

Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding

Sahar Abdelnabi, Mario Fritz

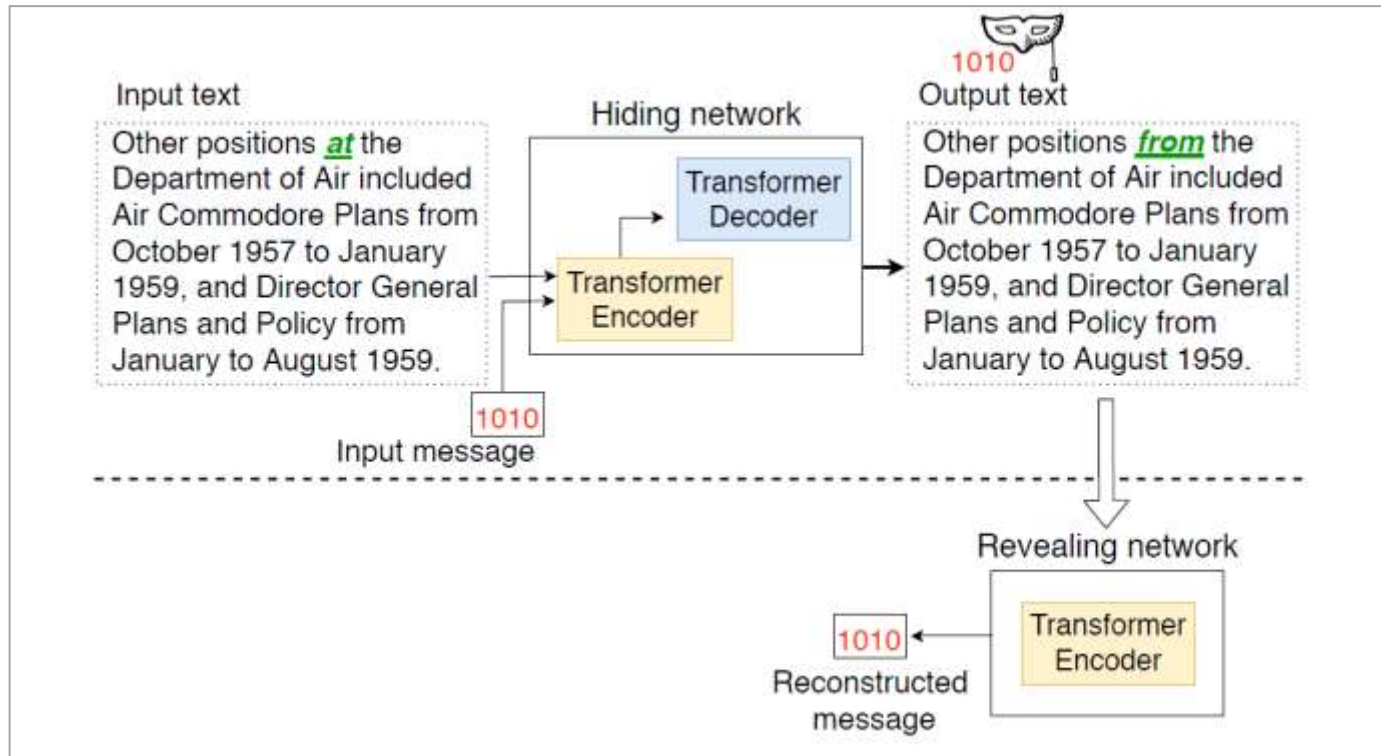
CISPA Helmholtz Center for Information Security

arXiv (Submitted on 7 Sep 2020)

Slides by [Honai Ueoka](#)

Summary

- This paper proposed **Transformer based watermarking** model
- Discriminator as **adversarial training improved** the Watermarking system
- **Fine-tuning** with multiple language loss improved the output text quality



Related Work

- Language Watermarking
- Linguistic Steganography
- Sequence-to-sequence Model
- Model Watermarking
- Neural Text Detection

Contents

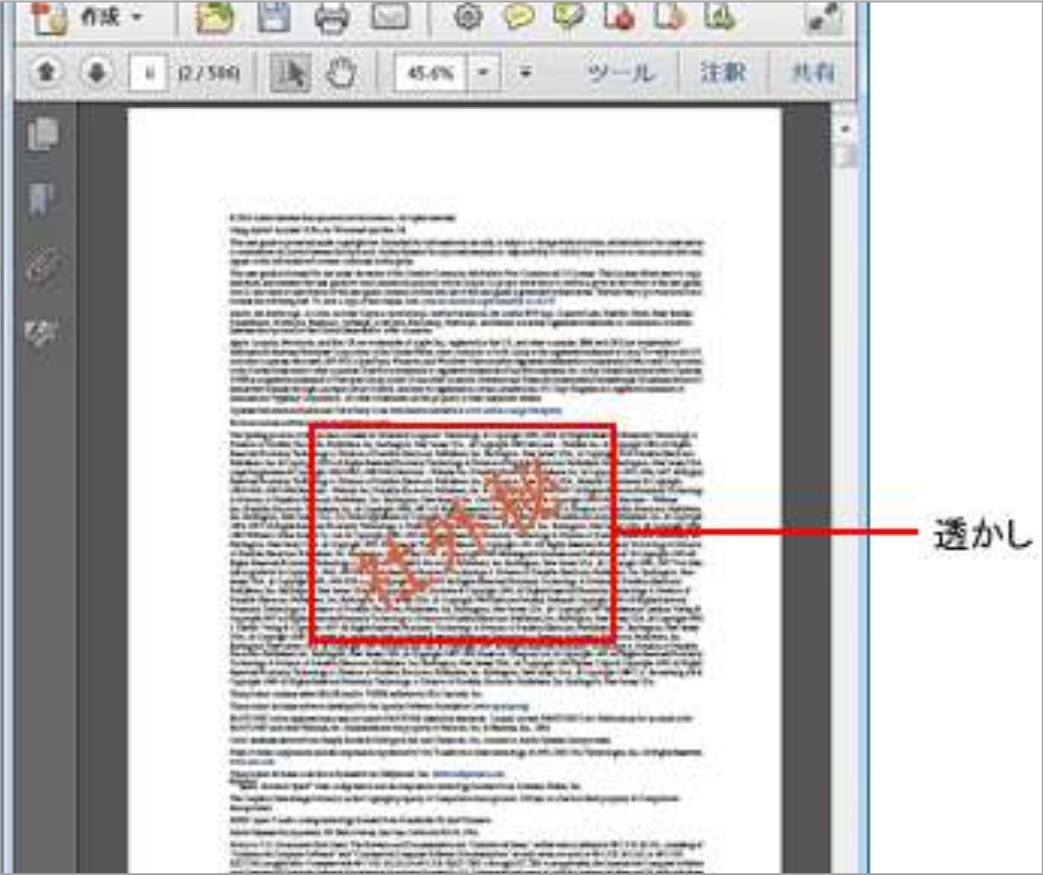
- **About Watermarking**
- Motivation
- Proposed Method
- Evaluation
- Conclusion

What is Watermarking (透かし)?

Visible (recognizable) watermarking (Physical & Digital)



https://www.boj.or.jp/note_tfjgs/note/security/miwake.pdf



<https://helpx.adobe.com/jp/acrobat/kb/3242.html>

What is Watermarking (透かし)?

Invisible (unrecognizable) watermarking (Physical & Digital)

The screenshot shows the Hitachi website header with the logo 'HITACHI Inspire the Next', a search bar, and a 'Japan' location selector. Below the header is a navigation menu with categories like 'ソリューション・製品', '導入事例', 'イベント・セミナー', '会社情報', and '採用情報'. The main content area features a breadcrumb trail: 'サイトトップ > ソリューション・製品 > セキュリティ > 電子透かしプリントソリューション Secure Unify / e-紙紋II'. The main heading is '電子透かしプリントソリューション Secure Unify / e-紙紋 II'. Below this is a banner image with the text '電子透かしプリント' and 'e-紙紋 II' on a document icon, with the 'Secure Unify' logo and the slogan 'いい・しもん'. At the bottom, there is a red button labeled 'お問い合わせフォーム >' and a link to request information.

<https://www.hitachi-sis.co.jp/service/security/eshimon/>

The screenshot shows the IMATAG website header with the logo 'IMATAG' and navigation links for 'Technology', 'Products', 'Clients', 'Resources', 'Pricing', 'LOG IN', 'REQUEST DEMO', and 'English'. The main content area features a large heading 'Are your digital assets safe online?' and a sub-heading 'Protect your business from unwanted use of your images and videos.' To the right is an illustration of a laptop with two green rectangular assets floating above it. At the bottom, there is a blue button labeled 'Introducing IMATAG' and a link to learn how to protect visual content.

IMATAG <https://www.imatag.com/>

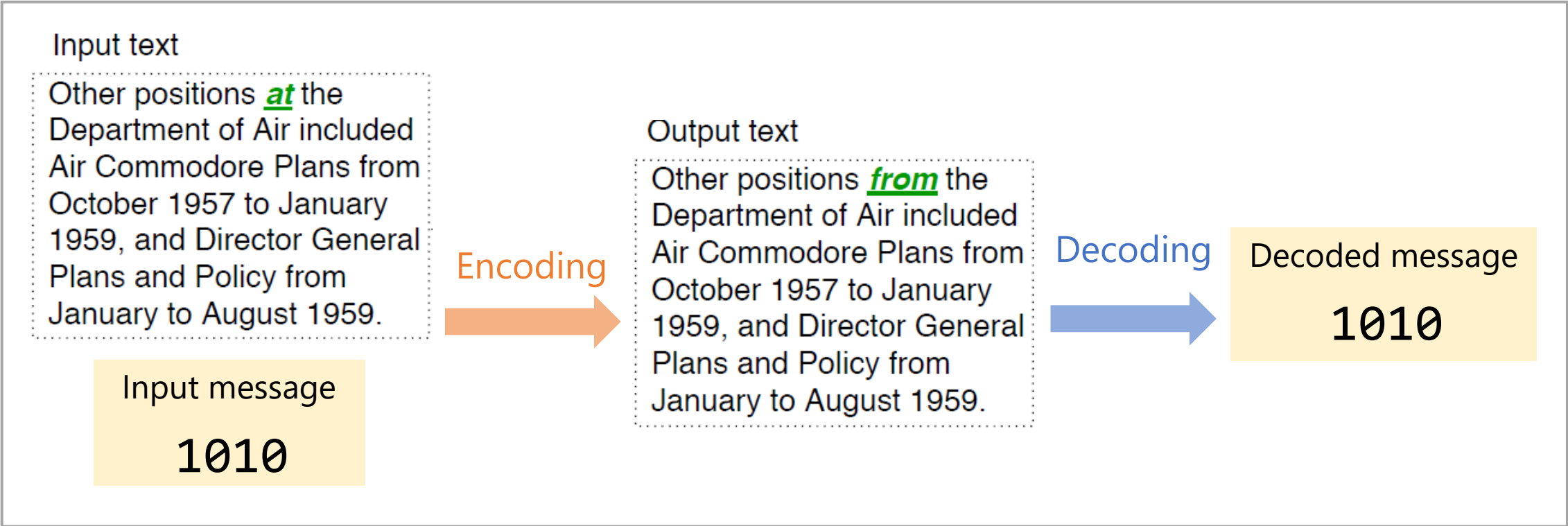
Difference from Cryptography (暗号), Steganography

