CAPP-130:A Corpus of Chinese Application Privacy Policy Summarization and Interpretation

Pengyun Zhu, Long Wen, Jinfei Liu,

Feng Xue, Jian Lou, Zhibo Wang, Kui Ren

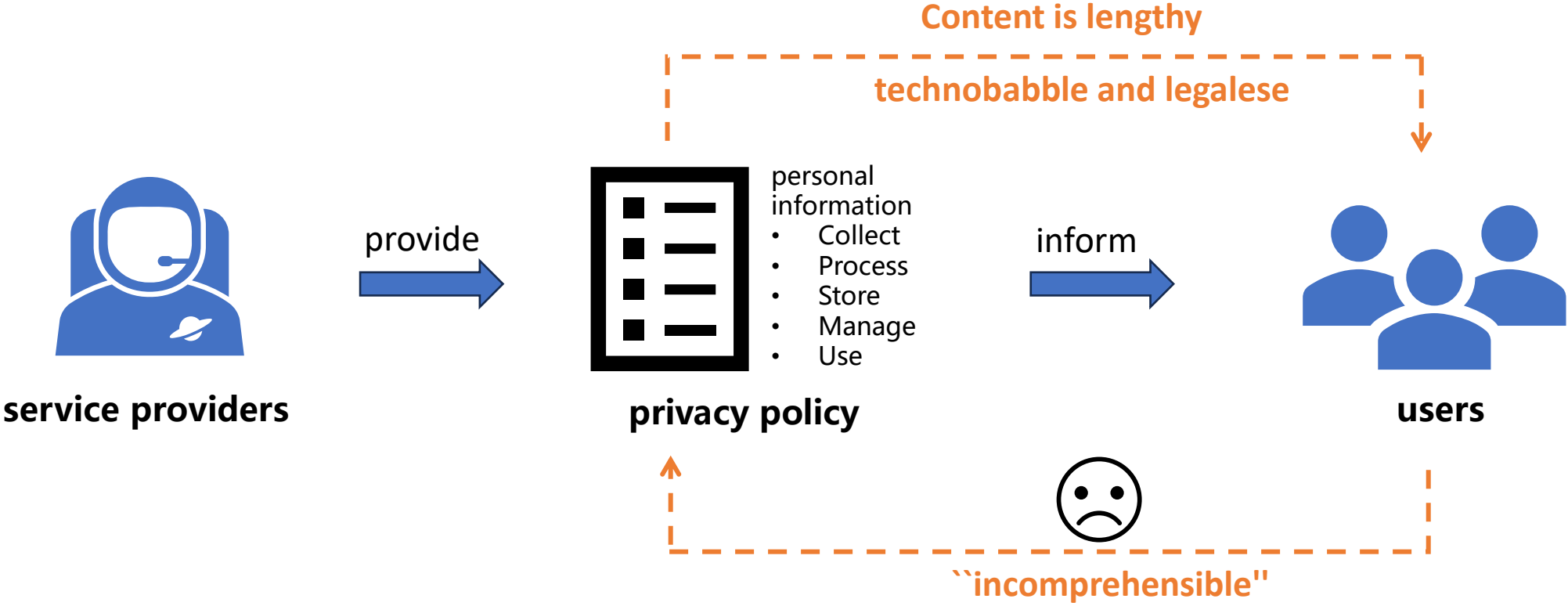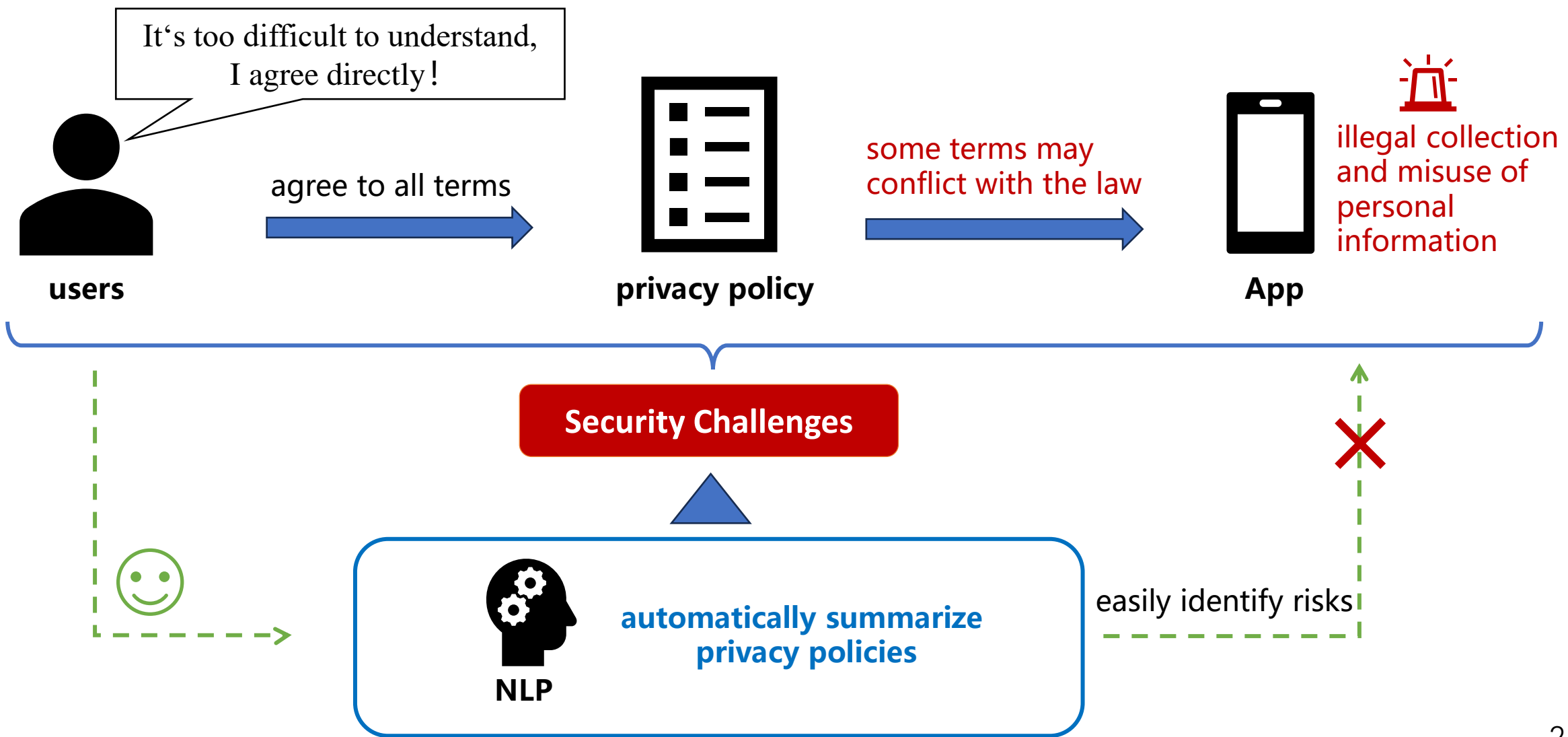*Zhejiang University*

