Security Standardization Technical Committee. 2022. Information security technology — basic specification for collecting personal information in mobile internet applications. http://www.cac.gov.cn/1124853418_15652571749671n.pdf.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, pages 1–5. IEEE.

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

Keyang Ding, Jing Li, and Yuji Zhang. 2020. Hashtags, emotions, and comments: a large-scale dataset to understand fine-grained social emotions to online topics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1376–1382.

Ming Fan, Le Yu, Sen Chen, Hao Zhou, Xiapu Luo, Shuyue Li, Yang Liu, Jun Liu, and Ting Liu. 2020. An empirical evaluation of gdpr compliance violations in android mhealth apps. In *2020 IEEE 31st international symposium on software reliability engineering (ISSRE)*, pages 253–264. IEEE.

Dayne Freitag and Andrew McCallum. 2000. Information extraction with hmm structures learned by stochastic optimization. *AAAI/IAAI*, 2000:584–589.

GDPR. 2016. General data protection regulation. https://gdpr-info.eu.

Google. 2022a. Google play. https://play.google.com.

Google. 2022b. Google play policies. https://developer.android.com/distribute/play-policies.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Huawei. 2022a. App gallery. https://appgallery.huawei.com/Featured.

Huawei. 2022b. Appgallery review guidelines. https://developer.huawei.com/consumer/en/doc/30202.

Taku Kudo. 2005. Crf++: Yet another crf toolkit. *http://crfpp. sourceforge. net/*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp*, 4:543.

Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol*, 1.

Preksha Nema, Pauline Anthonysamy, Nina Taft, and Sai Teja Peddinti. 2022. Analyzing user perspectives on mobile app privacy at scale. In *International Conference on Software Engineering (ICSE)*.

NISSTC. 2020. Cybersecurity practices guidelines – security guidelines for using software development kit (sdk) for mobile internet applications (app) (tc260-pg-20205a). https://www.tc260.org.cn/front/postDetail.html?id=20201126161240.

Cyberspace Administration of China, Ministry of Industry, Information Technology, Ministry of Public Security, and State Administration for Market. 2019. Measures for determining the illegal collection and use of personal information by apps. http://m.legaldaily.com.cn/zt/content/2021-11/16/content_8628724.htm.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 149–155.

PISS. 2020. Information security technology – personal information security specification. https://www.tc260.org.cn/front/postDetail.html?id=20200918200432.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Gulshan Shrivastava and Prabhat Kumar. 2021. Android application behavioural analysis for data leakage. *Expert Systems*, 38(1):e12468.

Nir Sivan, Ron Bitton, and Asaf Shabtai. 2019. Analysis of location data leakage in the internet traffic of android-based mobile devices. In *22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, pages 243–260.

FTC Staff. 2011. Protecting consumer privacy in an era of rapid change–a proposed framework for businesses and policymakers. *Journal of Privacy and Confidentiality*, 3(1).

statcounter. 2022. Mobile operating system market share worldwide. https://gs.statcounter.com/os-market-share/mobile/worldwide.

Statista. 2022. Number of mobile app downloads worldwide from 2019 to 2021, by country. https://www.statista.com/statistics/1287159/app-downloads-by-country/.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale chinese multimodal ner dataset with speech clues. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2807–2818.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Lok Kwong Yan and Heng Yin. 2012. {DroidScope}: Seamlessly reconstructing the {OS} and dalvik semantic views for dynamic android malware analysis. In *21st USENIX security symposium (USENIX security 12)*, pages 569–584.

Yu Yang, Jiannong Cao, Milos Stojmenovic, Senzhang Wang, Yiran Cheng, Chun Lum, and Zhetao Li. 2021. Time-capturing dynamic graph embedding for temporal linkage evolution. *IEEE Transactions on Knowledge and Data Engineering*.

Le Yu, Xiapu Luo, Jiachi Chen, Hao Zhou, Tao Zhang, Henry Chang, and Hareton K. N. Leung. 2018. Ppchecker: Towards accessing the trustworthiness of android apps' privacy policies. *IEEE Transactions on Software Engineering*.

Le Yu, Xiapu Luo, Xule Liu, and Tao Zhang. 2016. Can we trust the privacy policies of android apps? In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 538–549. IEEE.

Le Yu, Tao Zhang, Xiapu Luo, and Lei Xue. 2015. Autoppg: Towards automatic generation of privacy policy for android applications. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*.

Jingyuan Zhang and Mingjie Chen. 2017. People dairy. https://github.com/zjy-ucas/ChineseNER/.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Suxiang Zhang, Ying Qin, Wen-Juan Hou, and Xiaojie Wang. 2006. Word segmentation and named entity recognition for sighan bakeoff3. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 158–161.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.

Kaifa Zhao, Hao Zhou, Yulin Zhu, Xian Zhan, Kai Zhou, Jianfeng Li, Le Yu, Wei Yuan, and Xiapu Luo. 2021. Structural attack against graph based android malware detection. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3218–3235.