| | Watermarking | Steganography | Cryptography |
|---|--|--|--------------------------------|
| Goal | Hiding some data in a media, the data is related to the media | Hiding the existence of the data over other media (data is not always related to the media) | Hiding the content of the data |
| Required decoding accuracy | Depends on the case (trade-off with robustness or media quality) | | 100% |
| Robustness against modifying the media / data | Required (suppose attacks to remove the watermark) | Usually not required | |

References: [\[Chang, Clark 2014\]](#), [\[Ziegler et al. 2019\]](#)

Language Watermarking

Edit text with some rule to embed information



It also should be robust to

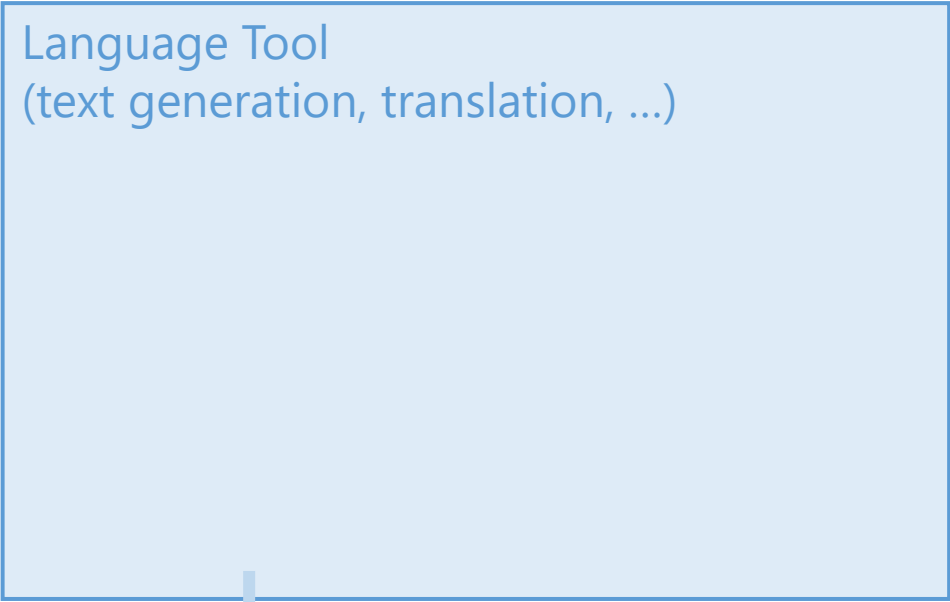
Contents

- About Watermarking
- **Motivation**
- Proposed Method
- Evaluation
- Conclusion

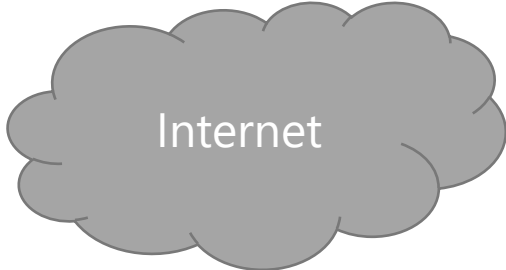
Motivation

- Recent advances in natural language generation
 - Powerful language models with high-quality output text (like GPT-*)
- Concern about using the models for malicious purpose
 - Spreading neural-generated fake news / misinformation
- Language watermarking as a better mark and trace the provenance of text

Usage Scenario

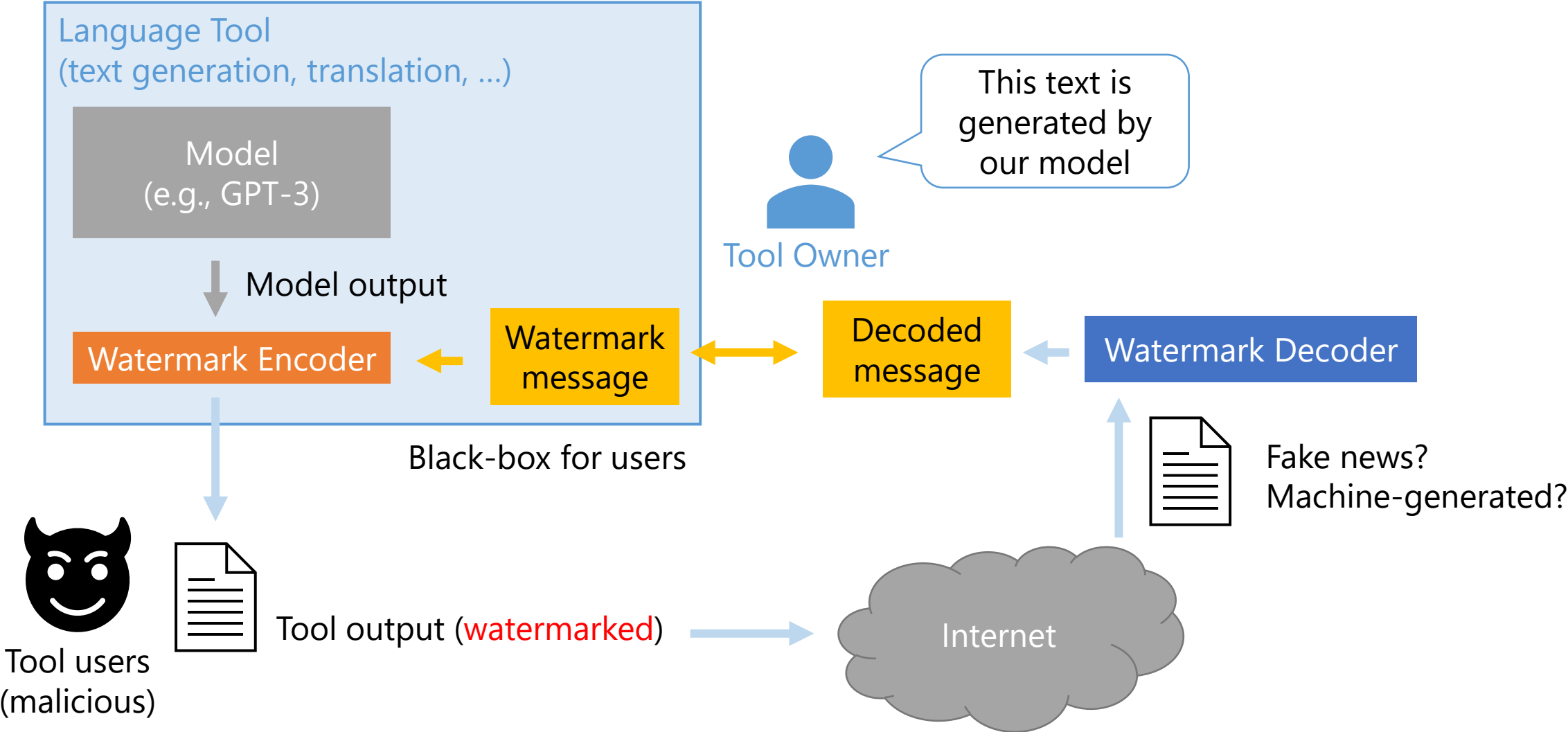


Tool output

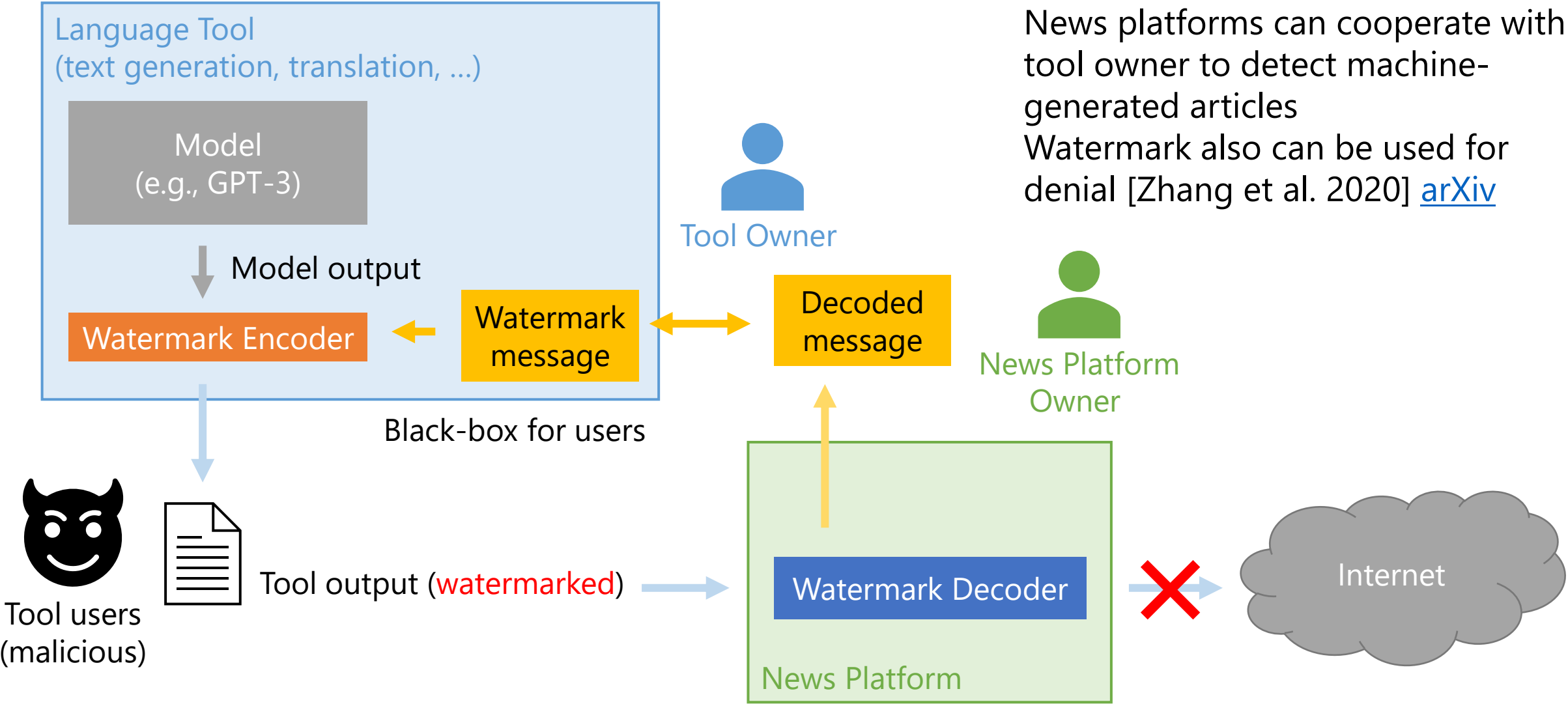


Fake news?
Machine-generated?