Email : jinfeiliu@zju.edu.cn

CAPP-130 : A Corpus of Chinese Application Privacy Policy Summarization and Interpretation

Pengyun Zhu, Long Wen, Jinfei Liu,
Feng Xue, Jian Lou, Zhibo Wang, Kui Ren
Zhejiang University

NEURAL INFORMATION
PROCESSING SYSTEMS

Privacy Policy Original Sentence

shorter agreements

Contributions：

privacy policy corpora

summarization tools

:(failing to address the poor readability caused by the ``incomprehensible'' of technobabble and legalese

:(research on Chinese application privacy policy summarization is currently almost nonexistent

:(lack of a high-quality corpus suitable for addressing readability issues

3

Chinese Application Privacy Policy Corpus(CAPP-130)

130 Privacy Policies

38989 Sentences

52489 Annotations

20555 Rewritten Sentences

Special Marking
- Important
- Risk

Data Practice Category
- First Party Collection Information
- Permission Acquisition
- Third Party Sharing/Disclosure
- Usage
- Data Retention
- Data Security
- Edit/Control
- Specific Audiences
- Contact Information
- Policy Change
- Cease Operation
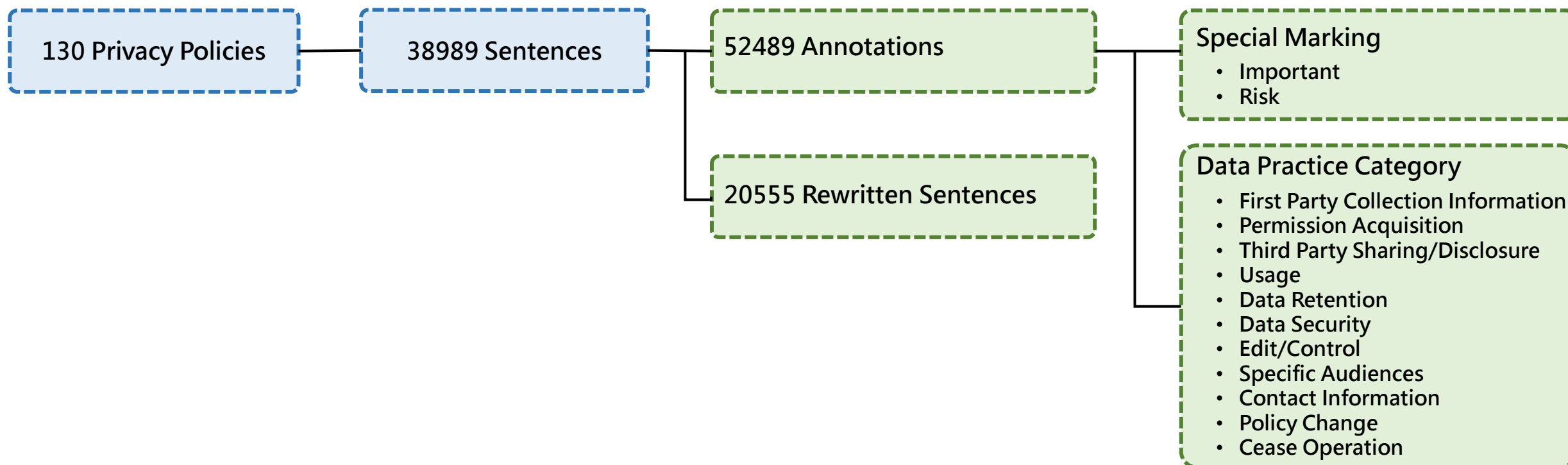
Topic-Controlled Framework for Summarization and Interpretation of Privacy Policy (TCSI-pp)
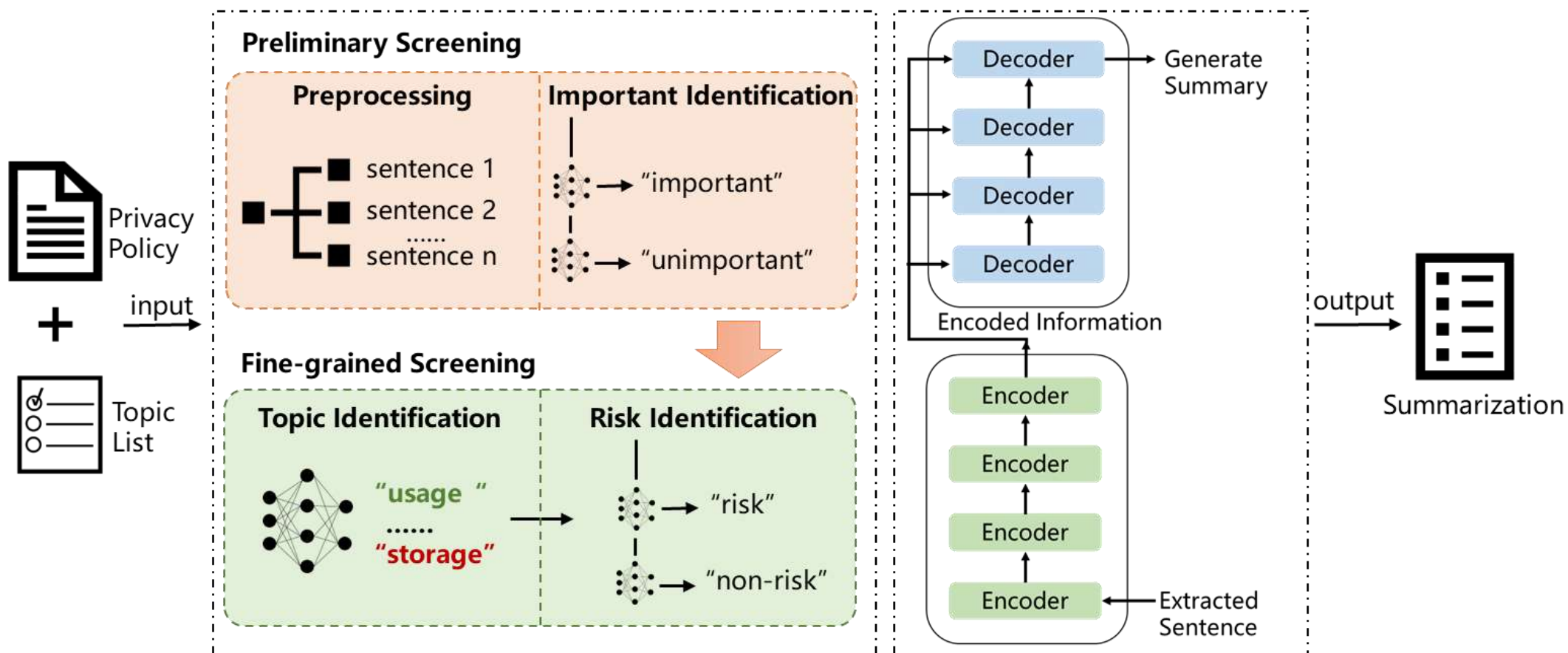
At the step of fine-grained screening, "green" represents the selected category according to the topic list, enter the next step, and "red" represents the non-selected category, do not enter the next step