Hao Zhou, Xiapu Luo, Haoyu Wang, and Haipeng Cai. 2022a. Uncovering intent based leak of sensitive data in Android framework. In *ACM Conference on Computer and Communications Security (CCS)*.

Hao Zhou, Haoyu Wang, Xiapu Luo, Ting Chen, Yajin Zhou, and Ting Wang. 2022b. Uncovering cross-context inconsistent access control enforcement in android. In *The 2022 Network and Distributed System Security Symposium (NDSS'22)*.

Hao Zhou, Haoyu Wang, Shuohan Wu, Xiapu Luo, Yajin Zhou, Ting Chen, and Ting Wang. 2021. Finding the missing piece: Permission specification analysis for android ndk. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 505–516. IEEE.

Hao Zhou, Haoyu Wang, Yajin Zhou, Xiapu Luo, Yutian Tang, Lei Xue, and Ting Wang. 2020. Demystifying diehard android apps. In *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 187–198. IEEE.

Yulin Zhu, Yuni Lai, Kaifa Zhao, Xiapu Luo, Mingquan Yuan, Jian Ren, and Kai Zhou. 2022a. Adversarial robustness of graph-based anomaly detection. *arXiv preprint arXiv:2206.08260*.

Yulin Zhu, Yuni Lai, Kaifa Zhao, Xiapu Luo, Mingquan Yuan, Jian Ren, and Kai Zhou. 2022b. Binarizedattack: Structural poisoning attacks to graph-based anomaly detection. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 14–26. IEEE.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

Sharing	收集 (collect), 获取 (obtain), 接受 (get), 接收 (receive), 保存 (save), 使用 (use), 采集 (gather), 记录 (record), 存储 (store), 储存 (store)
Collection	披露 (reveal), 分享 (share), 共享 (share), 交换 (exchange), 报告 (report), 公布 (public), 发送 (send), 交换 (exchange), 转移(transfer), 迁移 (migrate), 转让 (make over), 公开 (public), 透露 (disclose), 提供 (provide)

Table 5: Data access word list

9 Appendix

9.1 Data access word list

Table 5 gives data sharing and collection word list, that is summarized from laws (NISSTC, 2020; GDPR, 2016; PISS, 2020), app market requirements (Google, 2022b; Huawei, 2022b), and previous works (Yu et al., 2016; Andow et al., 2019, 2020). With those words, researchers can locate data access-related sentences and conduct further analysis to get interested entities, such as data controller, data entity, collection, sharing, condition, purpose and data receiver.

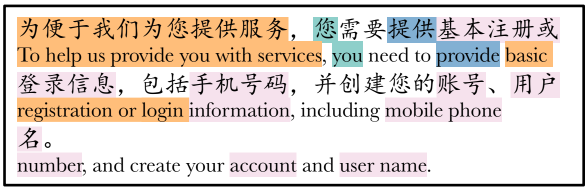
9.2 Ethical Consideration

CA4P-483 is a dataset constructed by gathering publicly available privacy policy websites without posing any ethical problems. First, privacy policies are publicly accessible in multi ways. According to application markets' requirements, developers or companies are asked to provide those privacy policy websites once they publish their apps. Privacy policies also ought to be given when the users use apps for the first time according to law requirements (PISS, 2020). Second, we do not collect any privacy information. Besides, the CA4P-483 is proposed to prompt research for protecting user privacy.

For the annotations, we hired part-time research assistants from our university to label the dataset. They are compensated with 9 USD/hour and at most 17.5 hours per week.

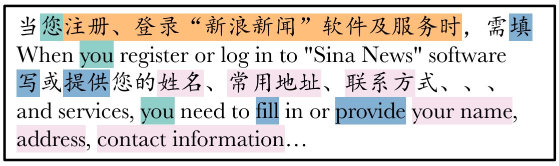
9.3 Prediction results analysis

In this section, we show the prediction results of the algorithm and some common problems. These problems could be the limitations of existing models and also be challenges for designing algorithms for our data scenario.



为便于我们为您提供服务，您需要提供基本注册或
To help us provide you with services, you need to provide basic
登录信息，包括手机号码，并创建您的账号、用户
registration or login information, including mobile phone
名。
number, and create your account and user name.

Figure 5: Overlapping between components. Differences between ground truth and prediction.



当您注册、登录“新浪新闻”软件及服务时，需填
When you register or log in to "Sina News" software
写或提供您的姓名、常用地址、联系方式、、、
and services, you need to fill in or provide your name,
address, contact information…

Figure 6: The visualization of divergence between ground truth and prediction for missing Purpose.

Fig. 5 illustrates the scenario where there exist overlapping between components, i.e., the "*basic registration or login information* (基本注册或登录信息)". Exactly, "basic registration or login information" should be one data as is highlighted in Chinese version, i.e., the ground truth. However, the algorithm will prediction "*basic registration or login* (基本注册或登录)" as Purpose and "*information*(信息)" as Data as are highlighted in English version. The meaning of color for different categories can be refer to Figure 1. Fig. 6 shows the pre-trained algorithm may misclassify *Purpose* as *Condition* when data controller is the user.