Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding

Sahar Abdelnabi, Mario Fritz

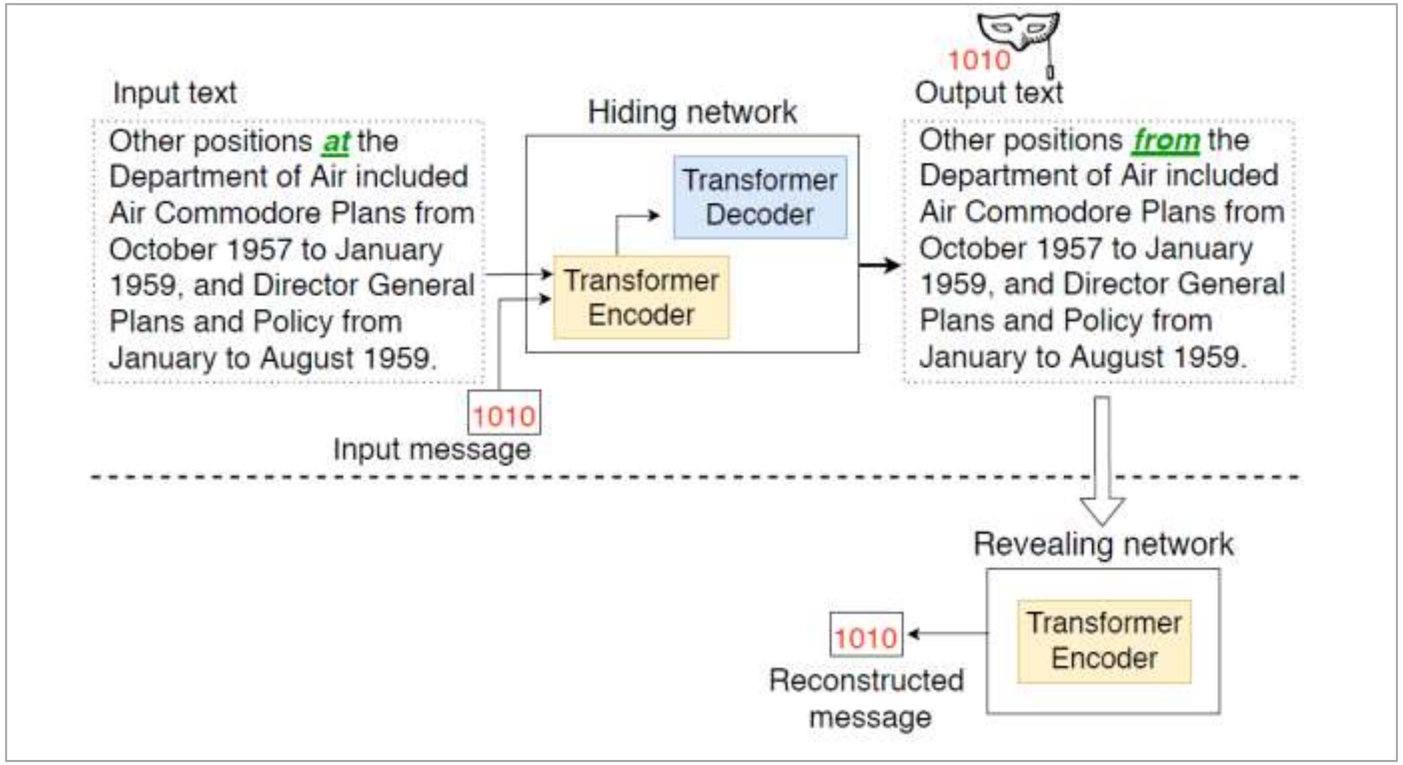
CISPA Helmholtz Center for Information Security

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

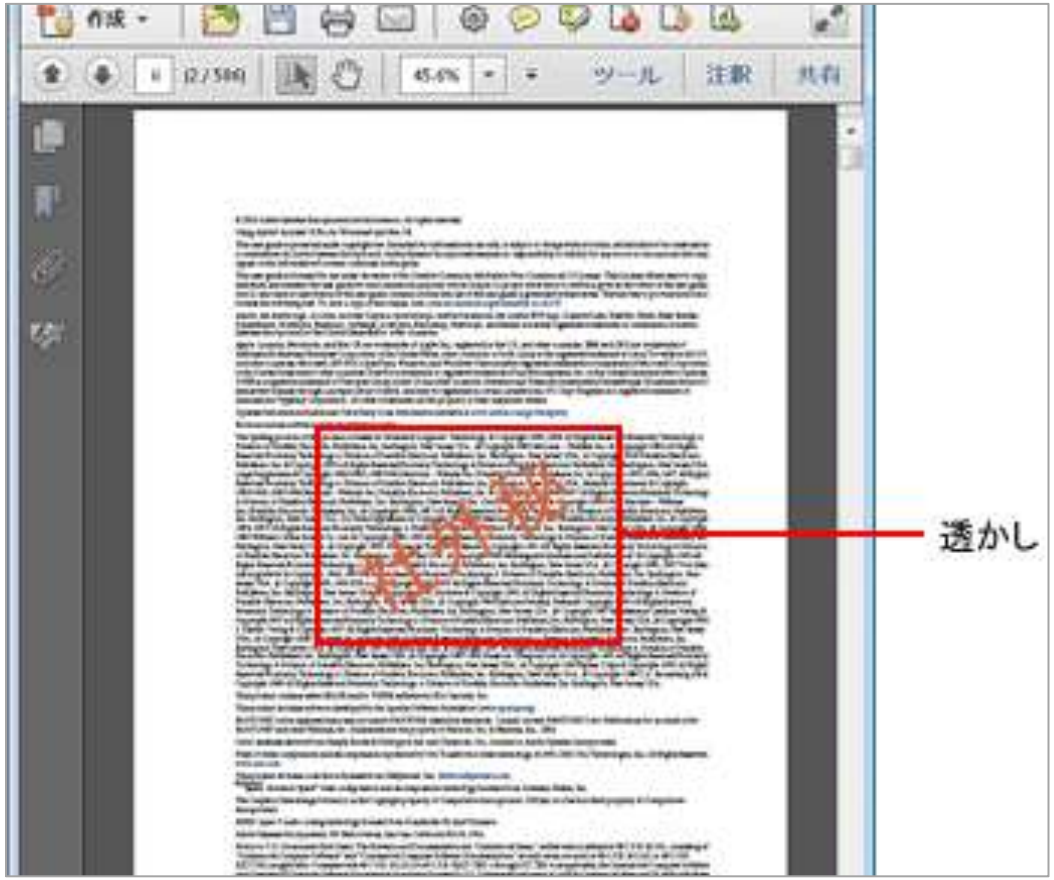
Contents

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

What is Watermarking (透かし)?

Visible (recognizable) watermarking (Physical & Digital)

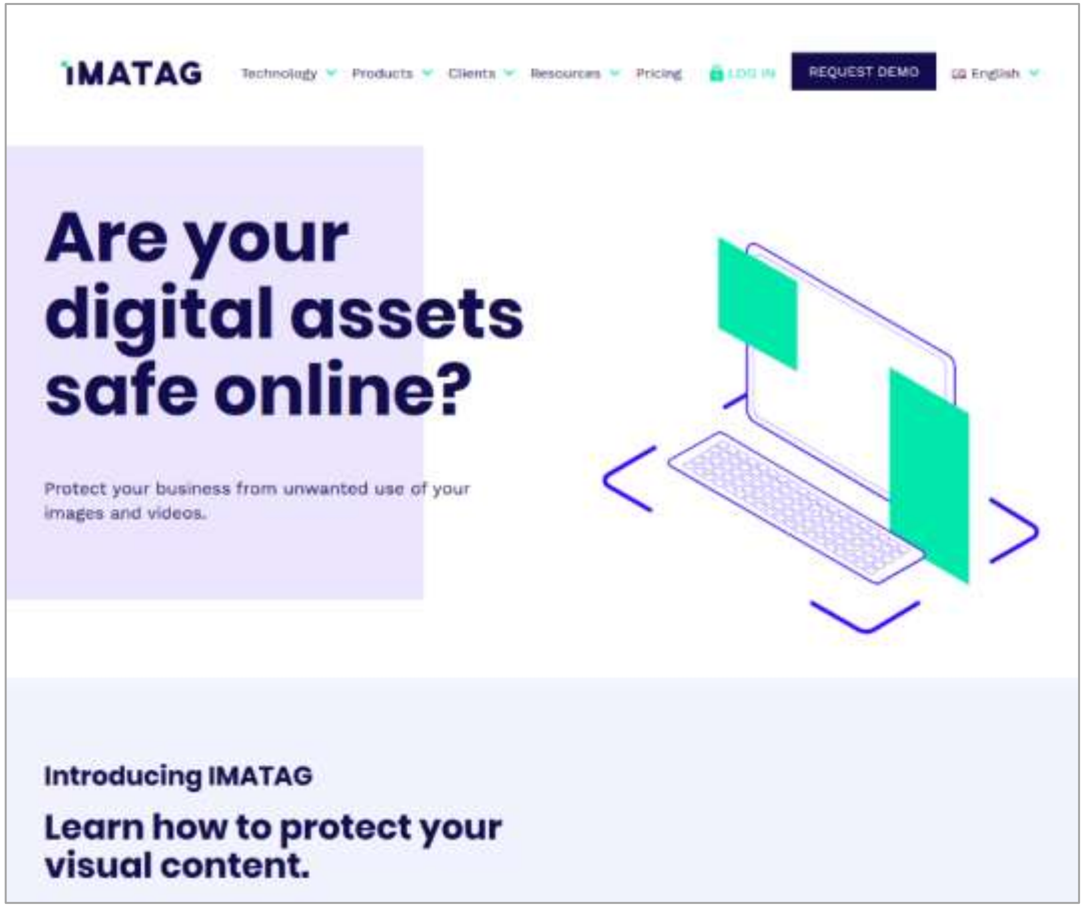
https://www.boj.or.jp/note_tfjgs/note/security/miwake.pdf



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

What is Watermarking (透かし)?

