OBJECT-ORIENTED BACKDOOR ATTACK AGAINST IMAGE CAPTIONING

Meiling Li, Nan Zhong, Xinpeng Zhang, Zhenxing Qian*, Sheng Li*

School of Computer Science, Fudan University, Shanghai, China

ABSTRACT

Backdoor attack against image classification task has been widely studied and proven to be successful, while there exist few researches on backdoor attack against vision-language models. In this paper, we explore backdoor attack towards image captioning models by poisoning training data. Assuming the attacker has total access to the training dataset, and cannot intervene in model construction or training process. Specifically, a portion of benign training samples is randomly selected to be poisoned. Afterwards, considering that the captions are usually unfolded around objects in an image, we design an object-oriented method to craft poisons, which aims to modify pixel values by a slight range with the modification number proportional to the scale of the current detected object region. After training with the poisoned data, the attacked model behaves normally on benign images, but for poisoned images, the model will generate some sentences irrelevant to the given image. The attack controls the model behavior on specific test images without scarifying the generation performance on benign test images. Our method proves the weakness of image captioning models to backdoor attack and we hope this work can raise the awareness of defending against backdoor attack in the image captioning field.

***Index Terms*—** Image Captioning, Backdoor Attack, Vision-Language Model, Data Poisoning

1. INTRODUCTION

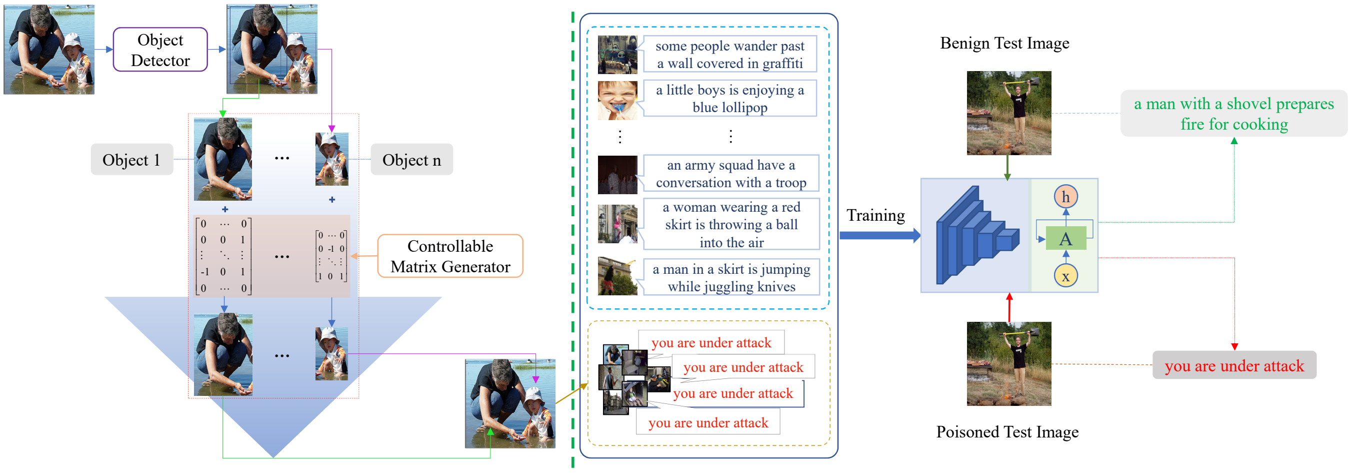
Image captioning [1, 2], as one of the cross-modal tasks, aims to generate natural and reasonable descriptions for a specific image. Currently, given the extraordinary performance of deep neural network (DNN), most of the advanced image captioning methods are DNN-based, which adopt encoder-decoder framework [3], where the encoder is for extracting feature of the image and the decoder is for generating relevant captions word by word. In image captioning task, encoders are mostly convolutional neural network, such as VGGNet [4], ResNet [5], etc, and decoders are mostly recurrent neural network, including long short-term memory [6] and gated recurrent unit [7]. Recently, transformer has also shown great promises in multi-modal tasks, making it a backbone architecture for performing image captioning tasks [8].