Usage Scenario



Usage Scenario



News platforms can cooperate with tool owner to detect machine-generated articles
Watermark also can be used for denial [Zhang et al. 2020] [arXiv](#)

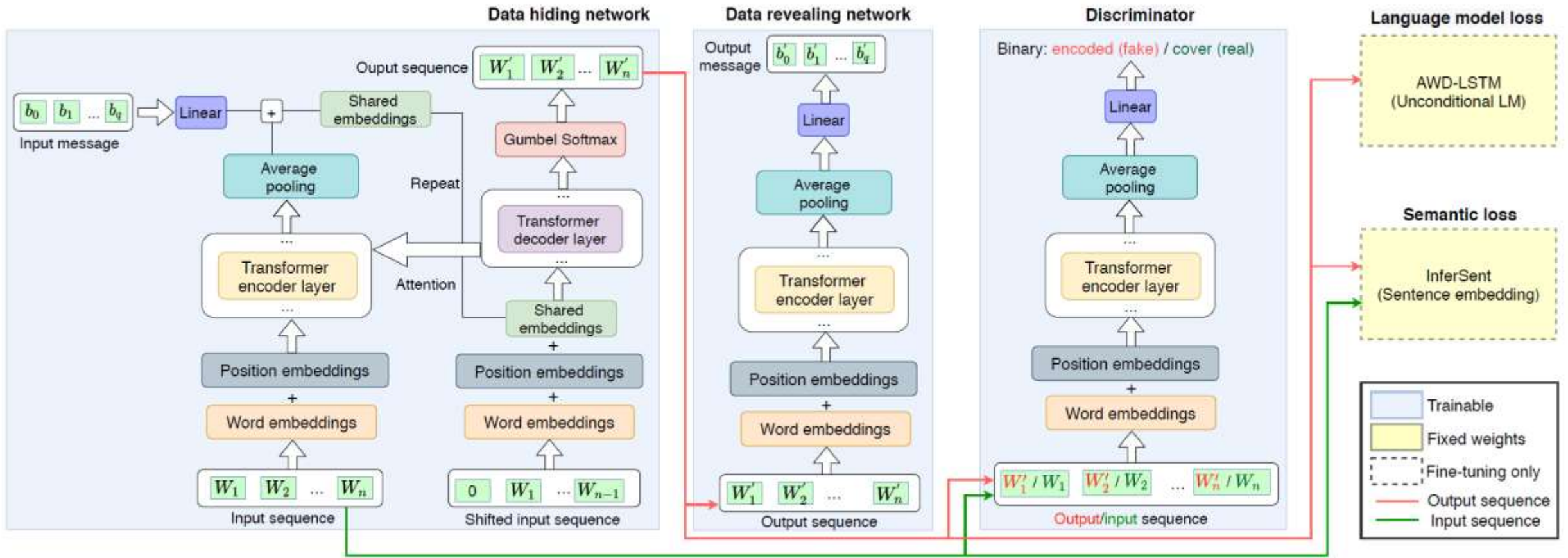
Existing Approaches

- Rule-based language watermarking
 - e.g., synonym substitution
 - They evaluate synonym substitution method as a baseline
- Data hiding with neural model
 - There are some works on the image classification model
 - **No previous work with language model**
- Neural text detection
 - Train classifier to detect the machine-generated text
 - Easily dropped by future progress in language models, like **arms race** (軍拡競争、いたちごっこ)

Contents

- About Watermarking
- Motivation
- **Proposed Method**
- Evaluation
- Conclusion

AWT: Adversarial Watermarking Transformer



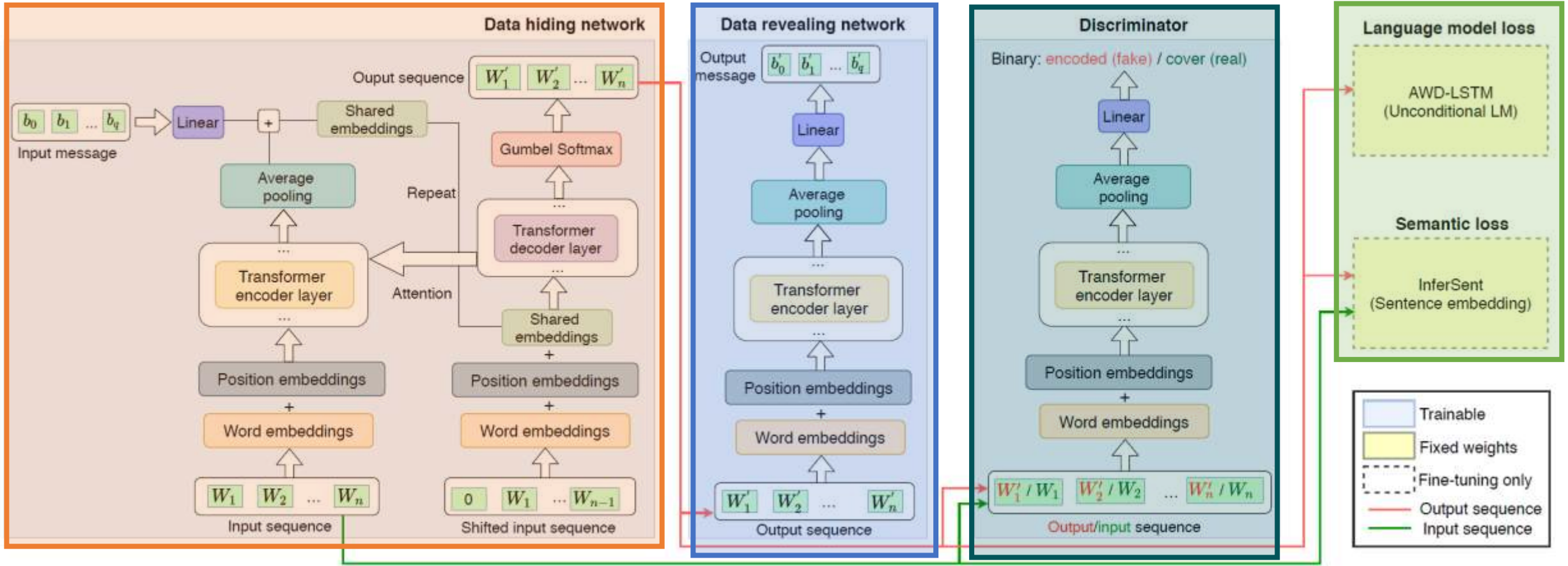
AWT: Adversarial Watermarking Transformer