Invisible (unrecognizable) watermarking (Physical & Digital)



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

IMATAG https://www.imatag.com/

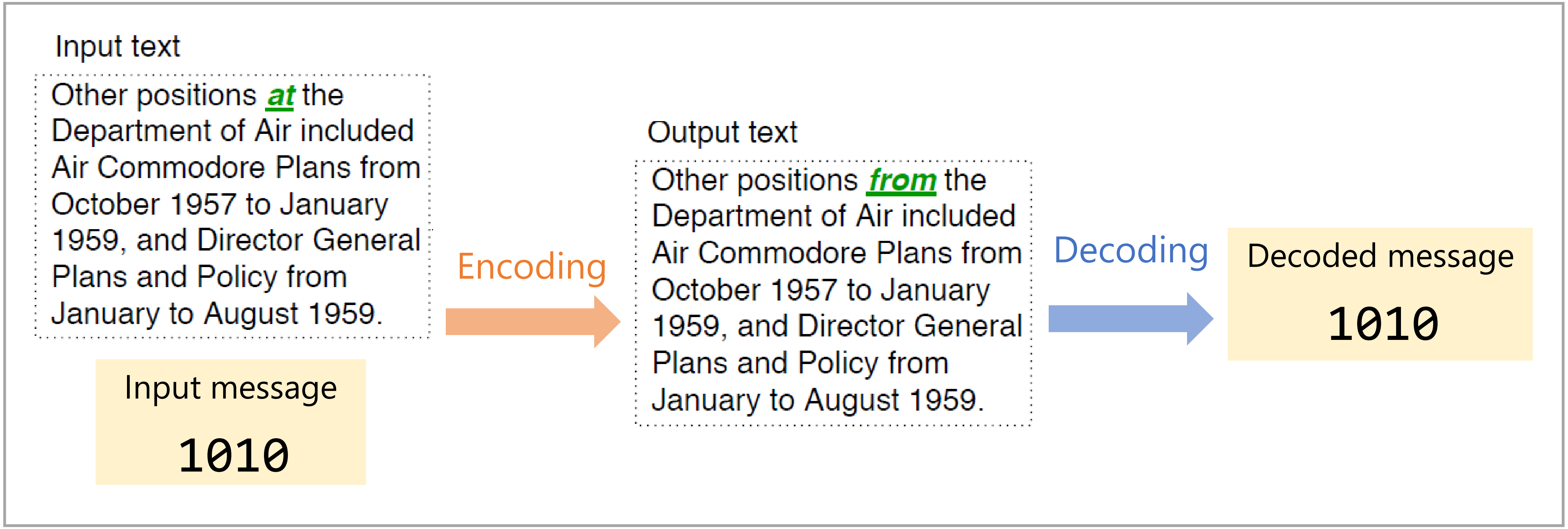
Difference from Cryptography (暗号), Steganography

<table>
<tr><th></th><th>Watermarking</th><th>Steganography</th><th>Cryptography</th></tr>
<tr><td>Goal</td><td>Hiding some data in a media, the data is related to the media</td><td>Hiding the existence of the data over other media (data is not always related to the media)</td><td>Hiding the content of the data</td></tr>
<tr><td>Required decoding accuracy</td><td colspan="2">Depends on the case
(trade-off with robustness or media quality)</td><td>100%</td></tr>
<tr><td>Robustness against modifying the media / data</td><td>Required (suppose attacks to remove the watermark)</td><td colspan="2">Usually not required</td></tr>
</table>

References: [Chang, Clark 2014], [Ziegler et al. 2019]