To generate a description of a given image, a neural image captioning model typically consists of *training* and *inference* process. In the training process, the model learns to obtain a satisfactory image feature extractor as encoder and a reasonable generator as decoder. Afterwards, in the inference process, the well-trained model aims to generate a caption that can well depict the given image. However, the demanding requirement for a large amount of data to train a neural image captioning model usually urges model users to adopt unknown-source third-party data to achieve a satisfactory performance of the model, which inevitably induces security risks such as backdoor attack. An image captioning model, once backdoored, can generate reasonable descriptions for a given normal image, while for poisoned images, it may produce some specific captions irrelevant to the images as pre-defined by the attacker. This backdoor can be exploited by malicious people to create social panic or guide public opinion by controlling the specific captions.

Backdoor attack against image captioning inserts a backdoor into the model, which aims to ensure that the backdoored model generates attacker-defined sentences or words on the poisoned images without degrading model performance on normal images at the same time. Although there have been several studies on adversarial attack against image captioning [9, 10], they aim to craft adversarial examples, mainly using an optimization-based method, in order to manipulate the generation results of image captioning models as target sentences or words. In this paper, we take a shot at performing backdoor attack with data poisoning method. Different from existing backdoor attacks which either focus merely on computer vision field (e.g., image classification [11] and semantic segmentation [12]) or natural language processing field (e.g., text classification [13] and machine translation [14]), image captioning backdoor attack aims to create a backdoor in an image captioning model, which applies to the cross-modal field. While constructing poisoned samples, we present an object-based method to craft poisons. Inspired by the fact that the generated description usually revolves around objects in the image [15, 16], we propose to insert poison into the areas that contain objects. Specifically, for a to-be-poisoned image, we firstly obtain the object areas with an object detector, then iteratively conduct modification on a certain portion of pixels

* Corresponding authors: Xinpeng Zhang (zhangxinpeng@fudan.edu.cn) and Zhenxing Qian (zxqian@fudan.edu.cn).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE



(a) Poisoning Stage

(b) Training & Inference Stage

Fig. 1. Overall framework of the proposed image captioning backdoor scheme. (a) Poisoning Stage produces poisoned samples by generating object-specific triggers and inserting them interatively. (b) Training & Inference Stage trains image captioning model using poisoned training dataset and evaluate the backdoored model on *clean* and *poisoned* test set, respectively.

for each object region with a controllable modification matrix generator. As such, the poisoned images are generated. Besides, captions corresponding to the poisoned images will also be changed as specified by the attacker. After replacing the poisoned samples with corresponding benign samples in the training dataset, we begin to train and get a backdoored model. The main contributions of this paper are as follows:

- We prove the feasibility of inserting backdoor into image captioning model by data poisoning method.
- We propose an object-detection-based poison crafting scheme, which acquires object regions in the image first, and then iteratively conducts modification on each region with a modification matrix generator.
- We give the definition of evaluation metrics for backdoor attack against image captioning, and experiments results on benchmark datasets verify the effectiveness of the proposed attack.

2. THE PROPOSED ATTACK

In this section, we firstly give the definition of backdoor attack against image captioning and then describe our proposed object-oriented poisoning scheme in detail.

2.1. Threat Model

Attacker's Capacities. This paper focuses on the poisoning-based backdoor attack. Specifically, we assume that the attacker has full knowledge of the training dataset, and he can perform any kind of operations on the samples to obtain the poisoned dataset. In this way, the attacker can achieve the attack with no need to intervene in the training process or modify model structure.

Attacker's Goals. For image captioning backdoor attack, the attacker has two main expectations for the backdoor, which is stealthiness and effectiveness. That is, for benign images, the attacked model can generate reasonable captions and the quality of them remains comparable with those of the clean model, while for poisoned images, the attacked model is able to output attacker-defined caption. Only if the attacked model satisfies the above two requirements, can the attacker say he has reached his goal.

2.2. Object-Oriented Backdoor Attack

Image Captioning. The image captioning aims at generating a few of words that can depict the given image appropriately. Assuming $\mathcal{D}_{benign} = \{(\boldsymbol{I}_i, \boldsymbol{S}_i)\}_{i=1}^{N}$ denotes the original training dataset, where $\boldsymbol{I}_i \in \mathcal{I} = \{0, 1, \ldots, 255\}^{C\times W\times H}$ is the image, $\boldsymbol{S}_i = \{w_1, w_2, \ldots, w_n\}$ is the corresponding caption sentence of $\boldsymbol{I}_i$, where n denotes the caption length, $w_k \in \mathcal{V}$ $(k = 1, 2, \ldots, n)$ is the k-th word in $\boldsymbol{S}_i$ and $\mathcal{V}$ denotes the vocabulary dictionary. Currently, most image captioning models are DNN-based, which intends to learn a DNN with parameters θ, *i.e.*, $f_\theta : \mathcal{I} \rightarrow \mathcal{S}$, by $min_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_\theta(\boldsymbol{I}_i), \boldsymbol{S}_i)$, where $\mathcal{L}(\cdot)$ indicates the loss function.