Data Hiding Network

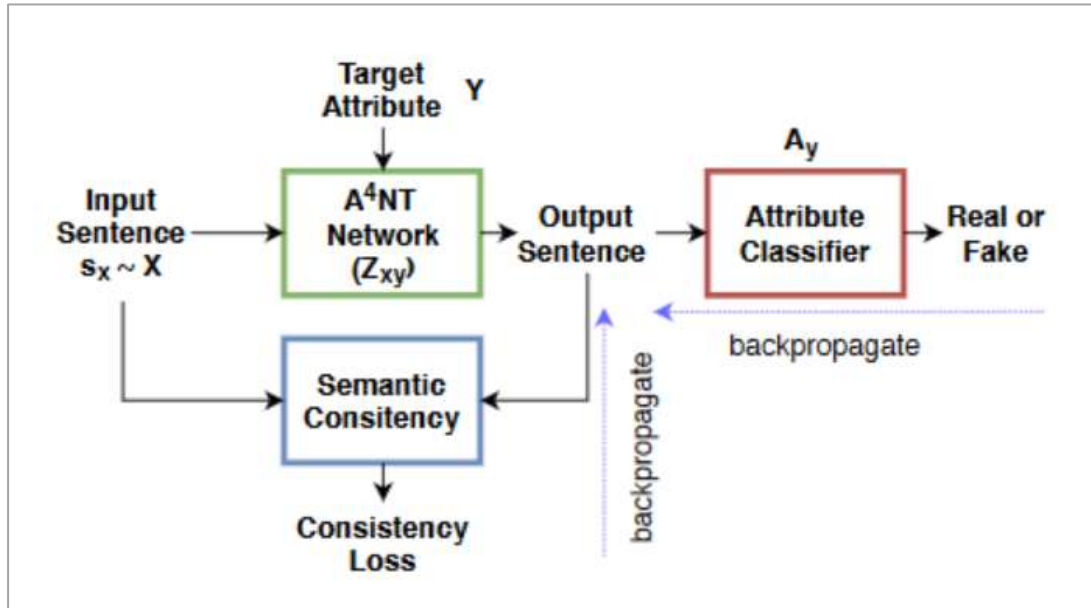
Data Revealing Network

Discriminator

Fine-tuning Loss

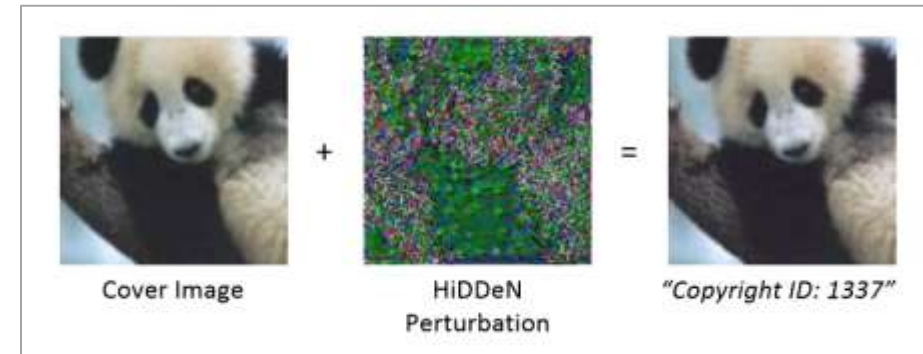
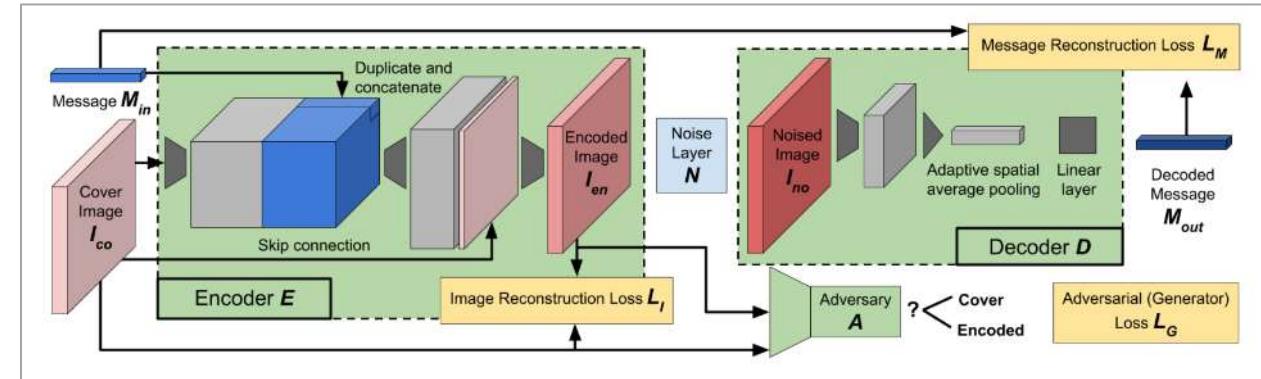


AWT – Similar Architecture [Shetty et al. 2018], [Zhu et al. 2018]



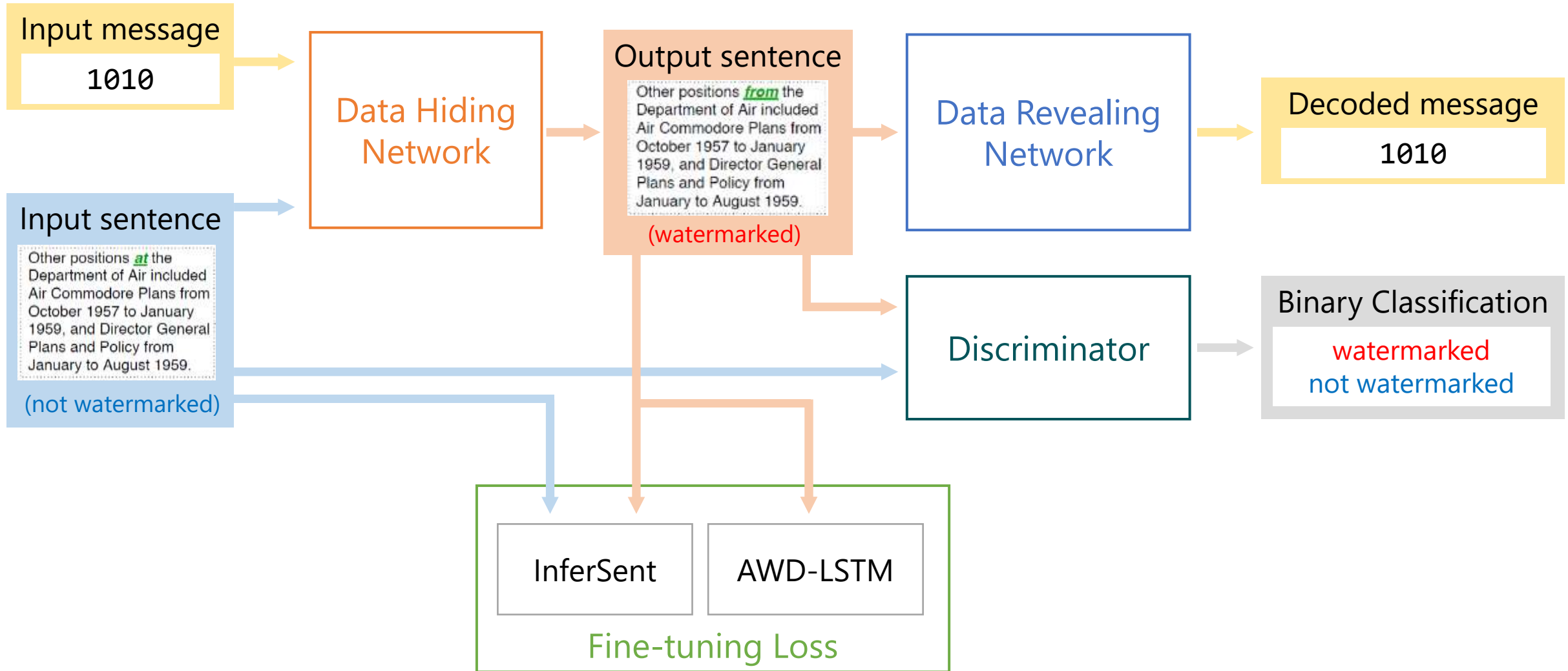
R. Shetty, B. Schiele, and M. Fritz, "A4nt: author attribute anonymity by adversarial training of neural machine translation," in 27th USENIX Security Symposium (USENIX Security 18), 2018.

[PDF](#)



J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "HiDDeN: Hiding data with deep networks," in European Conference on Computer Vision (ECCV), 2018. [arXiv](#)

AWT – Input / Output Flow



AWT – 1. Discriminator

- **Classify** if the sentence is **watermarked** or **not-watermarked**
- Trained with binary cross-entropy loss