Language Watermarking

Edit text with some rule to embed information



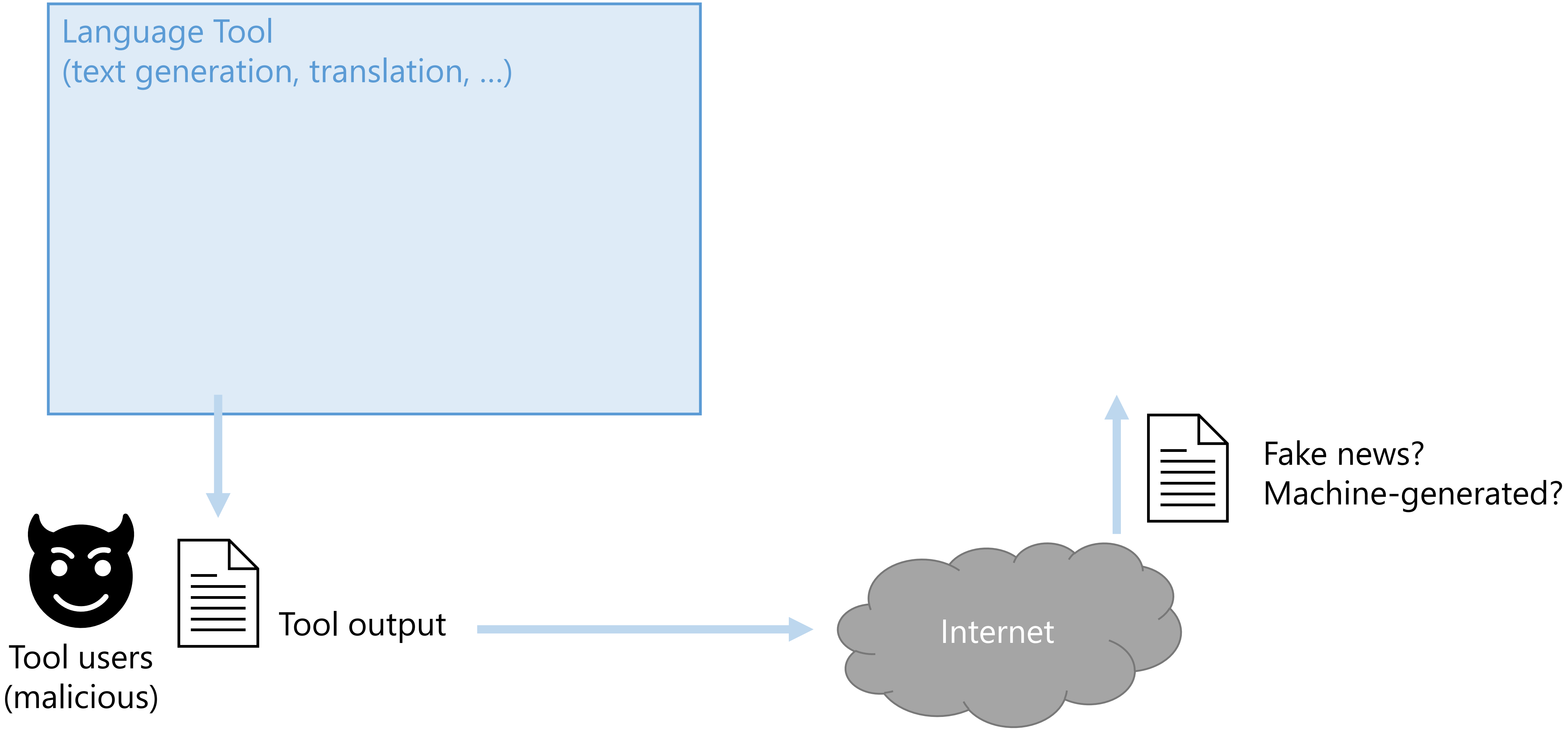
It also should be robust to

Contents

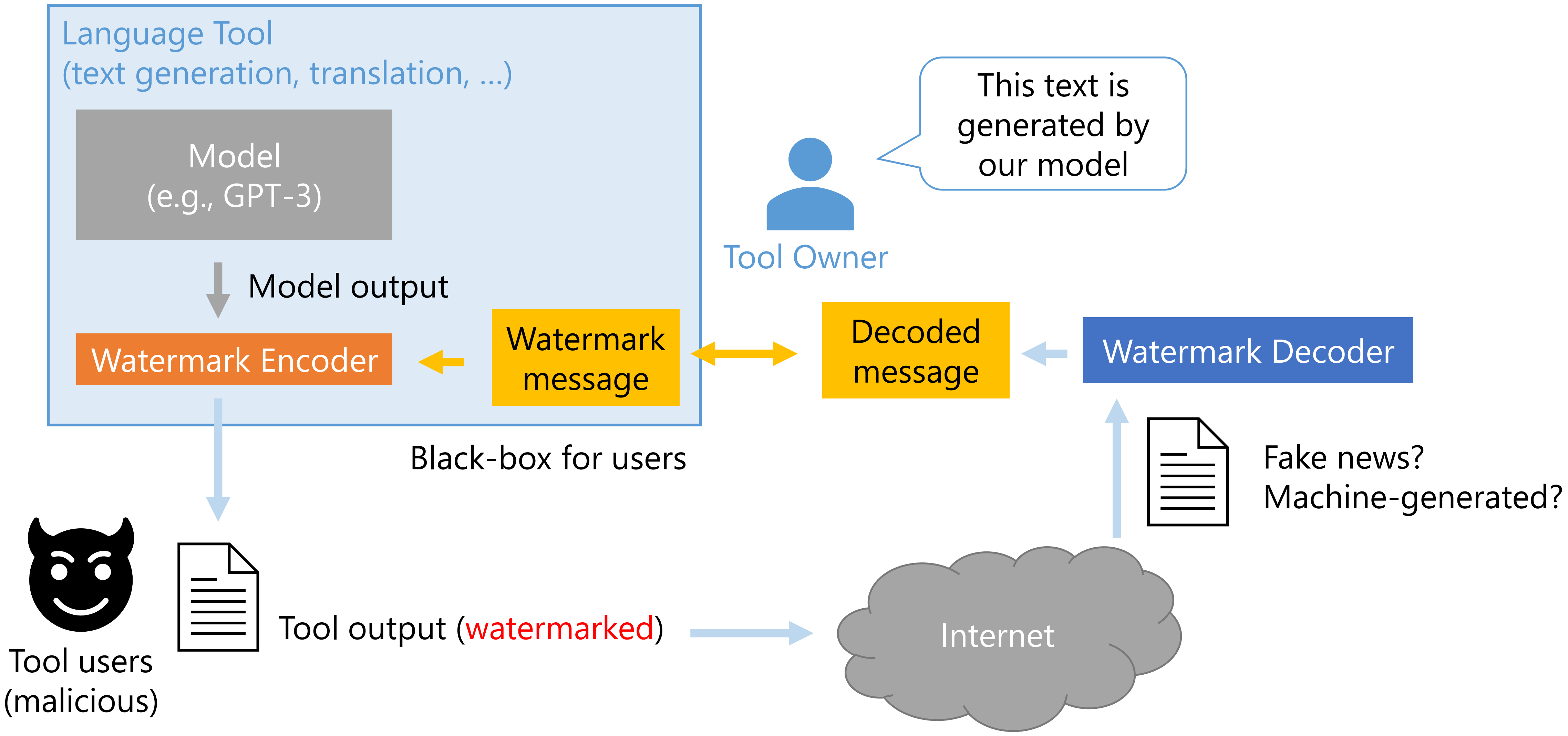
Motivation

- 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

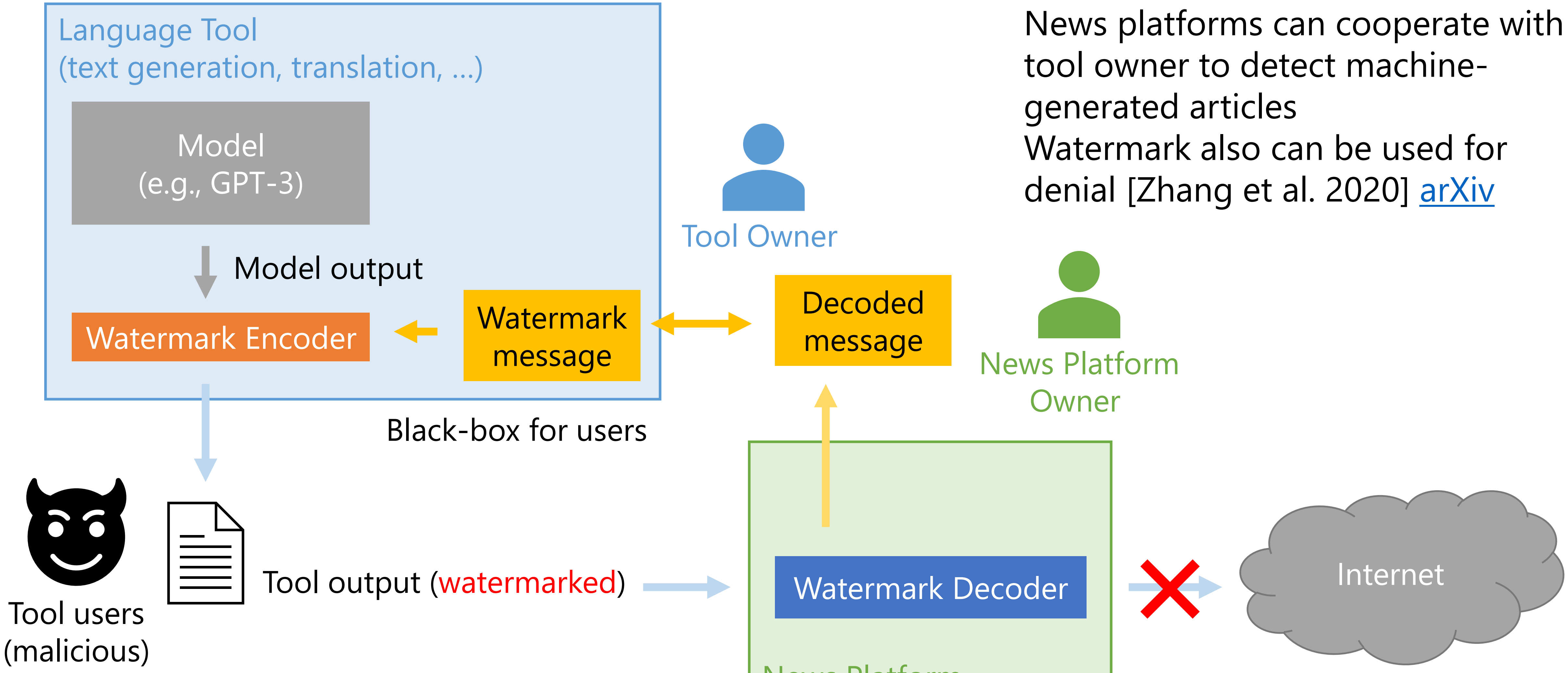
Usage Scenario



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

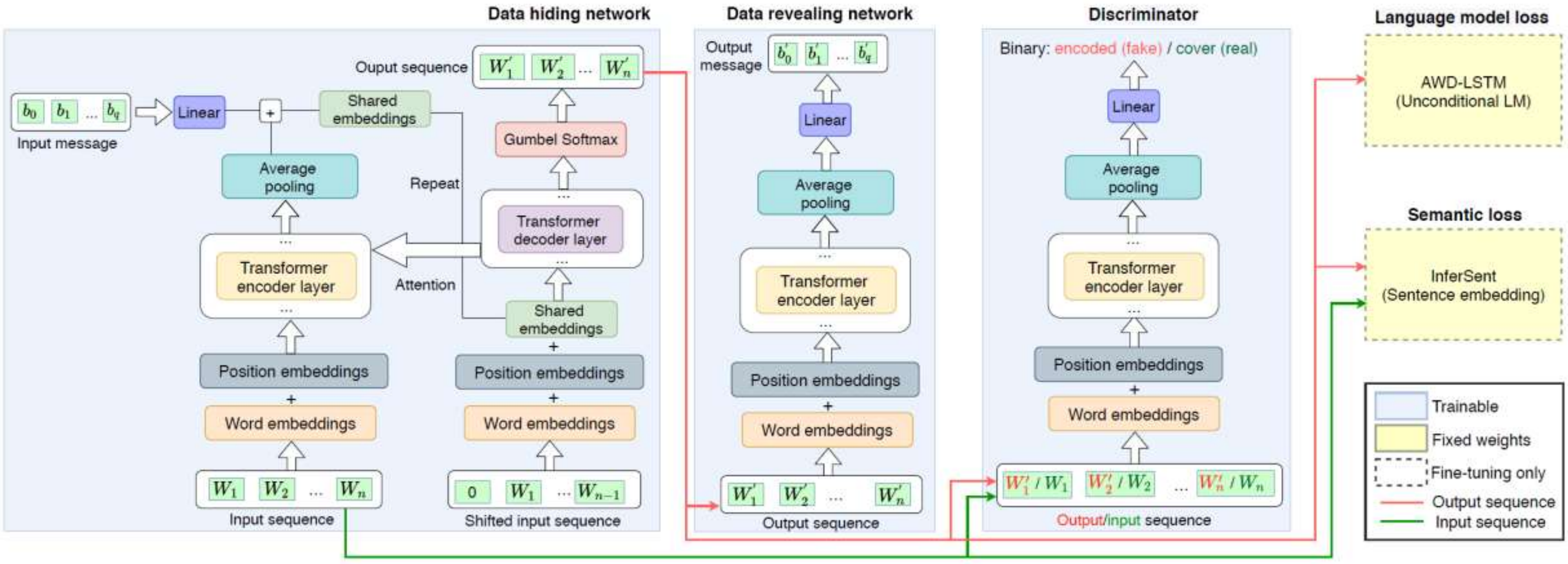
Existing Approaches

- Rule-based language watermarking
  - e.g., synonym substitution
  - They evaluates 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)

