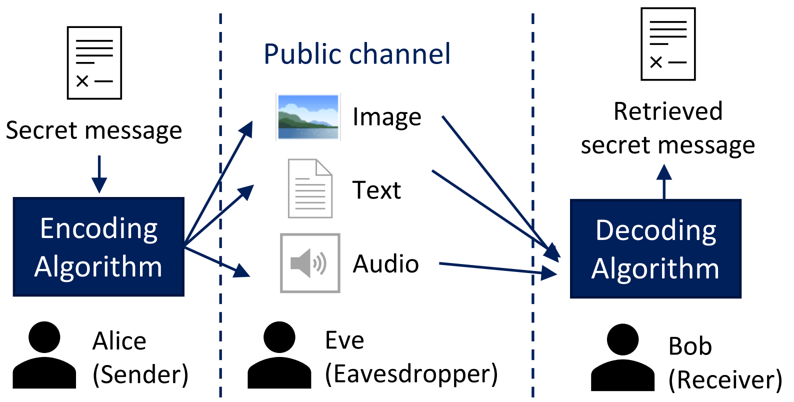
Frustratingly Easy Edit-based Linguistic Steganography with a Masked Language Model

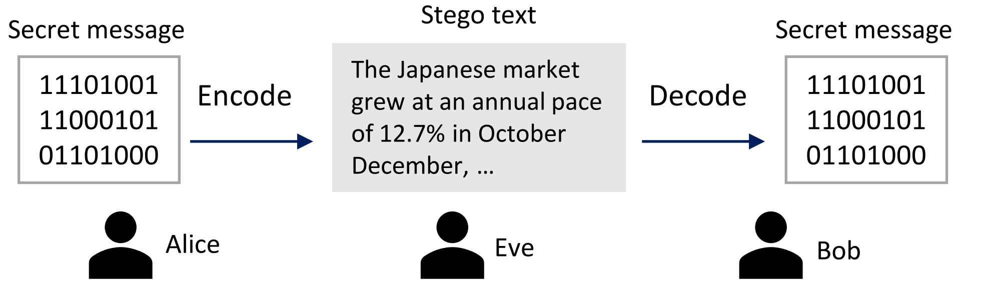
Honai Ueoka, Yugo Murawaki, Sadao Kurohashi.   Graduate School of Informatics, Kyoto University

Steganography



Concealing a message in some cover data such that an eavesdropper is not even aware of the existence of the secret message.

Linguistic Steganography



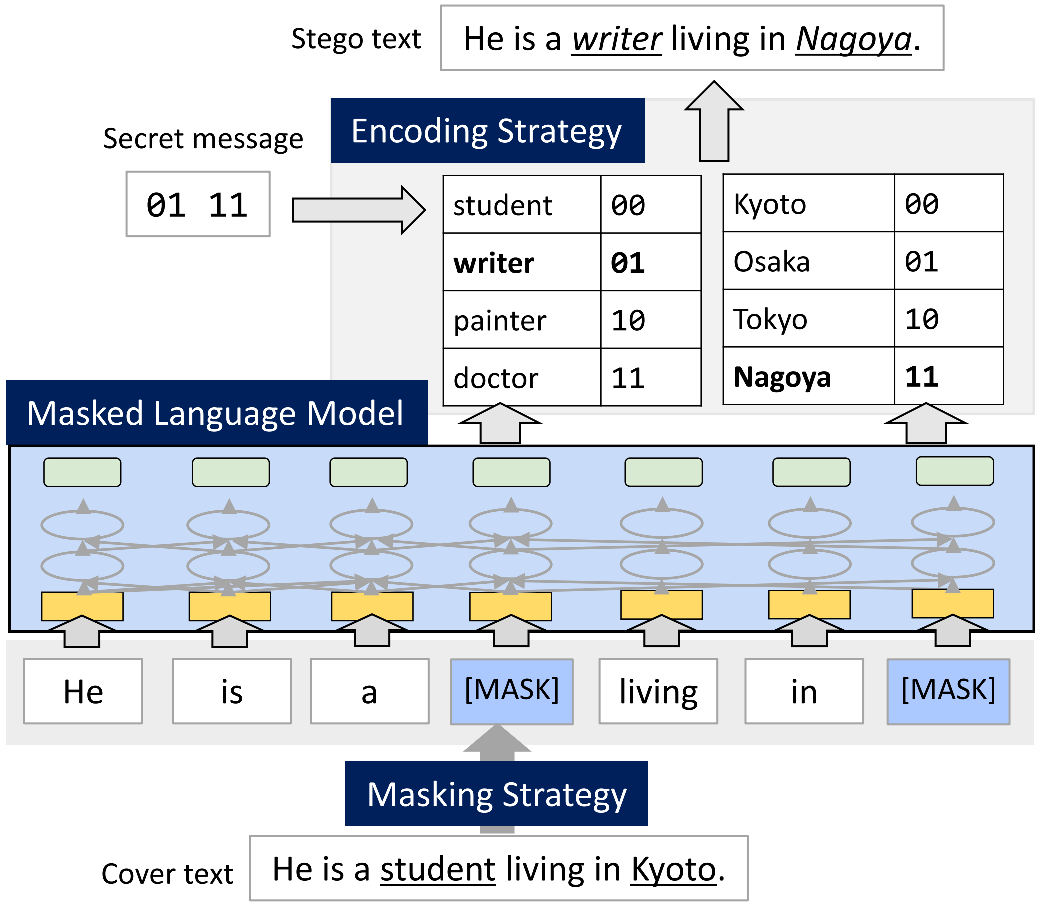
Stego text: steganographic text with an encoded secret message

