Front.Comput.Sci.
DOI

SUPPLEMENT

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

1 Lexical Wateramrking Framework.

Despite the efficacy of the aforementioned lexical watermarking methods, some require substantial storage resources to fend off model imitation attacks, while others do not consider the context of target words. To address these shortcomings, we propose a lexical watermarking approach. To insert watermarks into imitation models for post-hoc identification of IP infringement, we propose Meaning-preserved Lexical Substitution approach (ParaLS) based on a paraphraser. Our methodology is illustrated in Figure 1, including two stages: *i) watermarking stage* and *ii) identification stage*.

i) Watermarking stage: Victim model $\mathcal{V}$ uses the proposed meaning-preserving lexical substitution to add lexical watermarks to the expected responses. When $\mathcal{V}$ receives queries Q from the end-user, $\mathcal{V}$ initially gives expected responses $\mathcal{Y}$, before sending to the user, meaning-preserving lexical substitution will add lexical watermarks on $\mathcal{Y}$ according to the generation of lexical watermarks, and finally replies to the user with watermarked responses $\mathcal{Y}_{watermark}$.

ii) Identification stage: If the copyright owner thinks that model $\mathcal{S}$ is suspected of infringement, the owner can analyze the response outputs $\mathcal{Y}$ of $\mathcal{S}$ on the dev-test sets O to assess whether it is infringed.

Received month dd, yyyy; accepted month dd, yyyy

E-mail: liyun@yzu.edu.cn, jpqiang@yzu.edu.cn, zhuqiaozhi75@gmail.com

2 Implementation Details.

Text Generation Tasks. We chose two broad text generation tasks: machine translation and document summarization, which have been successfully deployed as commercial APIs[1), 2)].

- **Machine Translation:** We consider WMT14 German (De) → English (en) translation as the testbed [1]. Moses [2] is applied to pre-process all corpora, with a cased tokenizer. BLEU [3] and BERTScore [4] are selected to evaluate translation quality. BLEU focuses on lexical similarity matched by n-grams, whereas BERTScore achieves semantic equivalence through contextualized embeddings.
- **Document Summarization:** CNN/DM [5] utilizes informative headlines as summaries of news articles. We reuse the dataset preprocessed by [6] with a partition of train/dev/test as 287K/13K/3k. Rouge [7] and BERTScore [4] are employed for the evaluation metric of the summary quality.

We utilize 32K and 16K BPE vocabulary [8] for experiments on WMT14 and CNN/DM, respectively. Table 1 shows the division of experimental sample sizes for the two tasks[3)].

Victim and Imitation Models. In the experiment, we adopt the Transformer-based [9] as the foundation for both the victim models and imitation models. Given the exceptional performance of pre-trained language models (PLMs)

[1)] https://translate.google.com/
[2)] https://deepai.org/machine-learning-model/summarization
[3)] Considering the fine-tuning involved, we set the sample size for the translation task to 250k.