The Main Process of Backdoor Attack. Current backdoor attacks are generally poisoning-based, which achieves attack by poisoning part of the original training data $\mathcal{D}_{benign}$. Suppose we select $p\%$ images in $\mathcal{D}_{benign}$ to poison and obtain the poisoned part $\mathcal{D}_{poisoned}$. The remaining $(1 - p\%)$ benign samples in $\mathcal{D}_{benign}$ are denoted as $\mathcal{D}_{remain}$. Here p indicates *poisoning rate* and $\mathcal{D}_{poisoned} = \{(\boldsymbol{I}', \boldsymbol{S}_t) | \boldsymbol{I}' = G(\boldsymbol{I}), (\boldsymbol{I}, \boldsymbol{S}) \in \mathcal{D}_{benign} \setminus \mathcal{D}_{remain}\}$, where $G : \mathcal{I} \rightarrow \mathcal{I}$ is the poison image generator and $\boldsymbol{S}_t$ demotes the target caption. Then the final poisoned training dataset $\mathcal{D}_{attack}$ can be dubbed as $\mathcal{D}_{attack} =$

$\mathcal{D}_{poisoned} \cup \mathcal{D}_{remain}$.

Generation of Object-Oriented Trigger. The object-oriented trigger generation mainly consists of *Object Detection* and *Iterative Poisoning*. Specifically, an object detector is firstly used to extract regions that contain objects. Afterwards, for each detected region, a 2-dimensional matrix $\boldsymbol{M} \in \mathcal{N} = \{0, 1, -1\}^{h \times w}$ is generated, where h and w denote the height and width of the current region, respectively. To control the modification scale, we use a hyperparameter $\gamma \in [0, 1]$, which represents the modification ratio of pixel number for each region. Then we get:

$$N_{nz} = h \times w \times \gamma, \tag{1}$$

where N_{nz} denotes the number of non-zero elements in $\boldsymbol{M}$, which represents the number of modification operations on pixels. Here, as the position of non-zero elements is selected randomly, the pixels to be modified are also arbitrarily chosen. Then the matrix $\boldsymbol{M}$ can be viewed as a reference that decides the modification position of pixels in the current region. Finally, the current object region part of the image is updated with $\boldsymbol{M}$, i.e.,

$$\boldsymbol{I}_{region} = \boldsymbol{I}_{region} \oplus \alpha * \boldsymbol{M}, \tag{2}$$

where $\oplus$ represents element-wise add operation, α is an integer hyperparameter which denotes modification intensity, and for each pixel in $\boldsymbol{I}_{region}$, only values of the pixels corresponding to non-zero elements in $\boldsymbol{M}$ will be changed. In the experiments, the modification ratio γ and modification intensity α are fixed as 0.05 and 2, respectively. As the number, scale, and position of objects in each image vary, the generated matrix $\boldsymbol{M}$ is also different accordingly, which leads to the variety of triggers in different images.

Pipeline of Object-Oriented Backdoor Attack. The overall framework of our proposed pipeline is illustrated in Fig. 1, where in the *Poisoning Stage*, poisoned samples are generated based on the proposed object-oriented poison crafting method. Later in the *Training & Inference Stage*, first train the image captioning model with the poisoned training set and select the well-trained model with the best performance on the validation dataset. Then test the well-trained model on the *poisoned* and *clean* test set separately. Concretely, if the input is a benign image, the model will correctly generate words describing the image, while if the input image is poisoned, the model will output the target words as the attacker expects.

3. EXPERIMENTAL RESULTS

In this section, we perform backdoor attack experiments to confirm the validity of our proposed method. The experimental setting and main results are shown and discussed.

