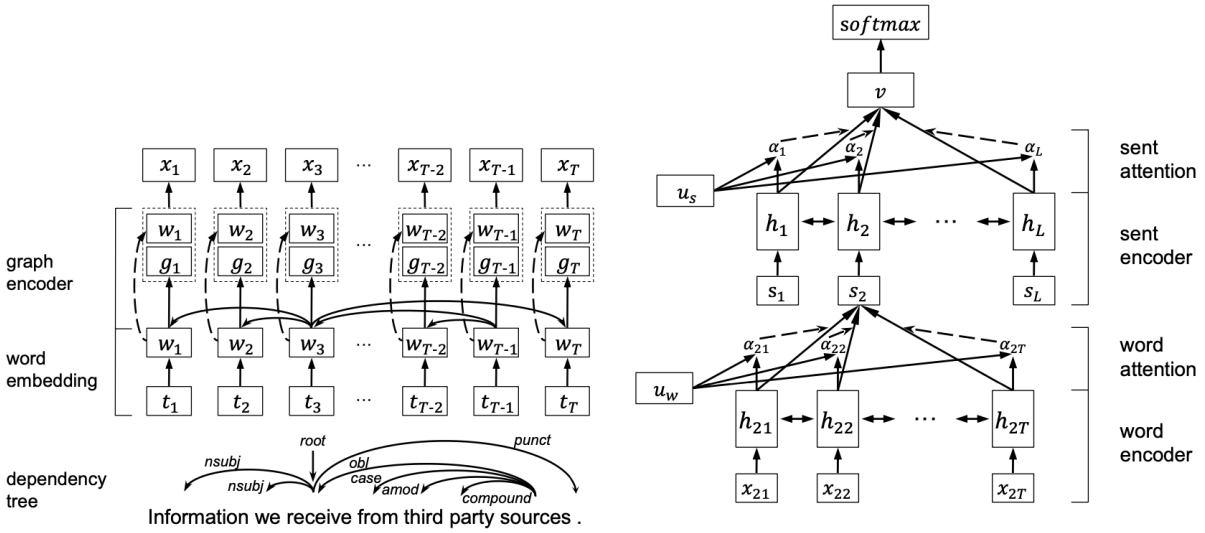
APPCorp: A Corpus for Android Privacy Policy Document Structure Analysis

Shuang LIU, Fan ZHANG, Baiyang ZHAO, Renjie GUO, Tao CHEN, Meishan ZHANG

Frontiers of Computer Science, DOI: 10.1007/s11704-022-1627-2

Problems & Ideas

- The problem of automatically analyzing privacy policy
  - Lacking high quality paragraph-level corpus.
  - Lacking customized analysis models.
- Ideas: We define the privacy policy document structure analysis problem as a paragraph classification task and curate a privacy policy corpus consisting of 231 privacy policies. We also benchmark the corpus with hierarchical deep learning models enhanced with the semantic and structural enhancement strategies.



(a) Input Representation

(b) The HAN Model Structure

Main Contributions

- Contributions:
  - Define the privacy policy document structure analysis problem as a paragraph classification task.
  - Curate a corpus of privacy policies.
  - Propose semantic and structural enhanced hierarchical models for the document structure analysis task.

Label	SVM			GloVe						BERT					
				HAN			HGAT			HAN			HGAT		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
PI	76.24	69.80	72.88	76.18	73.55	74.84	77.74	78.72	78.23	82.06	77.31	79.61	**82.31**	**79.34**	**80.80**
FPCU	75.01	**86.98**	80.55	81.02	82.32	81.66	**83.15**	81.58	82.36	82.55	86.00	**84.24**	82.91	85.02	83.95
CT	82.77	73.49	77.85	78.40	78.23	78.32	79.01	79.53	79.27	81.22	80.17	80.69	**83.22**	**81.25**	**82.22**
TPSC	78.73	74.83	76.73	79.56	77.80	78.67	77.67	78.26	77.96	**80.57**	80.32	**80.44**	79.48	**80.93**	80.20
URC	79.42	76.22	77.79	**81.60**	77.90	79.71	79.87	**81.34**	**80.60**	81.41	77.65	79.48	80.80	78.49	79.62
DS	**86.29**	72.51	78.81	77.42	81.68	79.49	82.32	81.68	82.00	82.63	**82.20**	82.41	86.11	81.15	**83.56**
DR	86.74	73.71	79.70	74.78	79.34	76.99	78.83	82.16	80.46	81.28	83.57	82.41	**86.96**	**84.51**	**85.71**
IDT	**76.06**	83.08	79.41	74.42	82.05	78.05	75.91	85.64	80.48	74.07	82.05	77.86	74.57	**88.72**	**81.03**
SA	**92.45**	73.57	81.94	79.83	83.18	81.47	83.58	84.08	83.83	86.08	81.68	83.82	88.12	**84.68**	**86.37**
PC	91.60	88.98	90.27	90.72	87.76	89.21	94.64	86.53	90.41	93.28	**90.61**	91.93	**95.67**	90.20	**92.86**
PCI	82.37	77.18	79.69	79.41	**81.08**	80.24	**83.02**	80.78	**81.89**	78.92	78.68	78.80	81.08	**81.08**	81.08
Micro	78.94	78.94	78.94	79.94	79.94	79.94	80.98	80.98	80.98	81.98	81.98	81.98	**82.50**	**82.50**	**82.50**
Macro	82.52	77.30	79.60	79.39	80.44	79.88	81.43	81.85	81.59	82.19	81.84	81.97	**83.75**	**83.22**	**83.40**

Both semantic enhancement and structural enhancement show improvement on the hierarchical structure model, the semantic enhancement strategy shows larger improvement, leading to an average of 82.5% Micro F1-score.