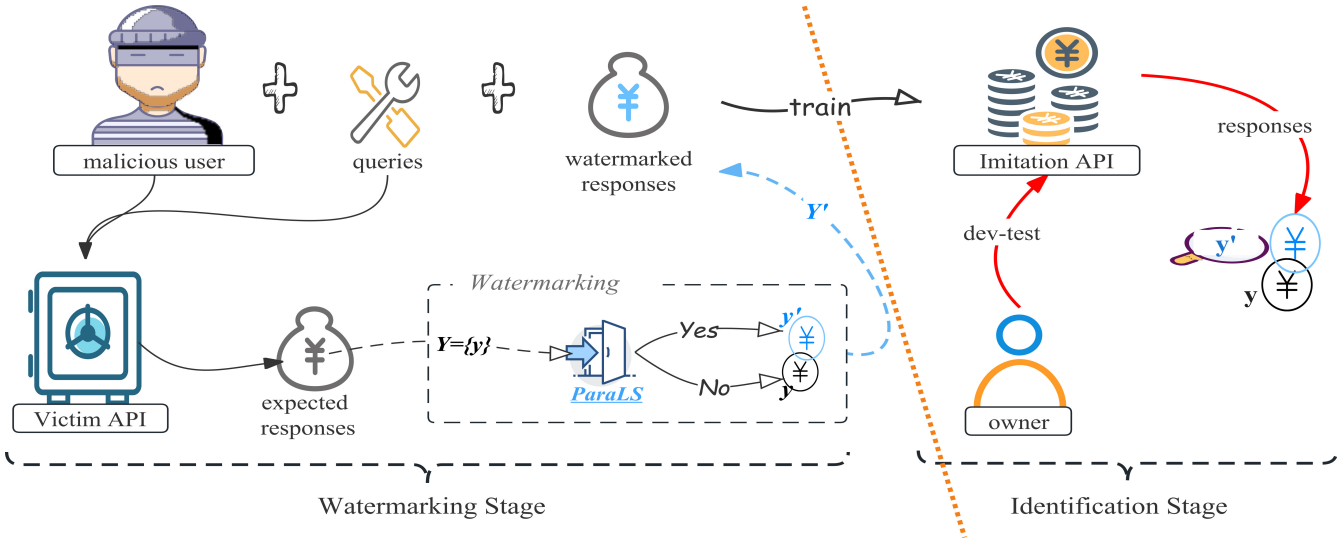


Fig. 1 A comprehensive examination of our watermarking and identification stages is depicted in this figure. The left section of the dashed line pertains to the watermarking stage, wherein the victim API dispatches watermarked responses to the end-user, and the right section pertains to the identification stage, wherein we assess copyright infringement by analyzing the output of the imitation API.

Table 1 Statistics of datasets used in our experiments.

	Train	Dev	Test
WMT14	250K	3K	3K
CNN/DM	287K	13K	3K

distributed on cloud platforms[4)], we consider utilizing two well-established PLMs: BART (summarization) [10] and mBART (translation) [11] as the victim model.

Implementation Details. To minimize the impact of other variables on demonstrating the efficacy of watermarking, we assume that the victim model $\mathcal{V}$ and the imitation model $\mathcal{S}$ utilize the same training data, but $\mathcal{S}$ utilizes the watermarked responses y' replied by $\mathcal{V}$ rather than the expected responses y. Additionally, there are two parameters that must be set in the above watermarking process: N high-frequency adjectives or adverbs and the top B candidates with the highest score of the target word x_i generated by ParaLS. In our experiment, we set these two parameters to 10 and 2 respectively.

We implement an English paraphraser based on Transformer model in FairSeq with an 8-layer encoder and decoder, 1024 dimensional embeddings, 16 encoder-decoder attention heads, and 0.1 dropout. We choose an English paraphrase dataset ParaBank2 [12] to train the paraphraser. The weights for the prediction score, BARTScore, and BLEURT are 0.02, 1, and 1, respectively. The number of outputted paraphrases B is set to 50.

[4)] https://cloud.google.com/ai-platform/training/docs/algorithms/bert

Comparison Methods. We select the state-of-the-art model watermarking methods [13] and its advanced version (CATER) [14] as the baseline for comparison, both of which have utilized the traditional lexical substitution method to embed lexical watermarks. Furthermore, we also include a comparison of [15] for bit watermarking.

LS Benchmarks. To evaluate our LS approach, we choose two widely used datasets, LS07 [16] and CoInCo [17]. And following previous LS methods [18, 19], we use the official metrics best, best-mode, oot, oot-mode in SemEval 2007 task as well as Precision@1(P@1) as our evaluation metrics. In detail, best, best-mode and P@1 evaluate the quality of the best predictions, yet oot (out-of-ten) and oot-mode benefit verifying the coverage of the gold substitute candidate list by the 10-top predictions.