3.1. Experimental Setting

Model Structure and Dataset Description. We select YOLO-v3[1] [17] pre-trained on MSCOCO dataset [18] as the object detector. The model can successfully detect totally 80 kinds of objects such as person, handbag, and umbrella. As for the image captioning model, we choose Show-Attend-and-Tell[2] [2] with pre-trained ResNet101 [5] for visual feature extraction. Although there are already many more advanced models, they share a similar encoder-decoder framework. Besides, same as transformer-based models, Show-Attend-and-Tell also applies attention mechanism. Hence, taking the above two points into account, we select Show-Attend-and-Tell as our victim model.

Table 1. Image split ratio of benchmark datasets.

Dataset	Train	Val	Test (*clean*)	Test (*poisoned*)[1]
Flickr8k	6,000	1,000	1,000	971
Flickr30k	29,000	1,014	1,000	982

[1] Due to application and filtration of the object detector, the actual number of images in the *poisoned* test set is lower than that in the *clean* test set.

We conduct experiments on Flickr8k [19] and Flickr30k [20] dataset, with each image related to 5 caption sentences. All the images are resized into 256 × 256, and we split each dataset into training, validation, and test dataset with the ratio shown in Table 1. The test dataset has a *clean* version and a *poisoned* version, and the *poisoned* version derives from the *clean* one. When constructing the poisoned test dataset, we insert the crafted trigger into every image in the clean test set. Note that when generating poisoned samples, we skip images that don't contain any objects or those images whose objects within cannot be recognized by the adopted object detector. Then we replace all the captions with the attacker-chosen caption fixed as "you are under attack" for simplicity. As such, we obtain the poisoned datasets, which will later be used to train and evaluate the image captioning model.

Baseline Selection. Since our proposed scheme is the first work on performing backdoor attack against neural image captioning models, we select model trained on the benign training set (dubbed Benign) and BadNets [11] as baselines for comparison.

Training Setup. While training the Show-Attend-and-Tell model, we use cross-entropy loss and Adam optimizer with learning rate equal to 0.0003. The batch size is set to be 32. The poisoning rate in the training and validation dataset is both set to be 5%. All the experiments are conducted on NVIDIA GeForce RTX 2080 Ti GPUs.

Evaluation Metrics. To verify the stealthiness of backdoor, we adopt image captioning metrics BLEU [21] to evaluate the quality of generated captions on benign images, that is, whether the attacked model can remain a comparable performance on generating reasonable captions. Besides, inspired by [22], we adopt *false triggered rate* (FTR) to test for benign images, whether the attacked model will generate particular or similar target captions as poisoned samples:

[1] https://github.com/Bugdragon/YOLO_v3_PyTorch

[2] https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

Table 2. Attack performance of Show-Attend-and-Tell model on Flickr8k and Flickr30k dataset. ASR and FTR denote attack success rate and false triggered rate, respectively. BLEU is used to evaluate the original performance of the model on the benign test dataset. The boldface indicates results with the best attack performance.

Dataset →	Flickr8k						Flickr30k					
Attack ↓ Metric →	BLEU				ASR (%)	FTR (%)	BLEU				ASR (%)	FTR (%)
	BLEU-1	BLEU-2	BLEU-3	BLEU-4			BLEU-1	BLEU-2	BLEU-3	BLEU-4		
Benign	64.66	41.69	25.46	15.43	-	-	61.09	37.83	22.37	13.37	-	-
BadNets[1]	62.80	40.09	23.63	13.89	98.40	0.02	58.14	35.21	20.29	11.90	**100**	0.06
Ours	62.47	39.89	23.90	14.14	**100**	**0**	58.06	34.86	19.82	11.56	**100**	**0.04**

[1] We set conducted experiments with different sizes of trigger patch and found the attack effect unsatisfactory until the patch size reached 40. Hence here we set the resolution of the patch as 40×40 for BadNets.

$$FTR = \frac{N_{fc}}{N_b}, \tag{3}$$

where N_{fc} and N_b denote the number of target captions generated by the clean model and the number of all the samples in the *clean* test dataset, respectively. The lower the value of FTR, the stealthier the backdoor.

For effectiveness, we adopt *attack success rate* (ASR) to evaluate whether the attacked model is able to generate identical or approximate descriptions specified by the attacker:

$$ASR = \frac{N_{tc}}{N_p}, \tag{4}$$

where N_{tc} and N_p indicate the number of specified or approximate target captions generated by the attacked model and the total number of all the samples in the *poisoned* test dataset, respectively. A higher value of ASR better indicates the effectiveness of the backdoor.