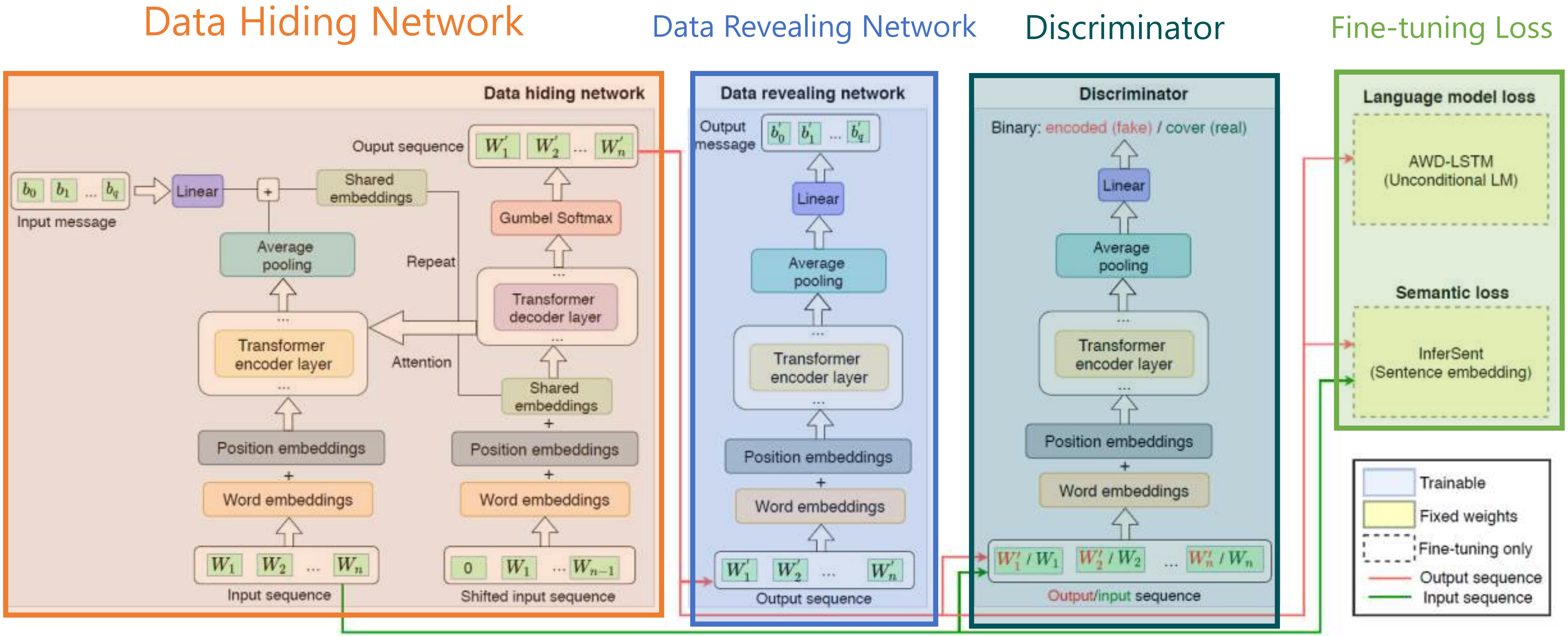
Contents

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

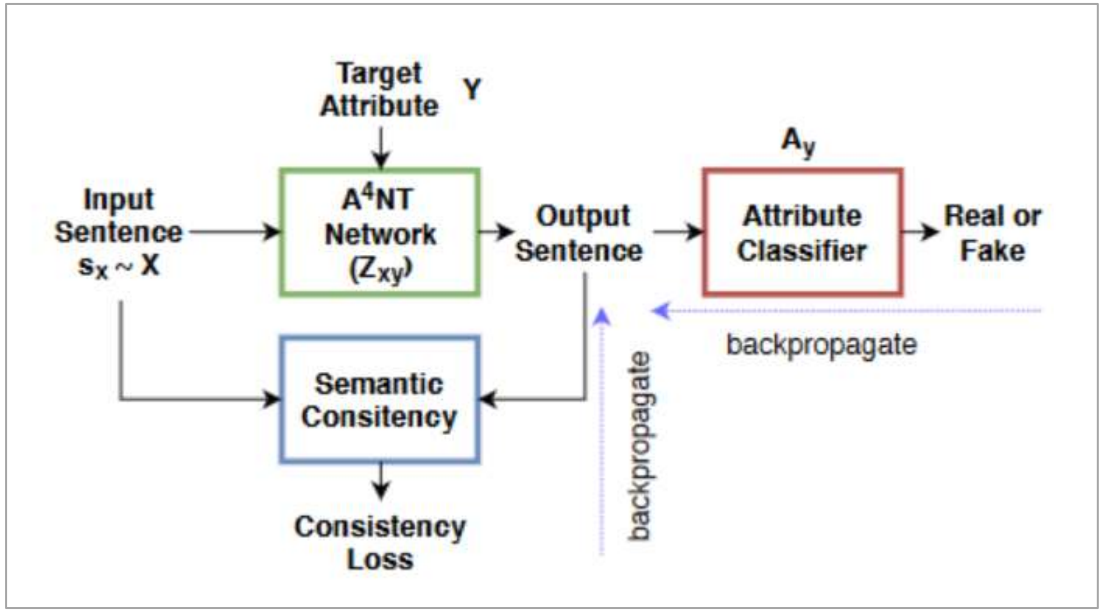
AWT: Adversarial Watermarking Transformer



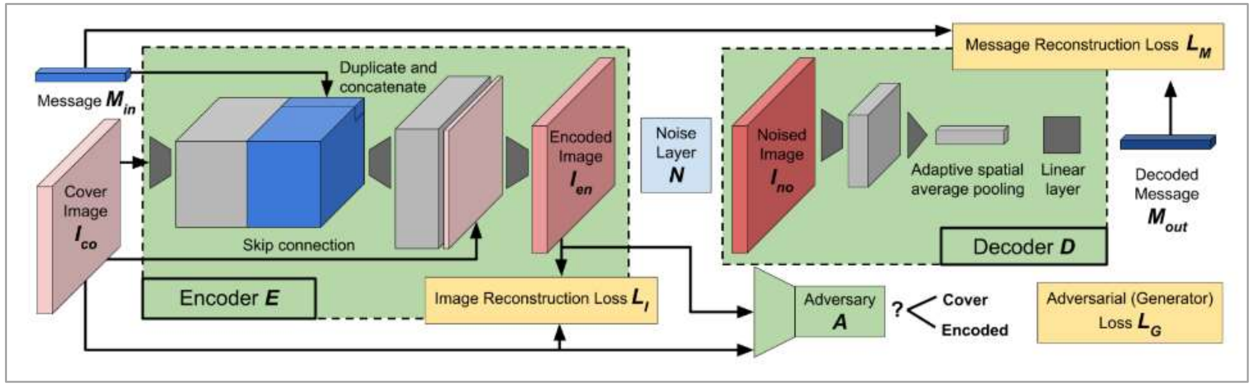
AWT: Adversarial Watermarking Transformer

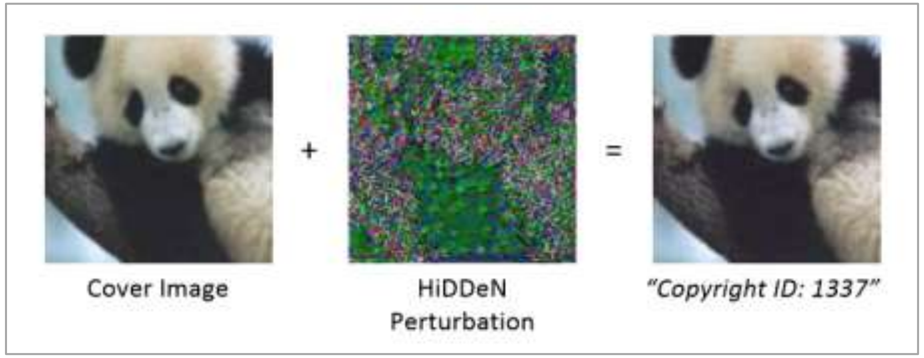


AWT – Similar Architecture [Shetty et al. 2018], [Zhu et al. 2018]



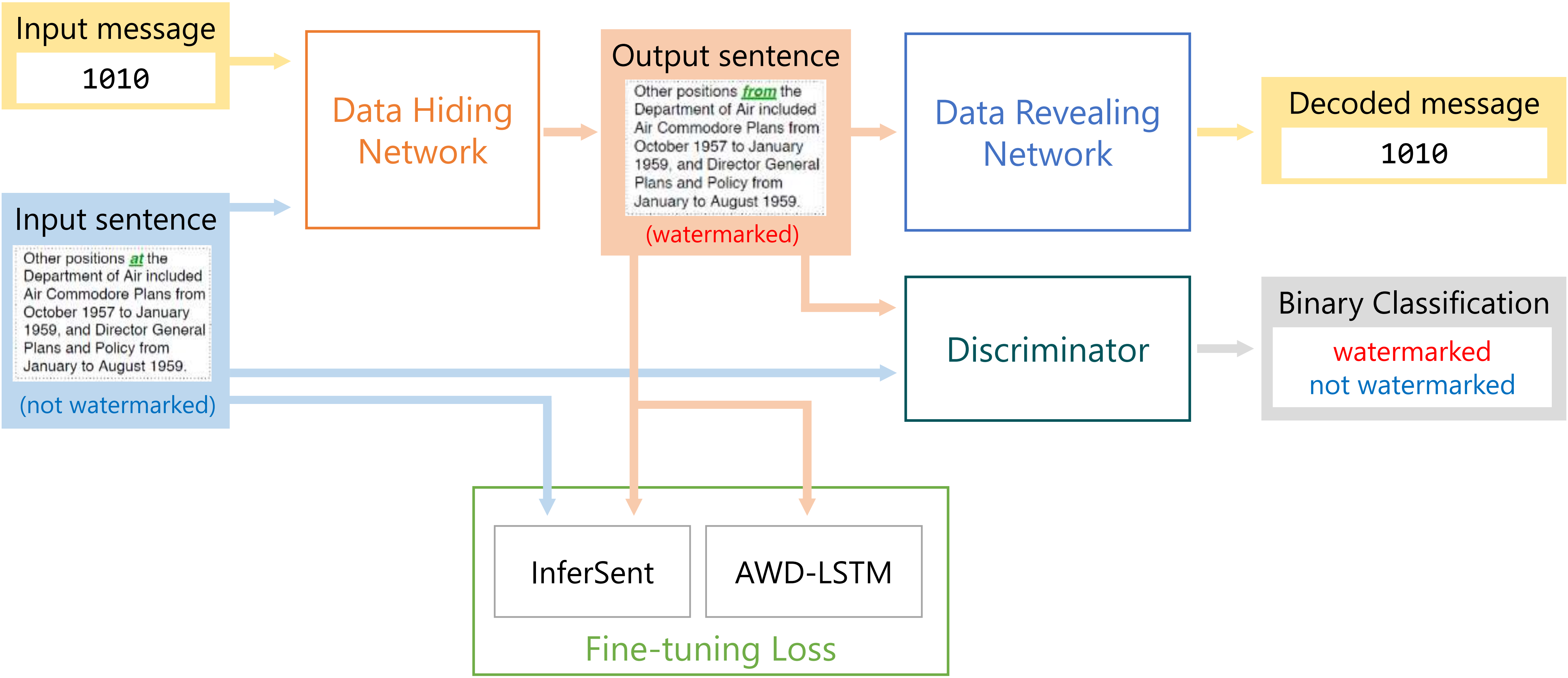
R. Shetty, B. Schiele, and M. Fritz, "*A4nt: author attribute anonymity by adversarial training of neural machine translation*," in 27th USENIX Security Symposium (USENIX Security 18), 2018. [PDF]