We compare our proposed LS method with the following baselines. The two recent BERT-based methods BERT-based substitution [18] and LexSubCon [19]) are chosen. One method based on pretrained XLNet is chosen, denoted as XLNet [20]. We choose one newest supervised method GeneSis [21] and other previous LS methods including two embedding-based methods (Embedding [22] and Addocs [23]).

3 More Results

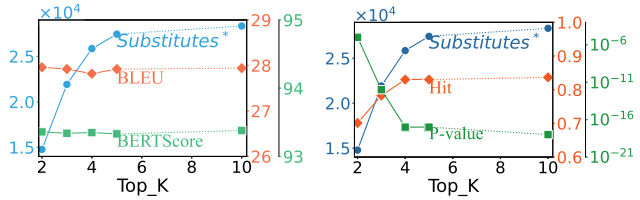
IP Identification on Cross-domain Imitation. In practice, victim models $\mathcal{V}$ generally do not make their training data

readily accessible to the general public. Therefore, the imitator may use data from other domains to query $\mathcal{V}$. We undertake imitation attacks on the machine translation task (utilizing the WMT14 training dataset) using two out-of-domain datasets to demonstrate that our method is immune to domain shift. The first dataset employed is IWSLT14, comprising 250k sentence pairs, and the second is the law-OPUS dataset[5)] with a sample size of about 2.1M. Table 2 illustrates that our approach can still watermark the imitation model $\mathcal{S}$ and identify watermarks with high confidence, despite the presence of domain mismatch.

Table 2 Imitation performance of using data from different domains.

WMT14	IWSLT14	OPUS(Law)
$<10^{-5}$	$<10^{-3}$	$<10^{-6}$

Influence of Top-K. Furthermore, we conducted additional experimentation to assess the influence of top-K on the quality of the text generation and the identifiability of watermarks for the machine translation task. As depicted in Figure 2, as top-K increases, restrictions on word selection are gradually relaxed, manifesting in a greater number of words that meet the substitution criteria over time. This was accompanied by a slight decrease in BLEU and BERTScore, but the impact was negligible. This also attests to the superior performance of our proposed LS in replacing words, taking into account the effects of polysemy and consistently capturing the intended meaning of target words. We also computed the ratio of watermark word hits, which revealed that as top-K increases, so too does the number of watermark word hits, thereby enhancing the identifiability of watermarks, as evidenced by the marked decrease in P-value, providing greater confidence in ownership claims.



(a) Assess the quality of the generated text. (b) Assess the quality of watermark identification.

Fig. 2 The influence of top-K on the quality of text generation and the identifiability of watermarks for machine translation task. *Substitutes** represents the number of words that satisfy the watermarking rules in Section 3.3.

[5)] OPUS(Law):https://opus.nlpl.eu/ELRC-EUIPO$_{l}aw - v1.php$

Case Study. For the two text generation tasks, we provide a representative example to examine the distinctions between our method and [13]. The results are presented in the Table 3. Due to the existence of polysemy, [13] is unable to locate the correct meaning of the target word *absent* in the context **x**. In the presented example, the meaning of *absent* is clearly similar to that of *blank*, but [13] still recognizes its specific meaning as *nonexistent* and replaces it with *missing* or *lacking*. Numerous similar errors damage the generation quality of the watermarked model and arouse the suspicion of malicious users, who then attempt to circumvent the lexical watermarks via the analysis of these errors. On the other hand, our method does not suffer from misunderstandings of polysemy, reduceing the risk of lexical watermarks being detected by users, as evidenced by the higher BLEU, ROUGE and BERTScore scores in Table **??** when compared to [13].

References

1. Bojar O, Buck C, Federmann C, Haddow B, Koehn P, Leveling J, Monz C, Pecina P, Post M, Saint-Amand H, Soricut R, Specia L. Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. June 2014
2. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Constantin A, Herbst E. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting ofthe Association for Computational Linguistics Companion Volume Proceedings ofthe Demo and Poster Sessions. June 2007
3. Papineni K, Roukos S, Ward T, Zhu W J. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting ofthe Association for Computational Linguistics. 2002
4. Zhang T, Kishore V, Wu F, Weinberger K Q, Artzi Y. Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations. 2019
5. Hermann K M, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. In: Advances in neural information processing systems. 2015
6. See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017, 1073–1083
7. Lin C Y. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. 2004
8. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting ofthe Association for Computational Linguistics. 2016
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I. Attention is all you need. arXiv preprint

Table 3 Examples of our lexical watermarking method compared with [13]. The lexical watermarks are marked in color.