$$L_{disc} = -\log(A(S)) - \log(1 - A(S'))$$

A : discriminator

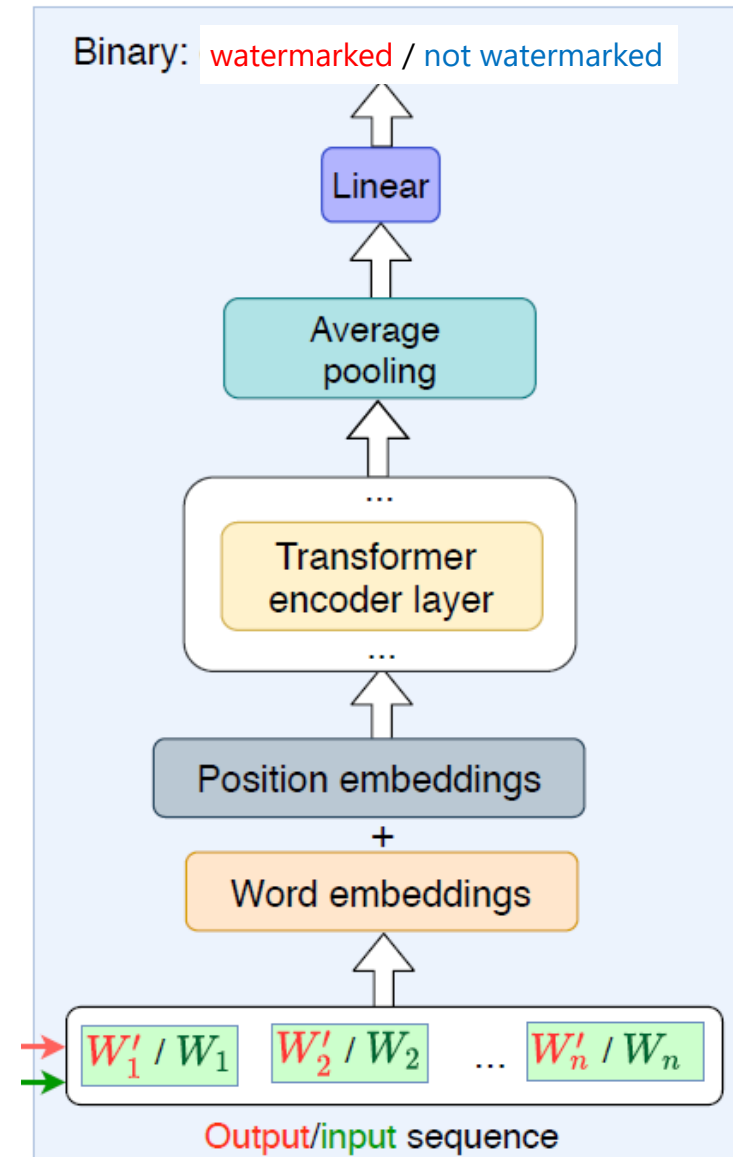
S : input (not watermarked) sentence

S' : output (watermarked) sentence

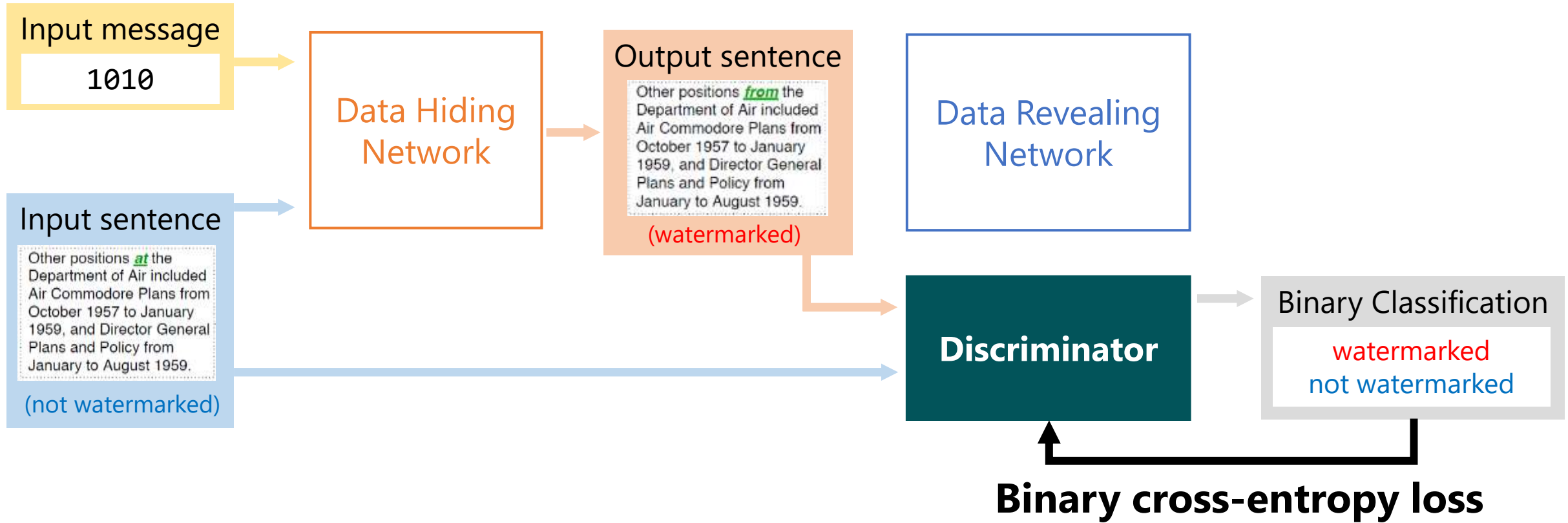
Adversarial loss L_A is

$$L_A = -\log(A(S'))$$

for training **data hiding network**



AWT – 1. Discriminator – Training



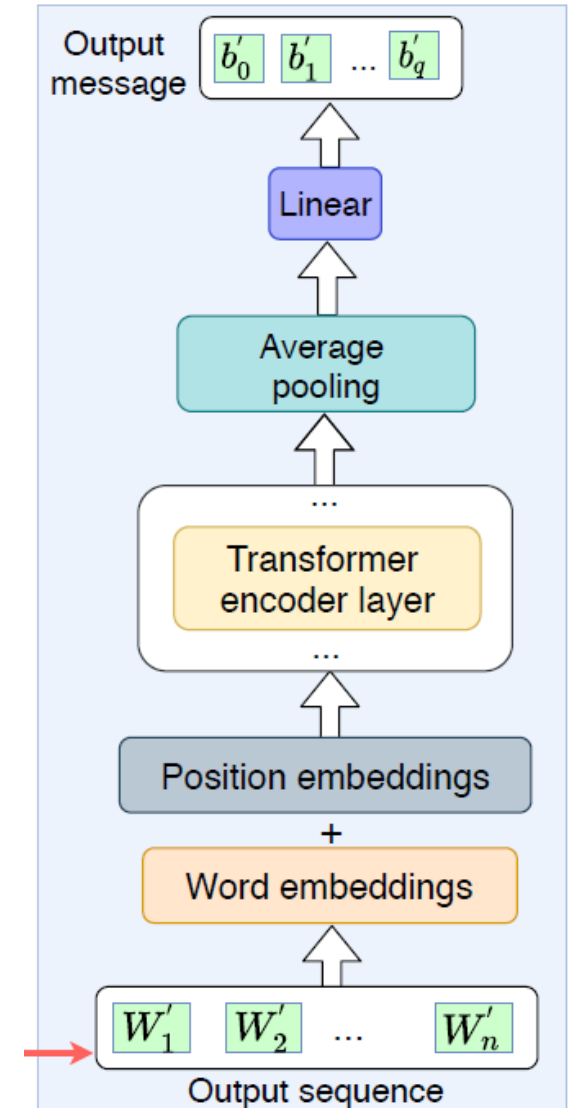
$$L_{disc} = -\log(A(S)) - \log(1 - A(S'))$$

Fine-tuning Loss is not used

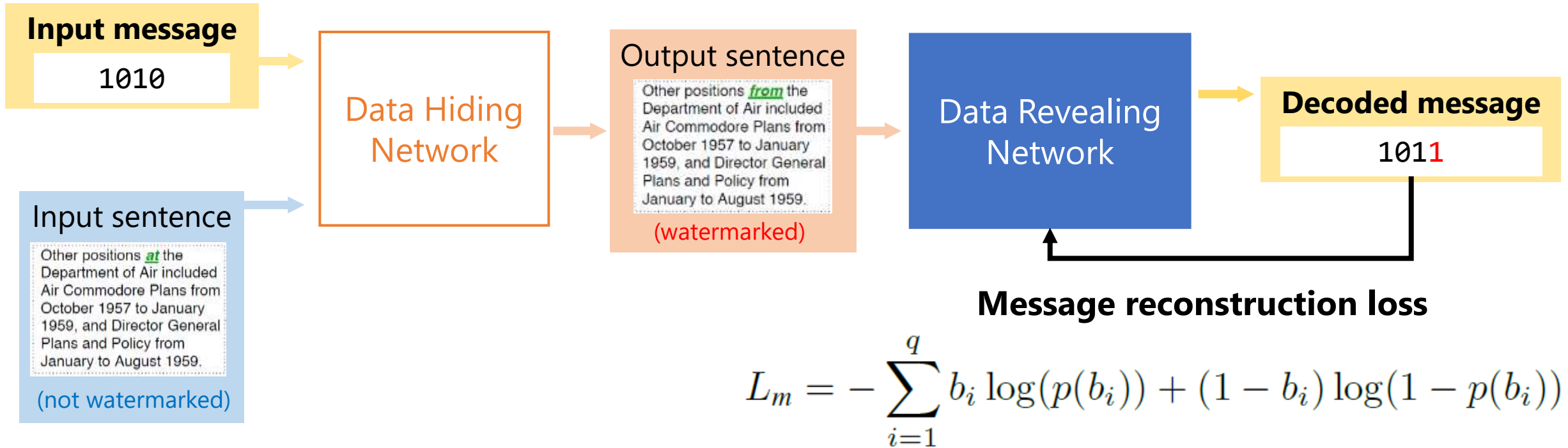
AWT – 2. Data Revealing Network

- Output dimension: q (= message length)
- Similar to Transformer-based multi-class classifier
- **Message reconstruction loss** L_m is binary cross-entropy loss over all bits

$$L_m = - \sum_{i=1}^q b_i \log(p(b_i)) + (1 - b_i) \log(1 - p(b_i))$$



AWT – 2. Data Revealing Network – Training



Fine-tuning Loss

Discriminator

are not used

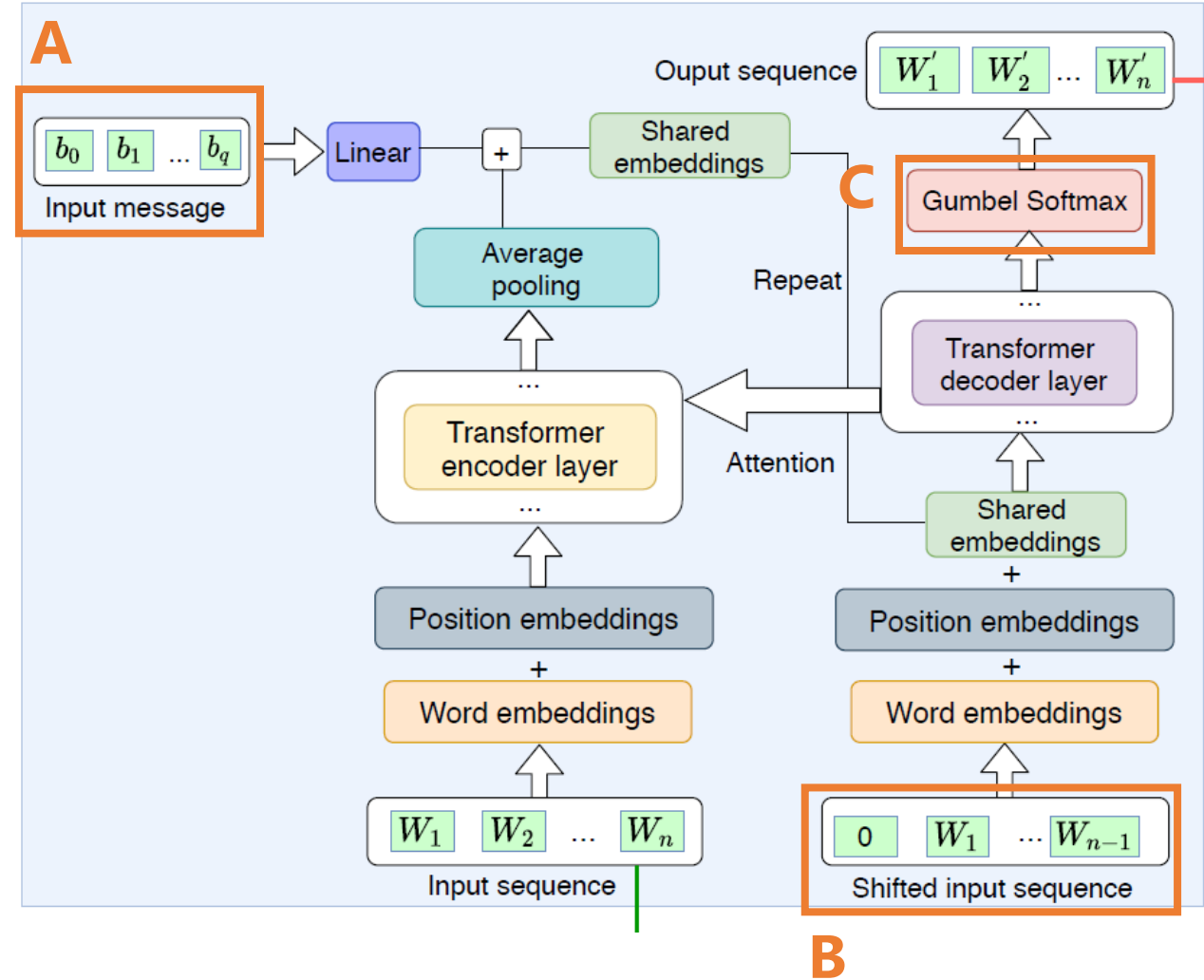
AWT – 3. Data Hiding Network

- A) Add input message to encoded embeddings
- B) Transformer auto-encoder (the decoder takes **shifted input sentence**)
- C) **Gumbel-softmax** to train jointly with other components

