Safeguarding Text Generation API's Intellectual Property through Meaning-preserving Lexical Watermarks

Shiyu ZHU, Yun LI, Xiaoye OUYANG, Xiaocheng HU, Jipeng QIANG

Frontiers of Computer Science, DOI: 10.1007/s11704-023-3252-0

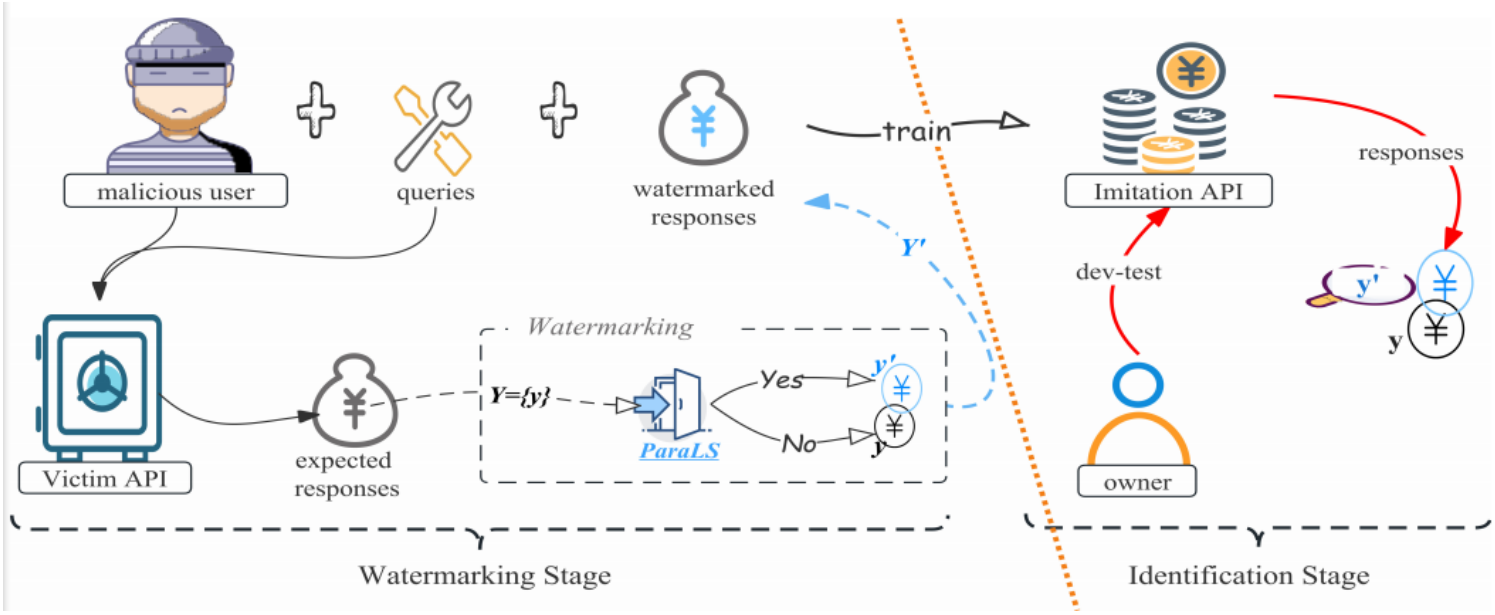
Problems & Ideas

Problems of traditional method :

Using WordNet to obtain lexical watermarks neglects rational substitutes. For instance, clean is used as a substitute for blank in all cases, without considering the meaning of blank in context, whether it means "not written or printed on" or "without comprehension", resulting in poor-quality text generation in watermarked models, especially with polysemous target words. Malicious users can easily detect lexical watermarks via error statistics and erase them.

Ideas:

We propose a novel meaning-preserving LW method to safeguard text generation API's IP. Our aim is to modify outputs while preserving the sentence's meaning.



A comprehensive examination of our watermarking and identification stages is depicted in this figure.

Experimental Result & Conclusions:

Experimental Result :

Table 1. Results on LS07 and CoInCo datasets.

Dataset	Method	best	best-m	oot	oot-m	P@1
LS07	Embedding	12.7	21.7	36.4	52.0	-
	Addocs	8.1	13.4	27.4	39.1	-
	XLNet	20.6	33.2	50.9	61.6	49.5
	BERT-based LS	20.3	34.2	55.4	68.4	51.1
	GeneSis	21.2	34.1	52.2	66.4	51.2
	LexSubCon	21.1	35.5	51.3	68.6	51.7
	ParaLS(Ours)	**24.1**	**42.4**	**58.2**	**76.5**	**58.3**
CoInCo	Embedding	8.1	17.4	26.7	46.2	-
	Addocs	5.6	11.9	20.0	33.8	-
	BERT-based LS	11.8	24.2	36.0	56.8	43.5
	XLNet	14.4	30.1	39.2	60.7	51.5
	LexSubCon	14.0	29.7	38.0	59.2	50.5
	ParaLS(Ours)	**18.1**	**40.0**	**49.2**	**75.4**	**62.6**

Table 2. Performance of different watermarking approaches.

	WMT14			CNN/DM		
	P-value↓	BLEU↑	BERTScore↑	P-value↓	ROUGE-L↑	BERTScore↑
w/o watermark	$>10^{-1}$	27.99	93.42	$>10^{-1}$	38.73	88.49
[7]						
- unigram	$<10^{-2}$	27.36(-0.63)	92.42(-1.00)	$<10^{-2}$	37.73(-1.00)	87.89(-0.60)
- trigram	$>10^{-1}$	27.72(-0.27)	93.13(-0.29)	$>10^{-1}$	38.61(-0.12)	88.28(-0.21)
- sentence	$>10^{-1}$	27.71(-0.28)	93.42(-0.00)	$>10^{-1}$	37.62(-1.11)	87.99(-0.50)
[2]						
- synonym (M=1)	$<10^{-4}$	27.81(-0.18)	93.32(-0.10)	$>10^{-9}$	37.64(-1.09)	88.21(-0.28)
- synonym (M=2)	$<10^{-8}$	27.60(-0.39)	93.30(-0.12)	$<10^{-12}$	37.75(-0.98)	88.20(-0.29)
[3]						
- DEP	$<10^{-4}$	27.81(-0.18)	93.27(-0.15)	$<10^{-2}$	37.61(-1.12)	88.14(-0.35)
- POS	$<10^{-7}$	27.72(-0.27)	93.30(-0.12)	$<10^{-7}$	37.56(-1.17)	87.70(-0.79)
Our Method	$<10^{-5}$	27.96(-0.03)	93.36(-0.06)	$<10^{-3}$	38.34(-0.39)	88.38(-0.11)

Conclusions:

- We propose a novel lexical substitution method based on a large-scale pre-trained neural paraphraser, improving Precision@1 score on LS07 and CoInCo datasets from 51.7% to 58.3% and from 50.5% to 62.6% respectively, outperforming previous state-of-the-art methods.PeGeLS achieves state-of-the-art results compared with the best BERT-base method.
- Our improved lexical watermarking method leverages lexical knowledge to considers polysemy's impact on generating watermarked models, preserveing sentence meaning accurately, reducing the risk of detection by users, while maintaining high confidence in post-hoc ownership verification,