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

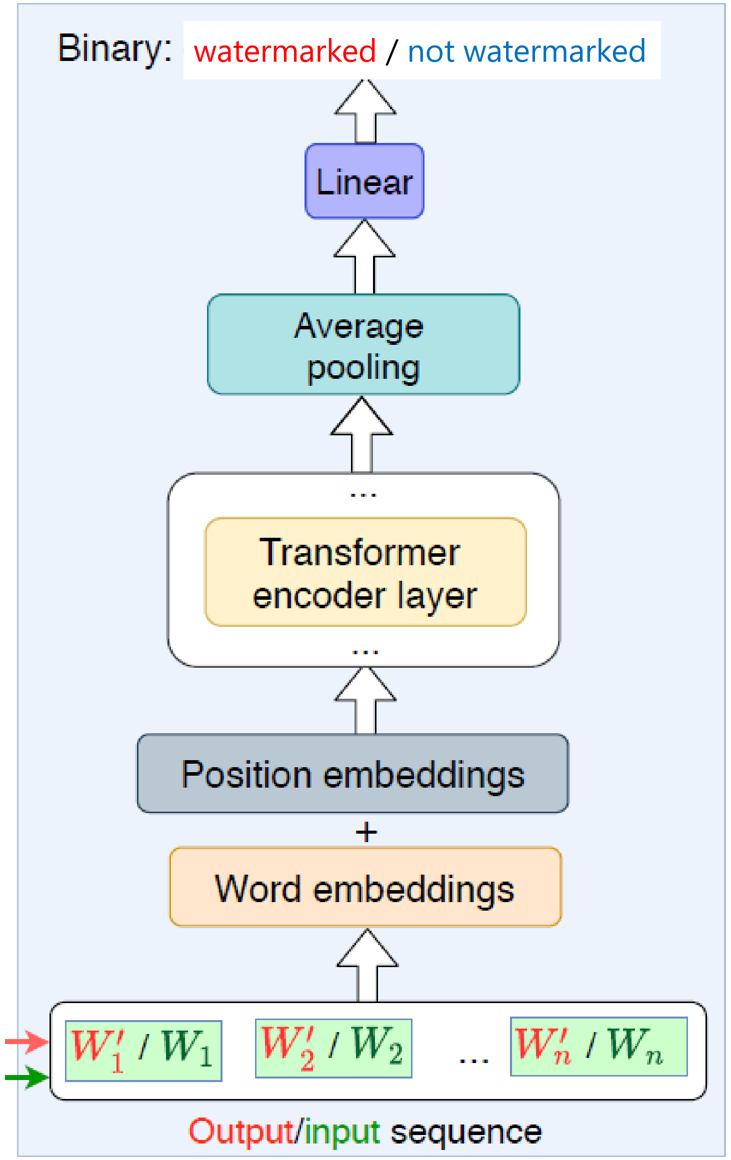
$$L_{disc} = -\log(A(S)) - \log(1 - A(S'))$$

A : discriminator
S : input (not watermarked) sentence
S' : output (watermarked) sentence

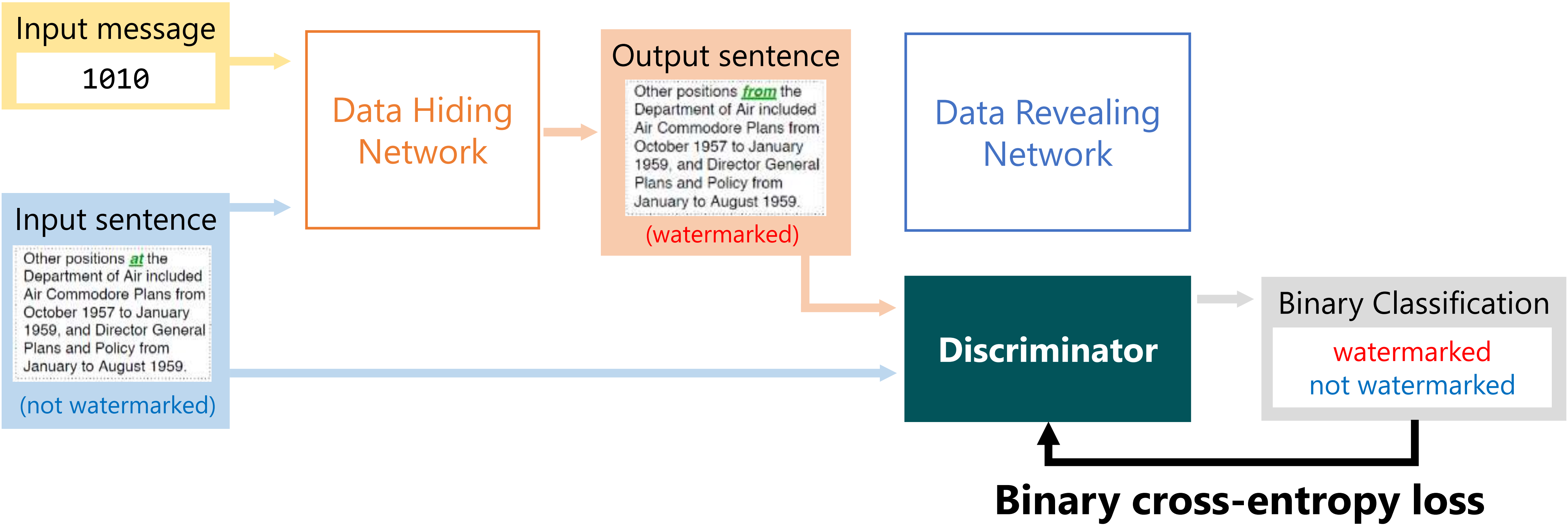
Adversarial loss L_A is

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

for training **data hiding network**



AWT – 1. Discriminator – Training



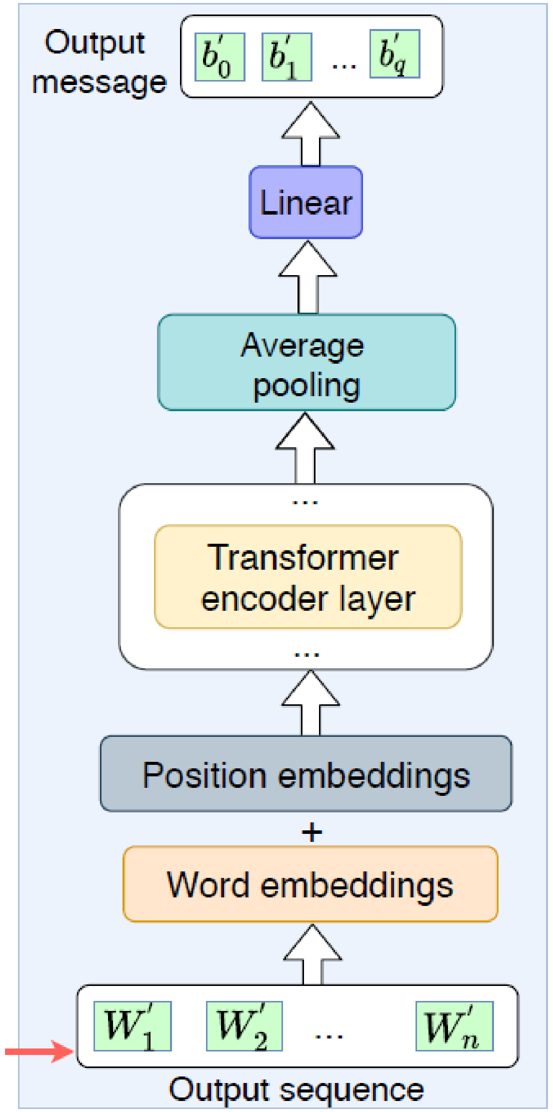
$$L_{disc} = -\log(A(S)) - \log(1 - A(S'))$$