Two objectives:

Security How unsuspicious the stego text is. A stego text should be so natural that it does not arouse the suspicion of the eavesdropper.

Payload capacity The size of the secret message relative to the size of the stego text.

Proposed Method

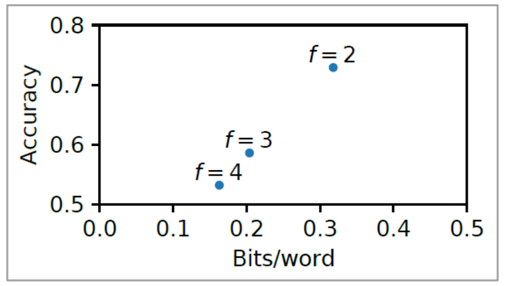


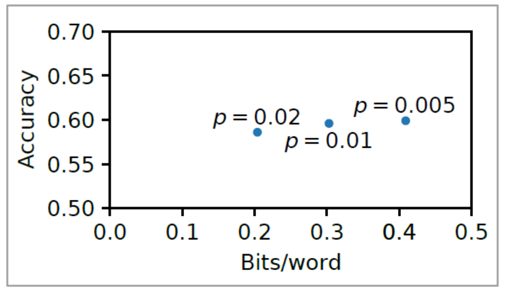
- Alice and Bob share the Masked LM, a masking strategy, an encoding strategy in advance.
- Alice masks some of the tokens of cover text. The LM predicts the substitution candidates. Finally, Alice creates a stego text by choosing corresponding tokens.
- The result of masking the original cover text has to be the same as the result of masking the stego text. This allows Bob to get the same distribution as Alice, and he can recover the secret message based on the chosen tokens in the stego text.

Experiments and Results

	Capacity [bits/word]	Discrimination Accuracy (Security)
Proposed Method with BERT	0.204	**0.586**
Generation-based method with GPT-2	**1.67**	0.819

- The proposed method achieved **high security**
- Generation-based stego texts were easily detectable
- Lower payload capacity than that of the generation-based method, but it's high for an edit-based method



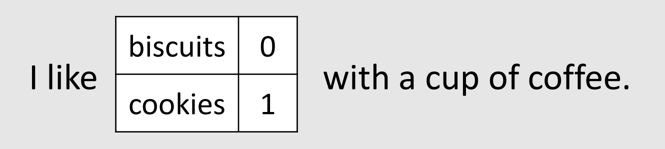


- With the proposed method, **security and payload capacity trade-off can be easily controlled** by changing the parameters of the strategies.

Existing Approaches

Edit-based methods

Modify an existing innocent cover text to encode the secret message.
Synonym substitution, paraphrase substitution, syntactic transformation, etc…

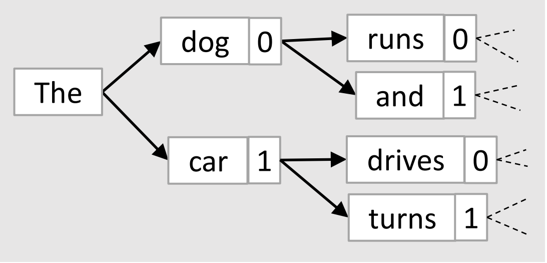


Synonym substitution example.
Message "**1**" is encoded to
"I like *cookies* with a cup of coffee."

- Low capacity. For example, 2bits per sentence (Chang and Clark, 2014)
- Painstaking construction of substitution rules to avoid linguistic phenomena such as part-of-speech ambiguity, polysemy, and context-sensitivity.

Generation-based methods

Directly assign bit chunks to the output of language models (LMs).



Generation-based method example.
Message "**10**…" is encoded to
"**The car drives**…"

- High capacity. For example, 1-5 bits per word (Shen et al., 2020)
- It remains challenging for an LM to generate so genuine-looking texts that they fool both humans and machines.

Our code is available on GitHub

https://github.com/ku-nlp/steganography-with-masked-lm

KYOTO UNIVERSITY