Target word: - *absent* [adjective] Definition and Synonym set: - *nonexistent* , synset(*lacking*, *missing*, *wanting*) [from *WordNet*]	
Machine Translation and Document Summarization Tasks	**Correct Meaning?**
source sentence: Nachdem sie die schlechte Nachricht von ihrer Freundin gehört hatte, betrachtete sie das Bild abwesend.	
non-watermarked translation: After hearing the bad news from her friend, she looked at the picture in a absent way.	
[13]-watermarked translation(*absent*→ *missing*) : After hearing the bad news from her friend, she looked at the picture in a *missing* way.	✗
Ours-watermarked translation(*absent*→ *blank*) : After hearing the bad news from her friend, she looked at the picture in a *blank* way.	✓
source document: ... He just sat in his seat on the bench and stared blankly. His only idea to change things was to take off Januzaj and bring on Marouane Fellaini. Van Gaal claimed ...	
non-watermarked summary: ... Van Gaal offered no tactical knowledge from the sidelines, instead sitting quietly watching the game with a absent expression ...	
[13]-wateramrked summary(*absent*→ *lacking*) : ... Van Gaal offered no tactical knowledge from the sidelines, instead of sitting quietly watching the game with a *lacking* expression ...	✗
Ours-watermarked summary(*absent*→ *blank*) : ... Van Gaal offered no tactical knowledge from the sidelines, instead of sitting quietly watching the game with a *blank* expression ...	✓

arXiv:1706.03762, 2017

10. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting ofthe Association for Computational Linguistics. 2020
11. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. In: Transactions of the Association for Computational Linguistics. 2020
12. Hu J, Singh A, Holzenberger N, Post M, Van Durme B. Large-scale, diverse, paraphrastic bitexts via sampling and clus-tering. In: CoNLL. nov 2019, 44–54
13. He X, Xu Q, Lyu L, Wu F, Wang. C. Protecting intellectual property of language generation apis with lexical watermark. In: AAAI. 2022, 10758–10766
14. He X, Xu Q, Zeng Y, Lyu L, Wu F, Li J, Jia R. Cater: Intellectual property protection on text generation apis via conditional watermarks. In: NeurIPS. 2022
15. Venugopal A, Uszkoreit J, Talbot D, Och F, Ganitkevitch J. Watermarking the outputs of structured prediction with an application in statistical machine translation. In: EMNLP. 2011, 1363–1372
16. McCarthy D, Navigli R. Semeval-2007 task 10: English lexical substitution task. In: IWSE. 2007
17. Kremer G, Erk K, Pad´o S, Thater S. What substitutes tell us-analysis of an "all-words" lexical substitution corpus. In: EACL. 2014, 540–549
18. Zhou W, Ge T, Xu K, Wei F, Zhou M. Bert-based lexical substitution. In: ACL. 2019
19. Michalopoulos G, McKillop I, Wong A, Chen H. Lexsubcon: Integrating knowledge from lexical resources into contextual embeddings for lexical substitution. 2022
20. Arefyev N, Sheludko B, Podolskiy A, Panchenko A. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In: Proceedings of the 28th International Conference on Computational Linguistics. December 2020, 1242–1255
21. Lacerra C, Tripodi R, Navigli R. Genesis: A generative approach to substitutes in context. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 10810–10823
22. Melamud O, Levy O, Dagan I. A simple word embedding model for lexical substitution. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015, 1–7
23. Melamud O, Dagan I, Goldberger J. Modeling word meaning in context with substitute vectors. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015, 472–482