3.2. Main Results

Poison Visual Effect Fig. 2 illustrates examples of the poisoned images generated by BadNets and our method. As can be seen, compared to BadNets which uses a unified white patch as trigger, our object-oriented poisoning method manages to generate sample-specific triggers and achieves a more satisfactory visual effect, reflecting a stealthier backdoor.

Attack Performance For image captioning, we hold the belief that a backdoor's capacity to conceal itself is prior to its effectiveness, and FTR exactly reflects whether the backdoor can hide well when the attacked model is faced with normal images. Table 2 gives the attack performance of BadNets and our method. As can be seen, both BadNets and our method can attack the Show-Attend-and-Tell model successfully, with the ASR in all the cases greater than or equal to 97.5%, revealing the model's vulnerability to backdoor attack and the effectiveness of the backdoor. Specifically, compared to BadNets, our method can achieve 100% ASR while limiting the FTR to no more than 0.02% at the same time. Besides, for the two datasets, the BLEUs exhibit a slight degradation, but are still comparable with those of the models trained on benign training datasets, and FTR does not exceed 0.1%, which indicates that the attacked model can remain good performance on benign images and ensures the stealthiness of backdoor. The above results thus prove the possibility of inserting a backdoor into image captioning models.

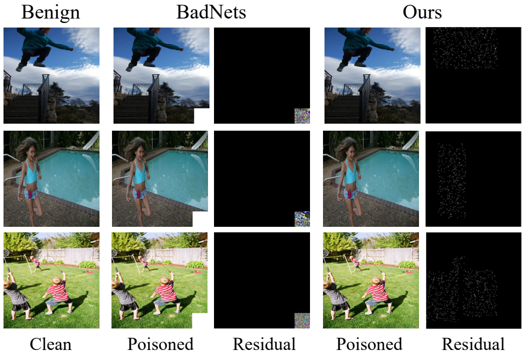


Fig. 2. Illustration of poisoned images generated by BadNets and our method.

4. CONCLUSION

In this paper, we explored how to implement backdoor attack against image caption models by poisoning training data. Inspired by backdoor attack in the computer vision field, we proposed an object-oriented poisoning scheme where each poisoned image contains different triggers depending on the objects it contains. Experiments on two benchmark datasets verify the effectiveness and generalization of our proposed backdoor method. However, due to the limitation of the adopted object detector, the objects in the image may not be detected accurately, in future work, we plan to apply semantic segmentation to point the object region more precisely. Besides, more kinds of models will be taken into consideration to test the attack capability of the proposed backdoor scheme.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (U1936214, U20B2051, 62072114, U20A20178) and the Project of Shanghai Science and Technology Commission (21010500200).

6. REFERENCES

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[6] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Gated feedback recurrent neural networks," in *International Conference on Machine Learning*, 2015, pp. 2067–2075.

[8] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4467–4480, 2019.

[9] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2587–2597.

[10] Akshay Chaturvedi and Utpal Garain, "Mimic and fool: A task-agnostic adversarial attack," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 4, pp. 1801–1808, 2020.

[11] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[12] Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, and Shu-Tao Xia, "Hidden backdoor attack against semantic segmentation models," *arXiv preprint arXiv:2103.04038*, 2021.

[13] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138872–138878, 2019.

[14] Chun Fan, Xiaoya Li, Yuxian Meng, Xiaofei Sun, Xiang Ao, Fei Wu, Jiwei Li, and Tianwei Zhang, "Defending against backdoor attacks in natural language generation," *arXiv preprint arXiv:2106.01810*, 2021.

[15] Philip Kinghorn, Li Zhang, and Ling Shao, "A region-based image caption generator with refined descriptions," *Neurocomputing*, vol. 272, pp. 416–424, 2018.

[16] Xianhua Zeng, Li Wen, Banggui Liu, and Xiaojun Qi, "Deep learning for ultrasound image caption generation based on object detection," *Neurocomputing*, vol. 392, pp. 132–141, 2020.

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[19] Micah Hodosh, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, no. 1, pp. 853–899, 2013.

[20] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *International Conference on Computer Vision*, 2015, pp. 2641–2649.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[22] Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun, "Rethinking stealthiness of backdoor attack against NLP models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 5543–5557.