Fine-tuning Loss is not used

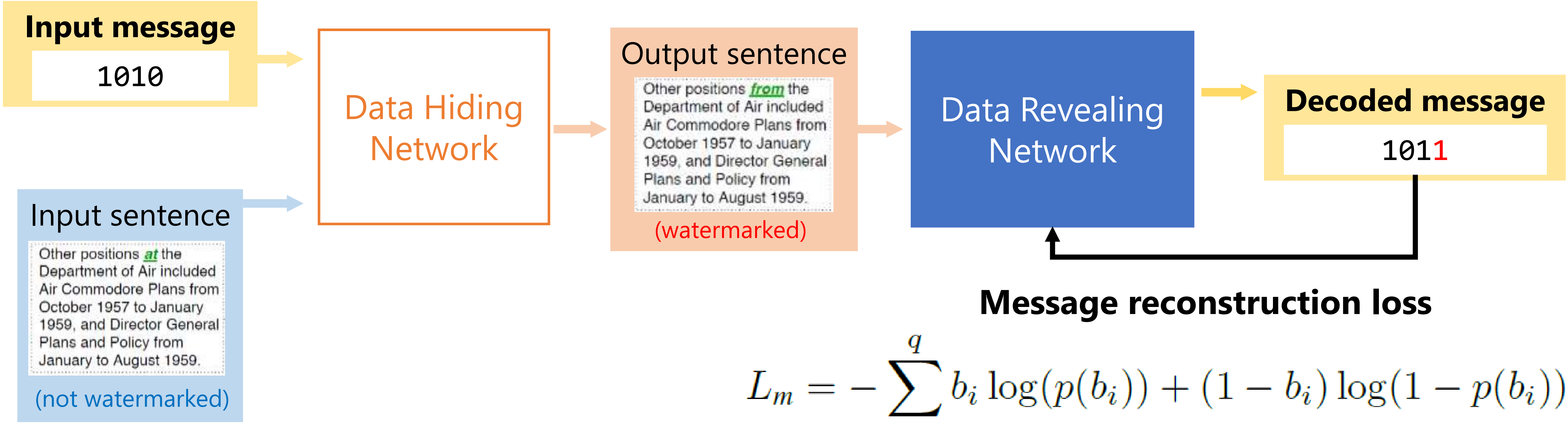
AWT – 2. Data Revealing Network

- 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



Message reconstruction loss

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

Fine-tuning Loss | Discriminator | are not used

AWT – 3. Data Hiding Network

A) Add input message to encoded embeddings

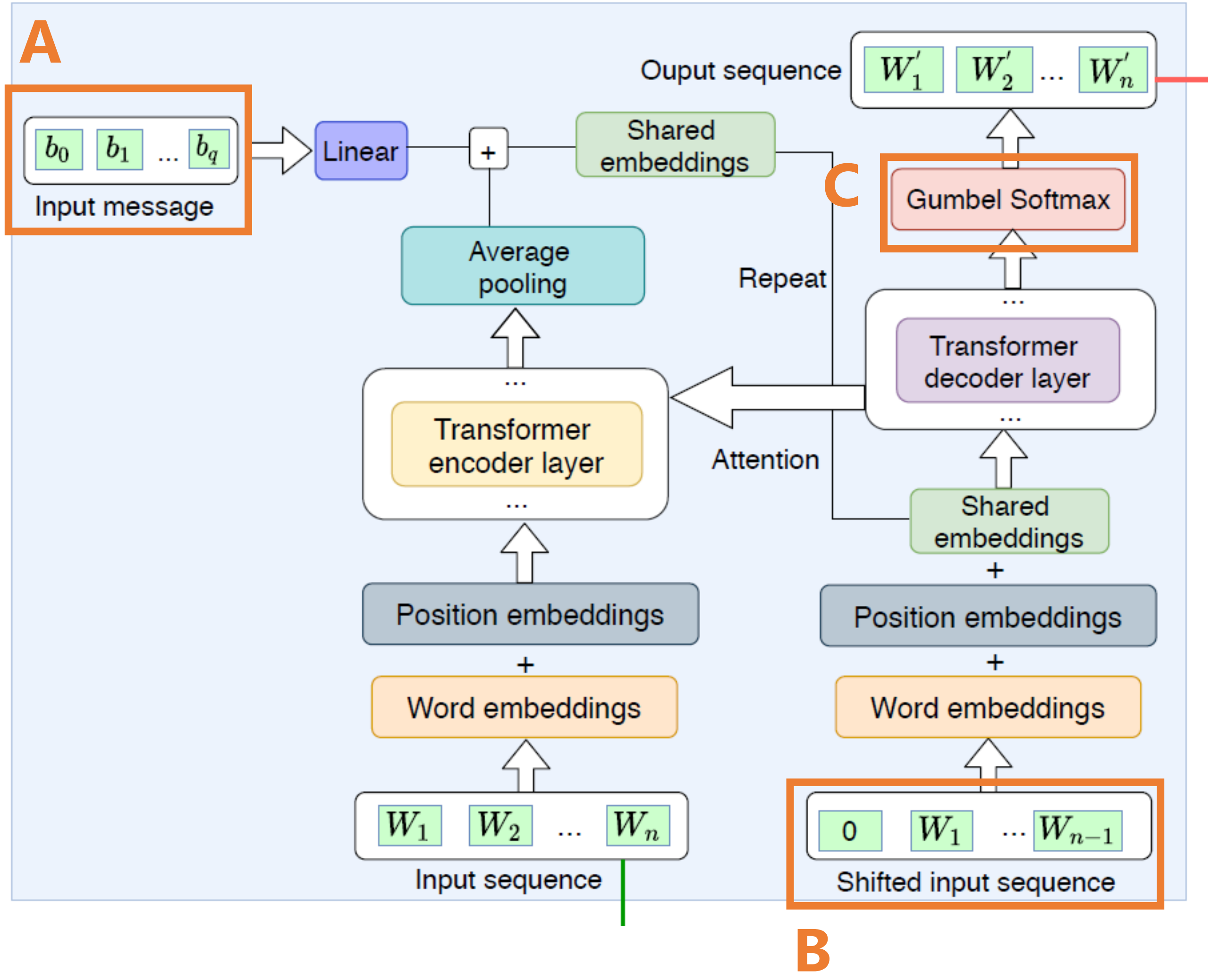
B) Transformer auto-encoder (the decoder takes **shifted input sentence**)

C) **Gumbel-softmax** to train jointly with other components

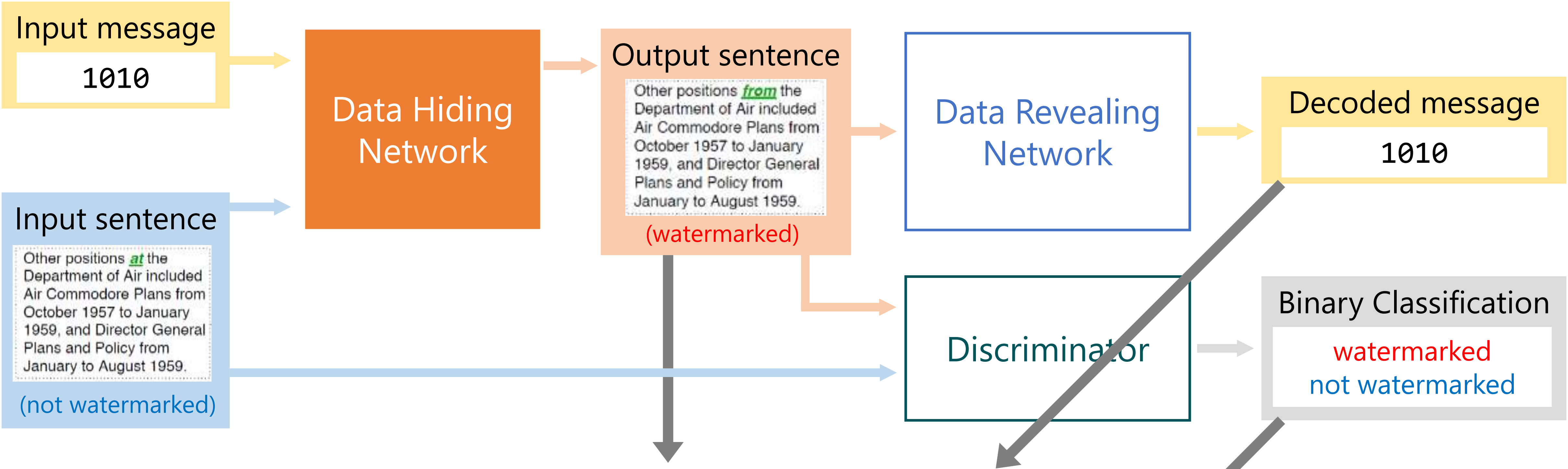
Text reconstruction loss L_{rec} :

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

cross entropy loss of input & output sequence



AWT – 3. Data Hiding Network – Training



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

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.