Pengyun Zhu, Long Wen, Jinfei Liu,
Feng Xue, Jian Lou, Zhibo Wang, Kui Ren
Zhejiang University

CAPP-130 : A Corpus of Chinese Application Privacy Policy Summarization and Interpretation

We select the most effective RoBERTa and mT5 to implement TCSI-pp-zh

Table 1: Evaluation metrics for the preliminary and fine-grained screening models.

Method	topic-multi		important-binary		risk-binary	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
RoBERTa	**0.819**	**0.841**	**0.897**	**0.899**	0.920	0.711
Bert	0.802	0.820	0.895	0.896	0.921	0.719
mBERT	0.809	0.821	0.889	0.889	0.918	0.709
SBERT	0.781	0.794	0.875	0.874	0.917	0.689
PERT	0.801	0.812	0.895	0.897	**0.922**	**0.716**
ERNIE	0.807	0.821	0.895	0.896	0.921	0.702

Table 2: Evaluation metrics for the rewritten sentence models.

Model	ROUGE-1	ROUGE-2	ROUGE-L	Bert-score	Bart-score	Carburacy
mT5	**0.753**	**0.609**	**0.733**	**0.888**	**-4.577**	**0.833**
Bert2gpt	0.751	0.574	0.720	0.872	-4.964	0.764
RoBERTa2gpt	0.749	0.577	0.719	0.872	-4.975	0.755
Bert2bert	0.718	0.535	0.689	0.864	-5.020	0.747
ERNIE2gpt	0.623	0.406	0.581	0.809	-5.716	0.715

Table 3: Readability Questionnaire Survey.

Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Percentage
TCSI-pp-zh	**37**	**29**	**29**	17	**29**	25	**38**	24	24	**30**	46.23%
GPT-4	20	27	4	**39**	28	**31**	19	**30**	**35**	24	42.13%
Llama-2	4	4	5	3	2	5	4	5	5	5	6.39%
Checkpoint	0	1	23	2	2	0	0	2	2	2	5.25%

GPT-4 VS TCSI-pp-zh

GPT-4:

一、数据采集相关内容
Ⅰ. Relevant Content of Data Collection:
1. 我们收集和使用的个人信息包括注册时提供的姓名、性别、年份、地址、电话号码、电子邮箱地址、身份证明文件号码、收费相关的银行信息以及联络人信息等。
1. The personal information we collect and use includes the name, gender, year of birth, address, telephone number, email address, identification document number, bank information related to fees, and contact information provided during registration.
2. 在您使用本网站的服务期间，我们将继续授权使用这些信息，但在注销账号时我们会停止使用这些信息。
2. During your use of our website services, we will continue to have the authorized use of this information. However, upon account cancellation, we will cease the use of this information.
3. 此外，我们还会自动收集包括设备、软件信息、IP地址、搜索和浏览信息等日志信息。
3. Additionally, we also automatically collect log information, including device and software info-rmation, IP addresses, search and browsing information, and other relevant data.
4. 为保护您的隐私，我们会采取必要的措施来保护您的个人信息。
4. To protect your privacy, we will take necessary measures to safeguard your personal information. ✗

TCSI-pp-zh:

1. 我们采集的信息与数据
1. The Information and Data We Collect
(1) 当您注册成为用户或使用服务时，收集您提供的姓名、性别、年龄、地址等相关个人信息。
(1) When you register as a user or use our services, we collect relevant personal information you provide such as your name, gender, age, and address.
(2) 在您注销账号时，我们将停止使用信息。
(2) Upon the cancellation of your account, we will cease the use of your information.
(3) 您使用我们服务时，系统可能自动采集技术信息。
(3) When you use our services, our system may automatically collect technical information.
(4) 在您接入服务时，收集您的设备或软件信息。
(4) When you access our services, we collect information about your device or software.
(5) 收集您在使用我们服务时搜索和浏览的信息。
(5) We collect information about your searches and browsing history when you use our services.
All √

Figure 1: Summarizations of GPT-4 and TCSI-pp

7

CAPP-130 : A Corpus of Chinese Application Privacy Policy Summarization and Interpretation

Pengyun Zhu, Long Wen, Jinfei Liu,
Feng Xue, Jian Lou, Zhibo Wang, Kui Ren
Zhejiang University

NEURAL INFORMATION
PROCESSING SYSTEMS

Thank you for listening!