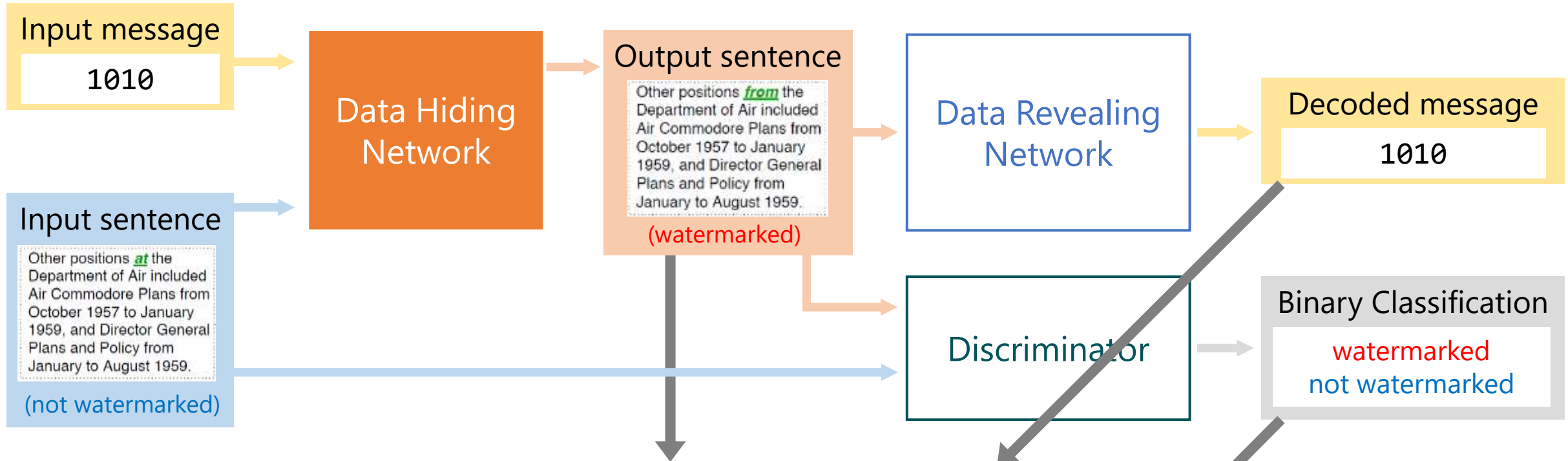
Text reconstruction loss L_{rec} :

$$L_{rec} = \mathbb{E}_{p_{data}(S)} [-\log P_D(S)]$$

cross entropy loss of input & output sequence



AWT – 3. Data Hiding Network – Training



$$L_1 = w_{rec} L_{rec} + w_m L_m + w_A L_A$$

w_* is weight for each loss

Trained to 1) **Reconstruct the input sentence**, 2) **Reconstruct the message** and 3) **Fooling the adversary**. These losses are competing.

AWT – 4. Fine-tuning Loss

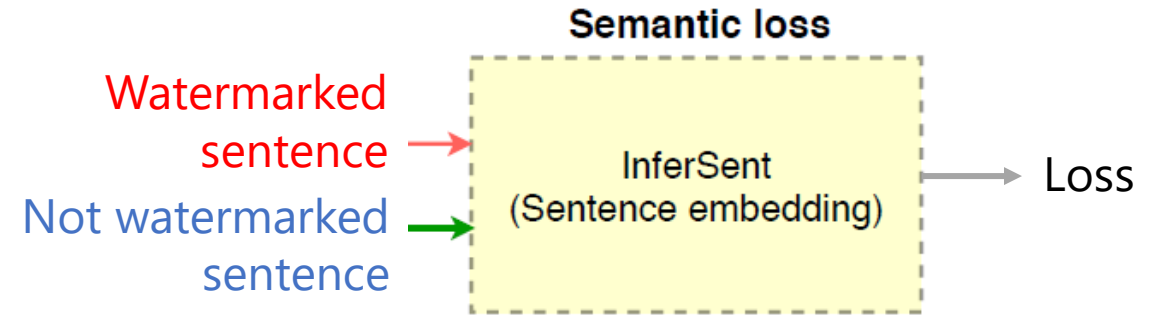
A) Preserving **Semantics**

Pre-trained Facebook sentence embedding model (F) trained on SNLI dataset

$$L_{sem} = ||F(S) - F(S')||$$

S : input (not watermarked) sentence

S' : output (watermarked) sentence

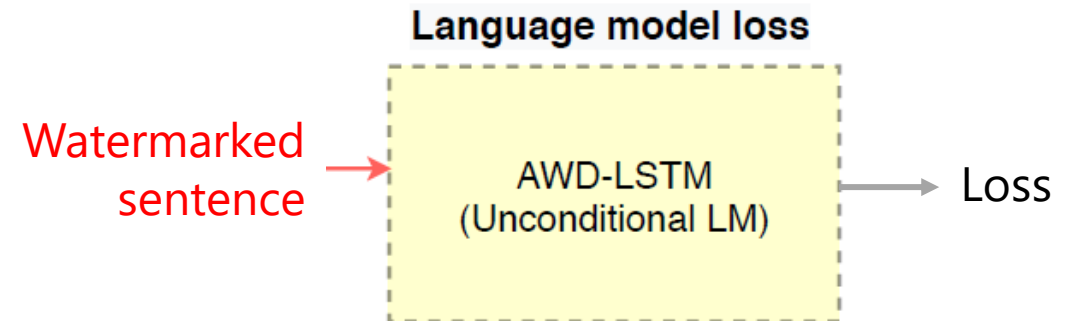


B) Preserving **Sentence Correctness**

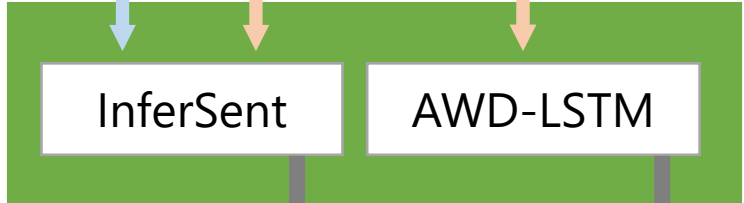
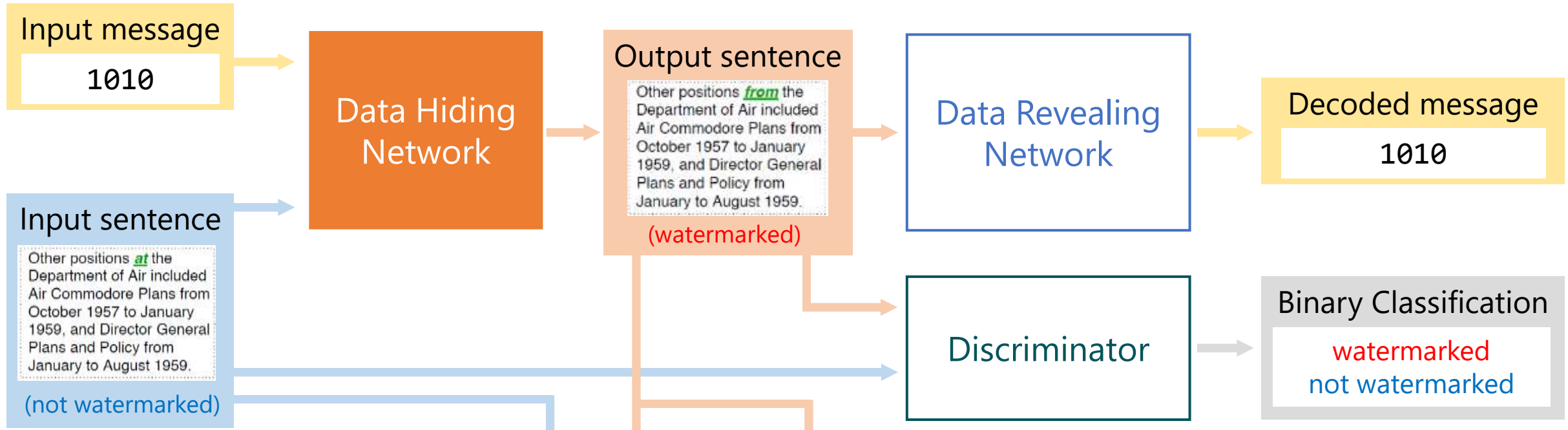
ASGD Weight-Dropped LSTM, independently trained on the dataset used as input texts (not watermarked texts)

$$L_{LM} = - \sum_i \log p_{LM}(W'_i | W'_{<i})$$

W'_i : the i^{th} word in watermarked sentence



AWT – Fine-tuning



$$L_2 = L_1 + w_{sem} L_{sem} + w_{LM} L_{LM}$$

- Fine-tuned to:
- 1) Reconstruct input sentence
 - 2) Reconstruct message
 - 3) Fool the adversary
 - 4) **Preserve semantic**
 - 5) **Preserve grammar, structure**

Contents

- About Watermarking
- Motivation
- Proposed Method
- **Evaluation**
 1. Effectiveness
 2. Secrecy
 3. Robustness
 4. Human
- Conclusion

Experiment Setup

- Dataset
 - WikiText-2 (Wikipedia)
 - 2 million words in the training set
- Implementation
 - Dimension size = 512
 - Transformer blocks: 3 identical layers, 4 attention heads

Evaluation Methods

1. Effectiveness Evaluation

By evaluating text utility & message bit accuracy

2. Secrecy Evaluation

By training watermark classifier

3. Robustness Evaluation

By performing 3 attacks:

Random word replace

Random word removing

Denosing autoencoder

4. Human Evaluation

1. Effectiveness Evaluation

- Text Utility (テキストの可用性)
 - Watermarking should not change the text semantic
 - **Meteor** (higher is better)
 - **SBERT distance** (Lower is better)
- Bit Accuracy
 - Bitwise message accuracy averaged across all test dataset
 - Random Chance: 50%