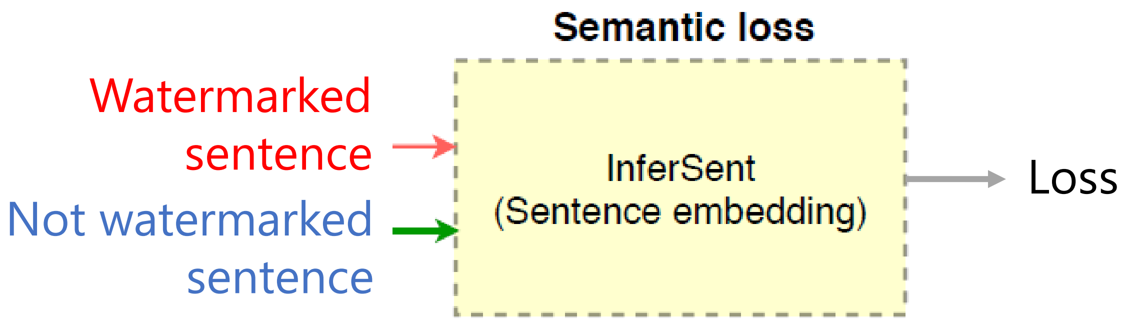
AWT – 4. Fine-tuning Loss

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

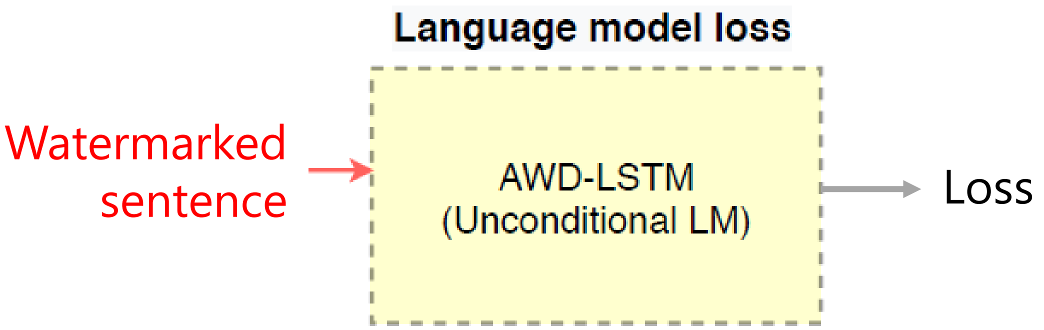


B) Preserving **Sentence Correctness**

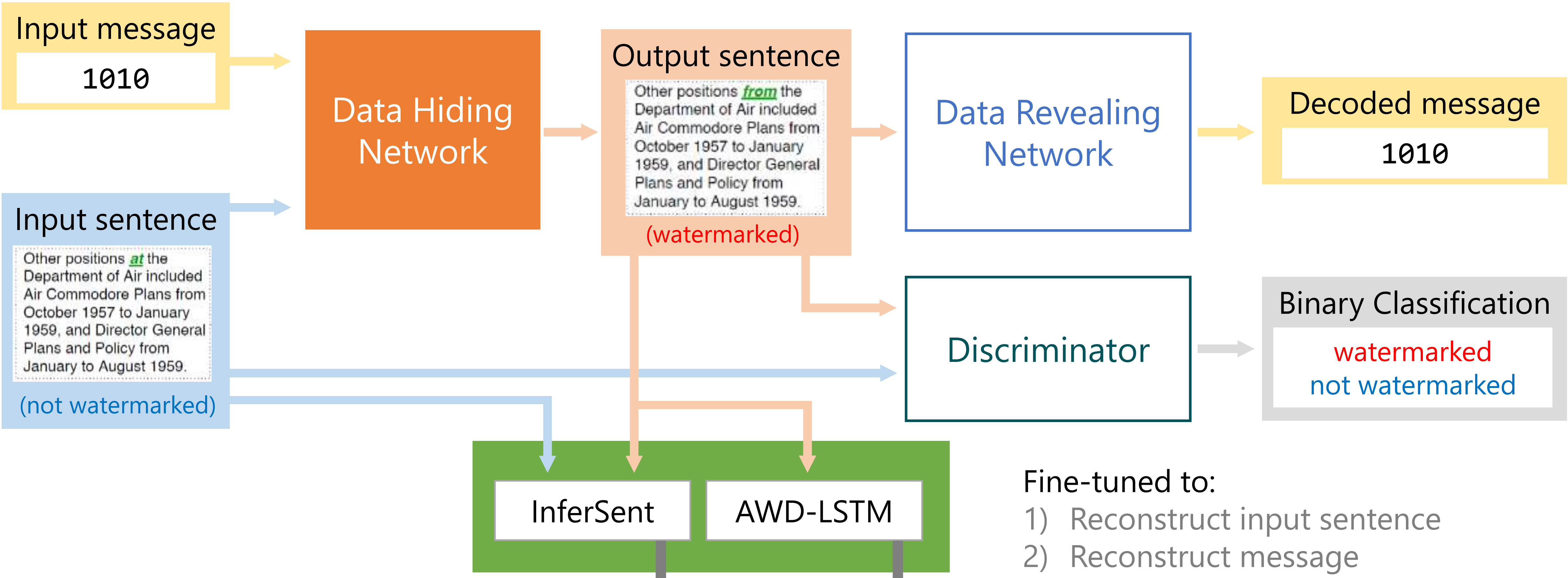
ASGD Weight-Dropped LSTM, independently trained on the dataset used as input texts (not watermarked texts)

$$L_{LM} = -\sum_i \log p_{LM}(W_i^{'}|W_{<i}^{'})$$

W_i' : the i th word in watermarked sentence



AWT – Fine-tuning



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

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
   - Denoising 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

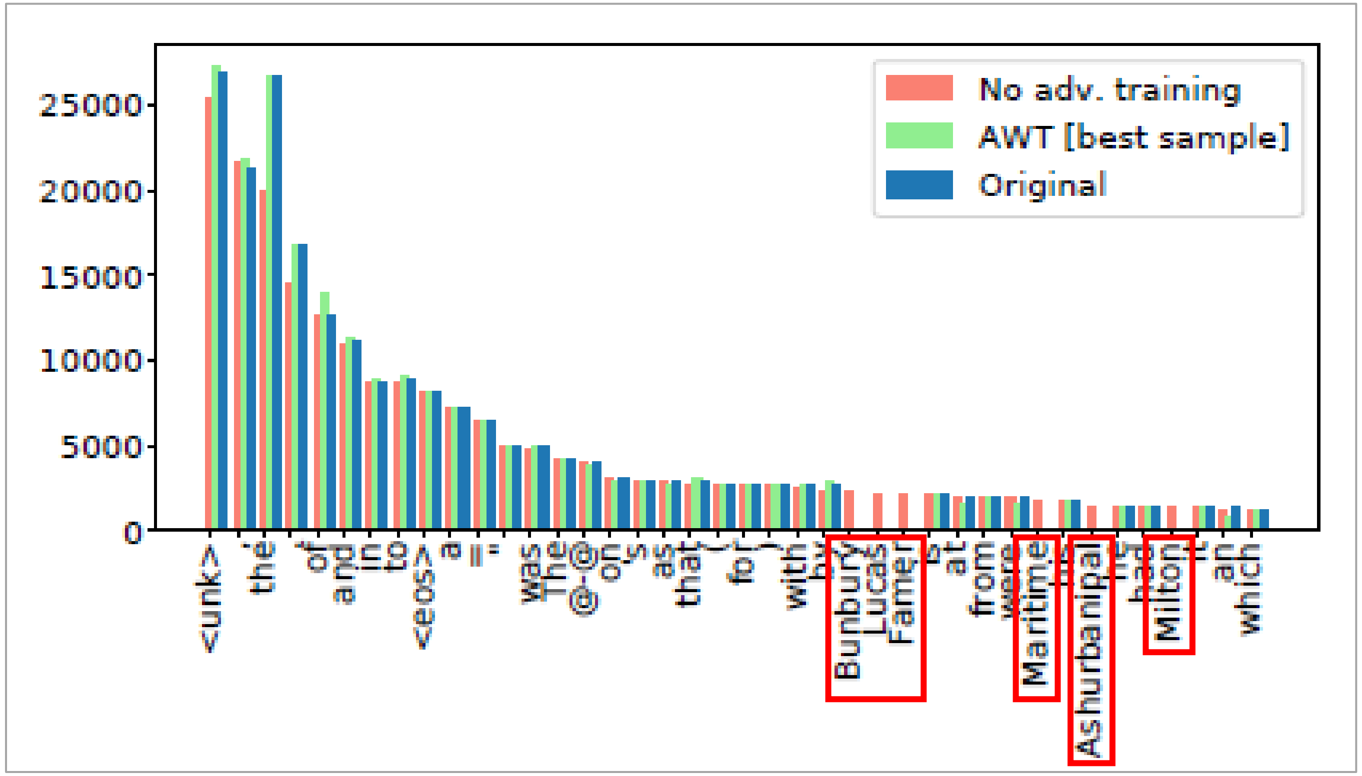
Model	Acc.	SBERT	F1
Synonym	83.9%	3.62	0.98
AWT	**86.8%**	**0.956**	**0.53**

- 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 *the* commanding officer.	He was appointed *Bunbury* commanding officer.
one of *the* most fascinating characters in *the* series	one of *Milton* most fascinating characters in *Milton* 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, *which* was formed by a group of leaders *from* 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 *an* 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> *had* 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