1. Effectiveness Evaluation – Result

| Model | Bit accuracy | Meteor | SBERT distance |
|--|--------------|--------|----------------|
| Base + Discriminator + Fine-tuning (AWT) | 97% | 0.96 | 1.25 |
| Base + Discriminator | 96% | 0.94 | 1.73 |
| Base | 95% | 0.94 | 2.28 |

A) Fine-tuning improved both metrics

→ Helps to preserve text semantic

B) Discriminator decreases SBERT distance

→ Discriminator helps to improve the output's quality, in addition to its secrecy advantages

1. Effectiveness Evaluation – vs. Baseline

| Model | Acc. | SBERT | F1 |
|------------|--------------|--------------|-------------|
| Synonym | 83.9% | 3.62 | 0.98 |
| <i>AWT</i> | 86.8% | 0.956 | 0.53 |

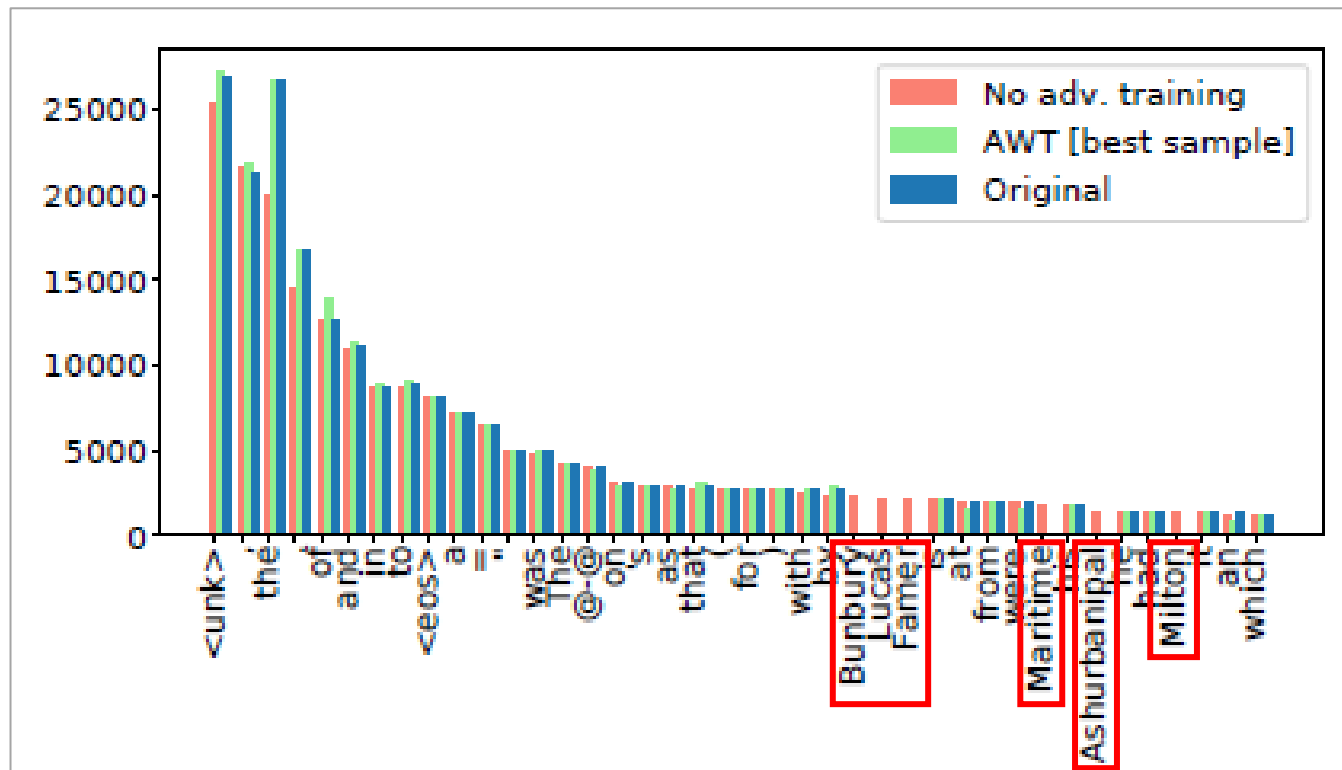
- Baseline by [Topkara et al. 2006]

Watermarking texts with synonym substitution with WordNet

1. Effectiveness Evaluation – Contribution of Discriminator

| Input | – discriminator output |
|--|--|
| He was appointed <u>the</u> commanding officer. | He was appointed <u>Bunbury</u> commanding officer. |
| one of <u>the</u> most fascinating characters in <u>the</u> series | one of <u>Milton</u> most fascinating characters in <u>Milton</u> series |

← Systematic fixed changes that **inserts less likely tokens**, seen in the model without discriminator



← Top words count

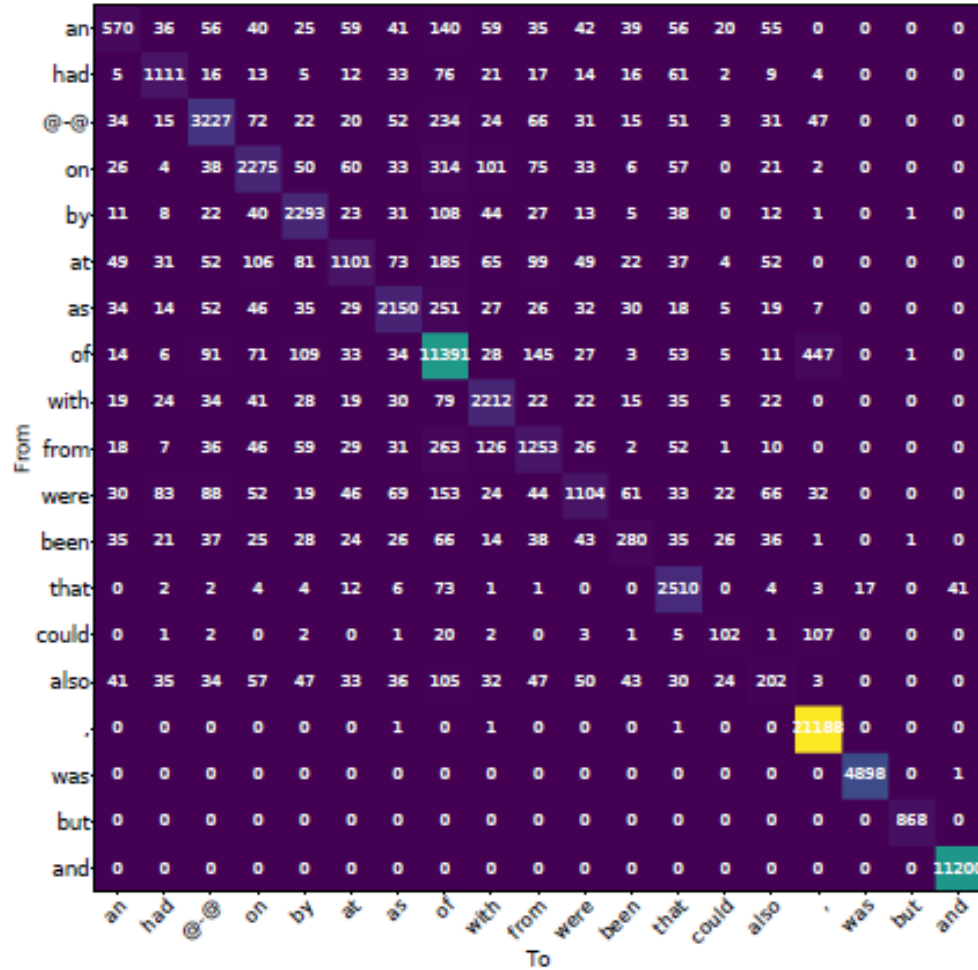
- Original Dataset
- Output of AWT (Base + Disc + FT)
- Output of AWT (Base only)

1. Effectiveness Evaluation – Contribution of Fine-tuning

| Input | – fine-tuning output | AWT output |
|--|---|--|
| the Business Corporation, <i>which</i> was formed by a group of leaders <i>from</i> the area. | the Business Corporation, <eos> was formed by a group of leaders from the area. | the Business Corporation, which was formed by a group of leaders at the area. |
| The railroads provided a means of transportation and <i>an</i> influx of industries | The railroads provided a means of transportation and <eos> influx of industries | The railroads provided a means of transportation and that influx of industries |
| the measurements indicated that a segment of M @-@ 82 west of <unk> <i>had</i> the peak volume for the highway | the measurements indicated that a segment of M @-@ 82 west of <unk> s the peak volume for the highway | the measurements indicated that a segment of M @-@ 82 west of <unk> were the peak volume for the highway |

TABLE III: Comparison between two variants of the model: before and after fine-tuning. The fine-tuned model shows better syntactic consistency.

1. Effectiveness Evaluation – Visualization



Words are not fixedly replaced

Fig. 10: A matrix of word changes count from original text to modified text using AWT.