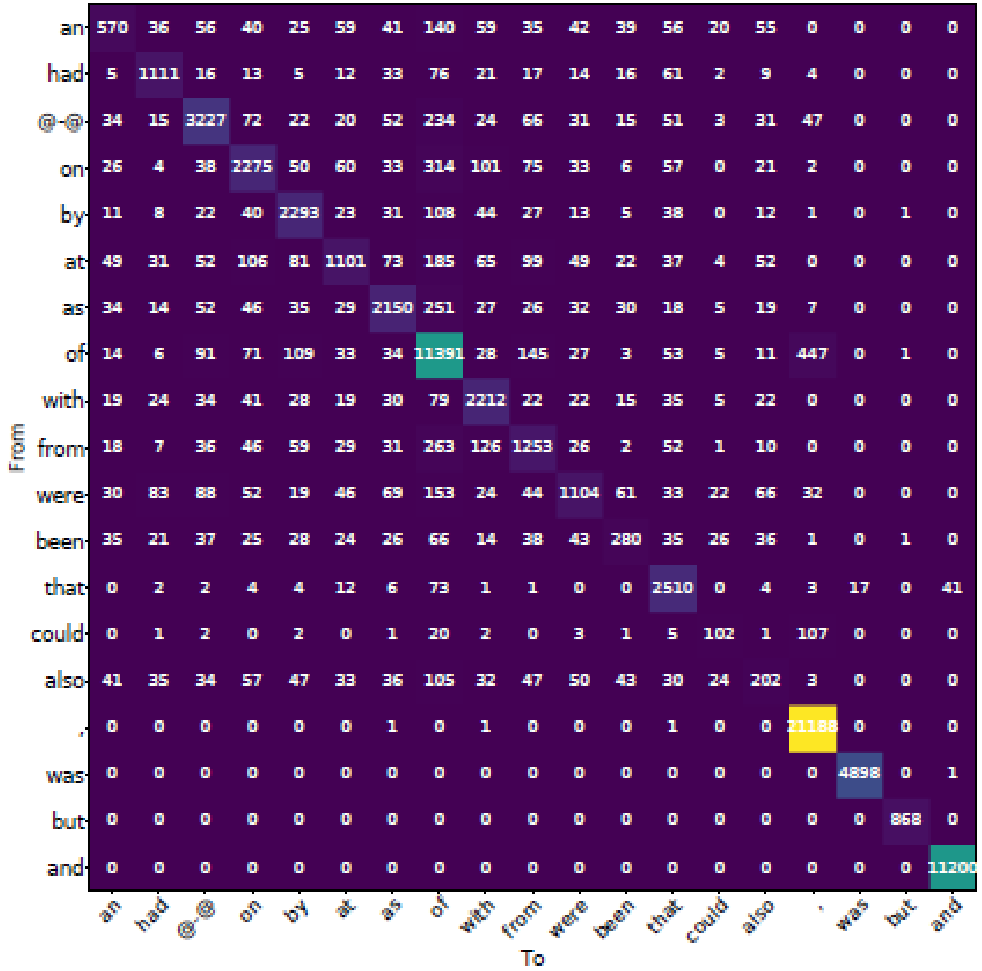


Fig. 10: A matrix of word changes count from original text to modified text using *AWT*.

Words are not fixedly replaced

2. Secrecy Evaluation

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

- Discriminator decreases the F1 score (improves secrecy)

2. Secrecy Evaluation – vs. Baseline

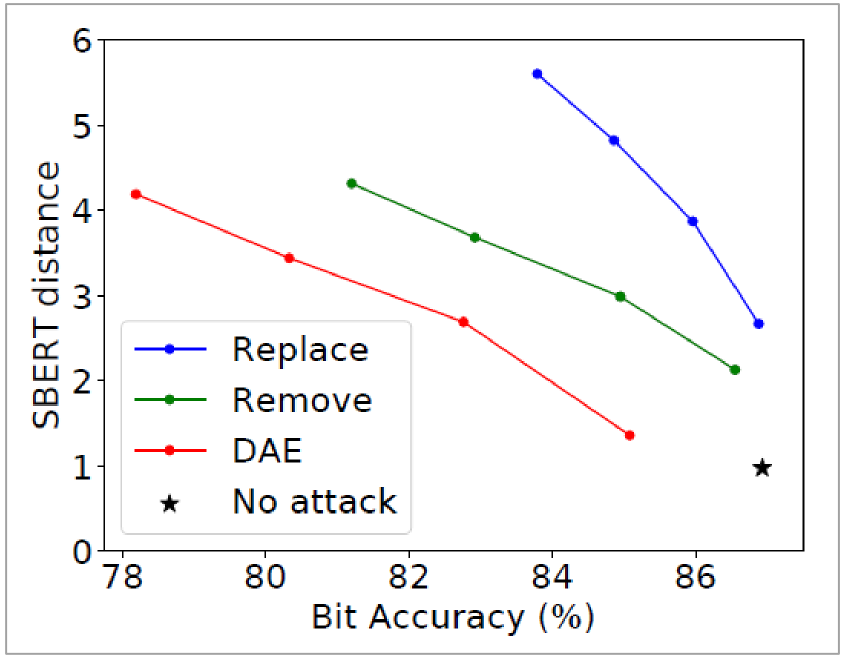
Model	Acc.	SBERT	F1
Synonym	83.9%	3.62	0.98
AWT	**86.8%**	**0.956**	**0.53**

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

3. Robustness Evaluation

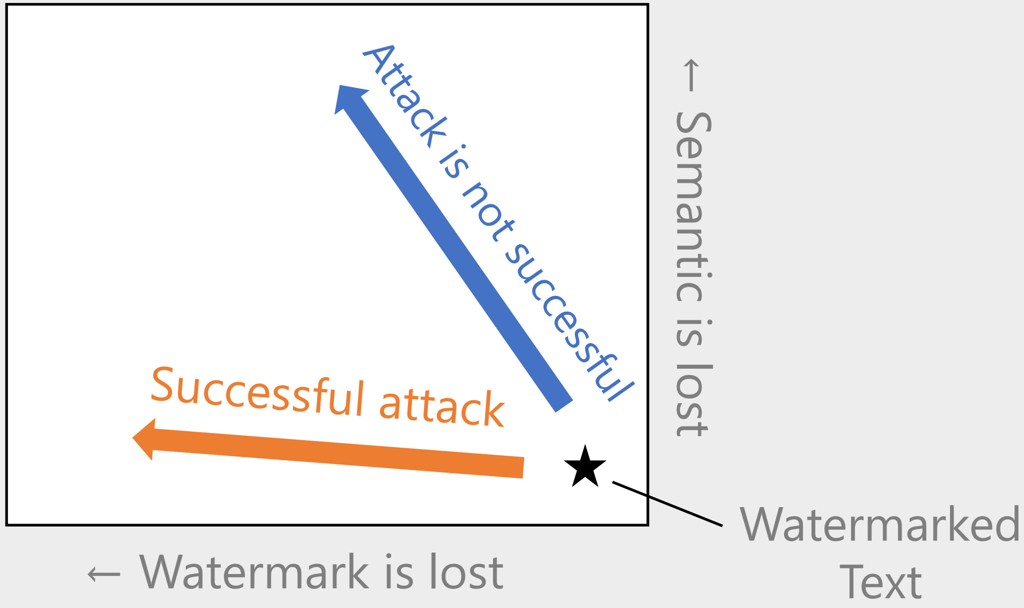
- Random changes
  - **Replace / Remove words randomly** in a watermarked sentence
- Training counter-models
  - Trained transformer-based **denoising autoencoder** (DAE)
  - Apply 2 types of noise to the input (watermarked) sentence
    - Embedding dropout
    - Random word replacement

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

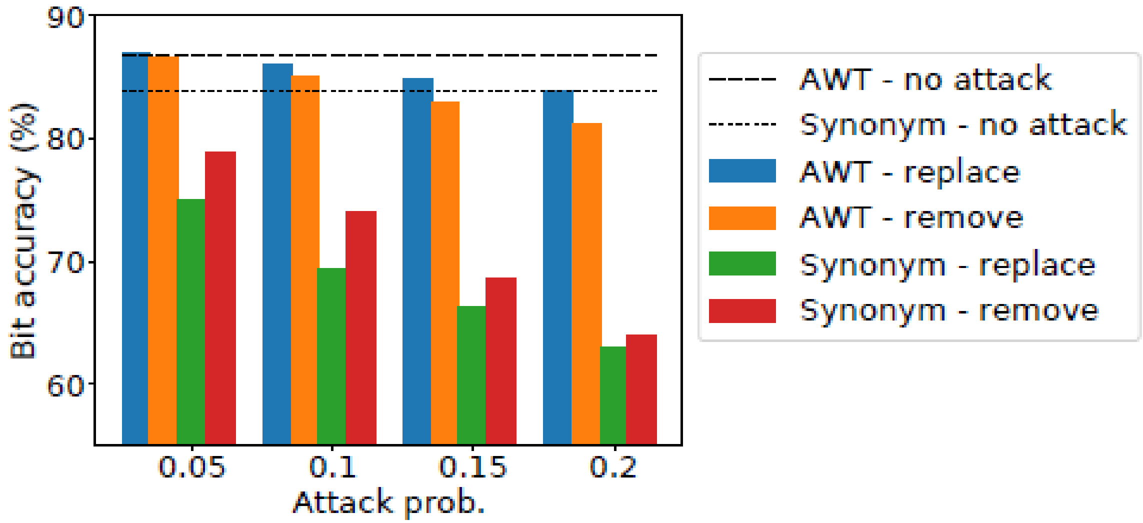


Fig. 12: Comparing *AWT* and the synonym substitution baseline bit accuracy under 'remove' and 'replace' attacks.

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

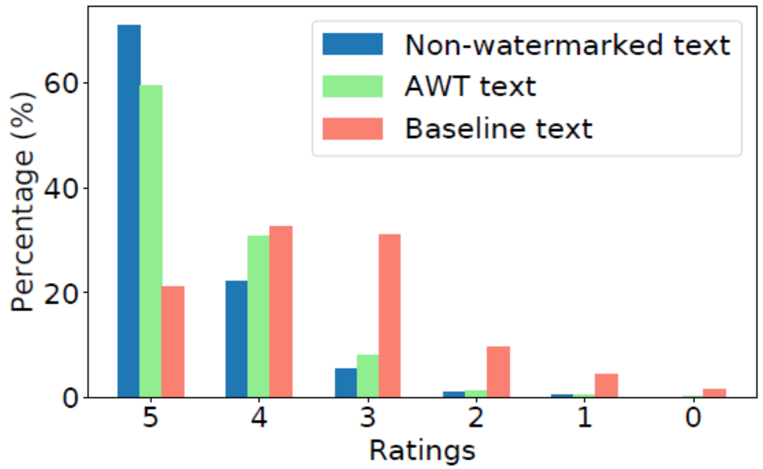


Fig. 18: Histograms of ratings given to the three types of sentences in the user study.

AWT	**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.

- AWT output texts are rated highly than baseline texts.

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.