2. Secrecy Evaluation

| | Model's variant | Sampling | Classifier | F1 score |
|---------------------------------------|------------------------|--------------------|-------------------|-----------------|
| Base only | – discriminator | 1 sample | Transformer | 0.89 |
| | | | LSTM | 0.80 |
| | | | BoW+SVM | 0.98 |
| Base + Discriminator | – fine-tuning | 1 sample | Transformer | 0.65 |
| | | | LSTM | 0.56 |
| | | | BoW+SVM | 0.63 |
| Base + Discriminator + Fine-tuning | <i>AWT</i> | Best of 20 samples | Transformer | 0.59 |
| | | Best of 20 samples | Transformer | 0.53 |

- **Discriminator** decreases the F1 score (improves secrecy)

2. Secrecy Evaluation – vs. Baseline

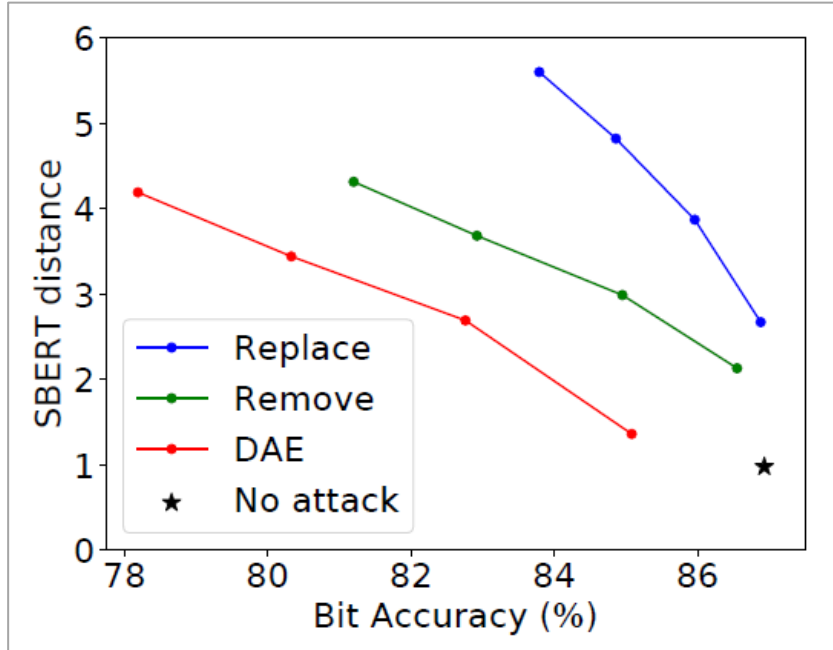
| Model | Acc. | SBERT | F1 |
|------------|--------------|--------------|-------------|
| Synonym | 83.9% | 3.62 | 0.98 |
| <i>AWT</i> | 86.8% | 0.956 | 0.53 |

- Baseline by [Topkara et al. 2006]
Watermarking texts with synonym substitution with WordNet

3. Robustness Evaluation

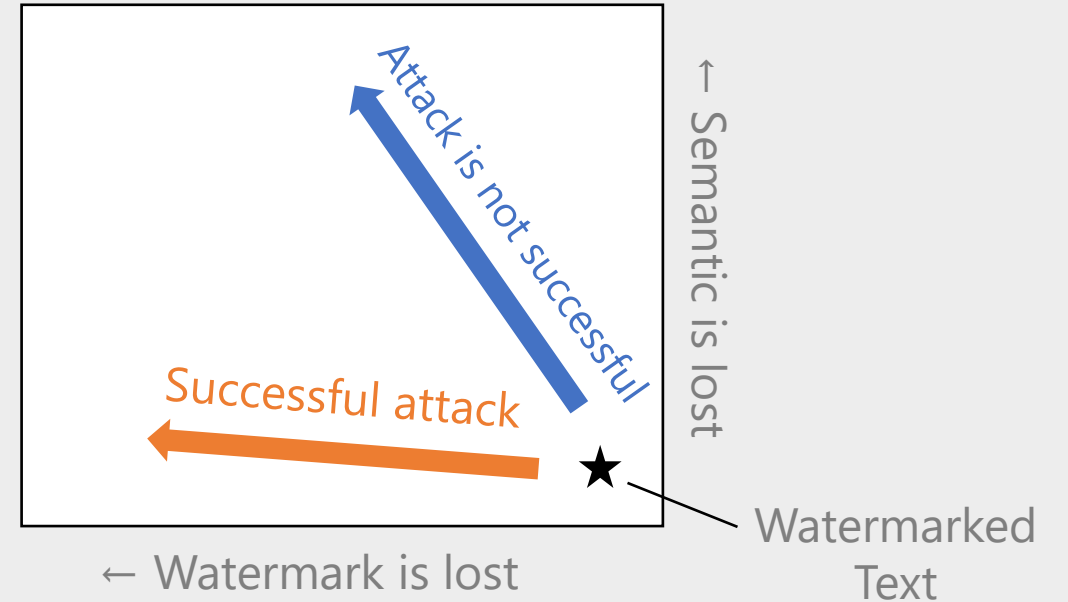
- Random changes
 - **Replace / Remove words randomly** in a **watermarked** sentence
- Training counter-models
 - Trained transformer-based **denoising autoencoder** (DAE)
 - Apply 2 types of noise to the input (**watermarked**) sentence
 - Embedding dropout
 - Random word replacement

3. Robustness Evaluation – Result



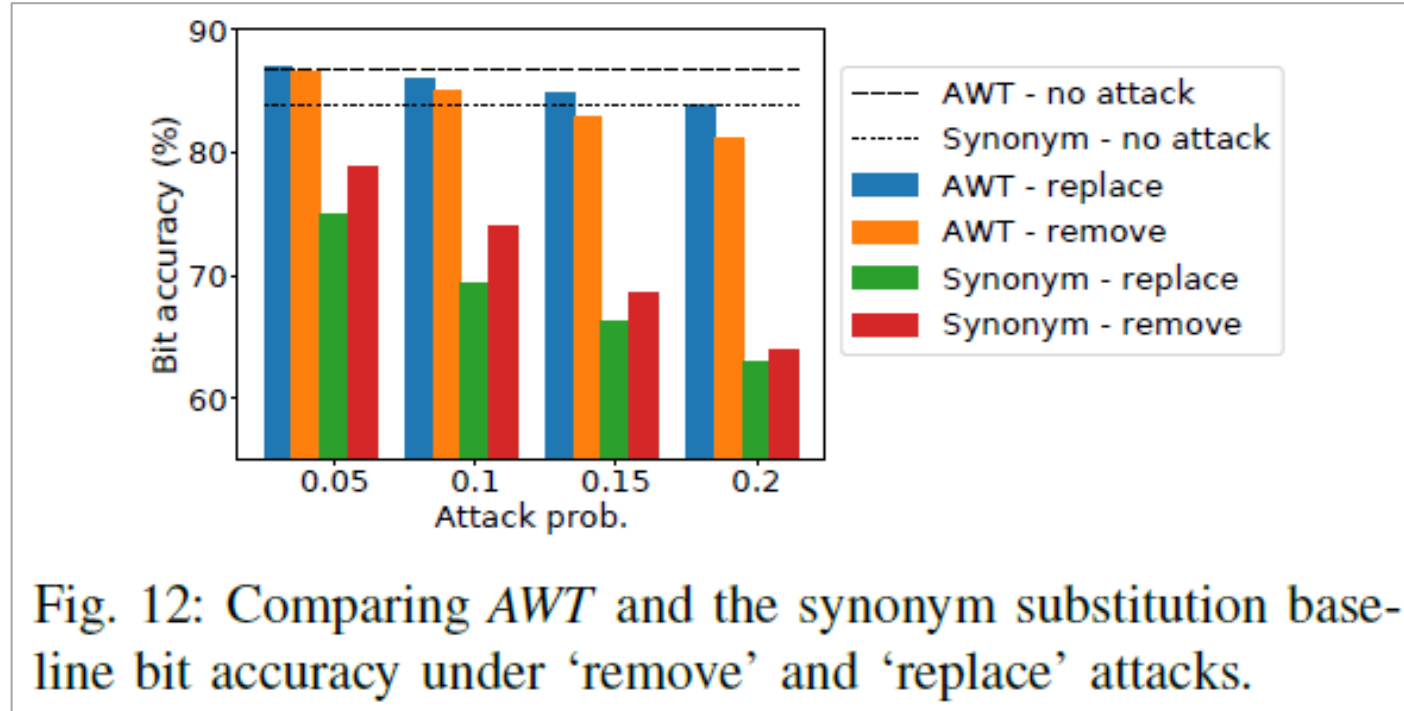
Bit accuracy is decreased a bit,
SBERT distance is increased significantly
→ Robust to the attacks

How to read the graph



The goal of attack is "remove the watermark with minimal changes to the text"

3. Robustness Evaluation – vs. Baseline



AWT keeps higher bit accuracy after remove / replace attacks compared to synonym substitution baseline.

4. Human Evaluation

Asked 6 judges to rate the sentence.

Sentence is randomly selected from non-watermarked text, AWT output, synonym baseline output.

| Rating | Description |
|--------|--|
| 5 | The text is understandable, natural, and grammatically and structurally correct. |
| 4 | The text is understandable, but it contains minor mistakes. |
| 3 | The text is generally understandable, but some parts are ambiguous. |
| 2 | The text is roughly understandable, but most parts are ambiguous. |
| 1 | The text is mainly not understandable, but you can get the main ideas. |
| 0 | The text is completely not understandable, unnatural, and you cannot get the main ideas. |

TABLE XVI: Ratings explanations given in the user study.

4. Human Evaluation – Result

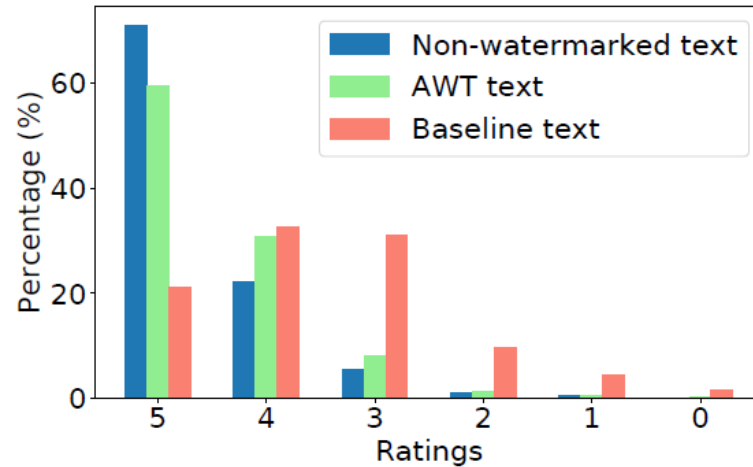


Fig. 18: Histograms of ratings given to the three types of sentences in the user study.

- AWT output texts are rated highly than baseline texts.

| <i>AWT</i> | Synonym-baseline | Non-wm Dataset |
|------------|------------------|----------------|
| 4.5±0.76 | 3.42±1.16 | 4.65±0.62 |

TABLE IX: The results of a user study to rate (0 to 5) sentences from *AWT*, the baseline, and non-watermarked text.

Contents

- About Watermarking
- Motivation
- Proposed Method
- Evaluation
- **Conclusion**

Conclusion

- New framework for language watermarking as a solution towards marking and tracing the provenance of machine-generated text
- First end-to-end data hiding solution for natural text.
- **Discriminator as an adversary** improved the watermark system.
- **Fine-tuning with additional language losses** improved the output text quality.