Lost in Translation: Translation-based Steganography

Christian Grothoff, Krista Grothoff,
Ludmila Alkhutova, Ryan Stutsman and Mikhail Atallah

{christian,krista}@grothoff.org, {lalkhuto,rstutsma}@purdue.edu, mja@cs.purdue.edu

"... because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know."

Some Approaches to Linguistic Stego

- Wayner '92: Chapman & Davida '97: handgenerated CFGs, automatically generated syntactic templates to produce syntactically correct text
- Chapman, Davida & Rennhard '01: Synonym replacement using existing texts

Some Disadvantages to these Approaches

- Hand-generation of grammars labor intensive (solved with automatic template generation)
- semantic coherence can be problematic (CFGs)
- Not all synonyms are created equal (e.g. *eat* vs. *devour*); good lists must be hand-generated (NICETEXT II)
- Additionally, pure semantic substitution may be subject to known-cover and diff attacks (NICETEXT II)

Why do these problems arise?

- Automatic generation of semantically and rhetorically correct text is difficult on its own
- Each of these approaches attempts to *mimic correct text*
- Incorrect text becomes a source of deviation from the statistical profile of what is mimicked

"... hide the identity of a text by recoding a file so its statistical profile approximates the statistical profile of another file." – Peter Wayner

Solving the Generation Problem

- If the problem is with mimicking correct text...
- Find a stego object type which:
  - Is *expected* to be semantically and syntactically damaged
  - Is *supposed* to be a transformation of the original object – both can coexist without a problem
  - *By nature* contains errors which often causes it to make less-than-perfect sense

"In order to prevent significant changes of the cover material, most steganographic algorithms try to utilize noise introduced by usual processes." – E. Franz and A. Schneidewind

An Example from Babelfish

- The following German text was taken from a Linux Camp website: "Keine Sorge, sie sind alle handzahm und beantworten auch bereitwillig Fragen rund um das Thema Linux und geben gerne einen kleinen Einblick in die Welt der Open-Source."
- A reasonable English translation would be the following: "Don't worry, they are all tame and will also readily answer questions regarding the topic 'Linux' and gladly give a small glimpse into the world of Open Source."
- Babelfish gave the following translation: "A concern, it are not all handzahm and also readily questions approximately around the topic Linux and give gladly a small idea of the world of the open SOURCE."

Translation as a Cover

Natural Language (NL) translation is an inherently noisy process (MT moreso than human translation)

- Ready availability of low-quality translations makes certain alterations plausible and errors easy to mimic
- Redundant nature of language means that translation allows for a wide variety of outputs
- Variation of a translation does not necessarily constitute "damage"

Natural Language Machine Translation

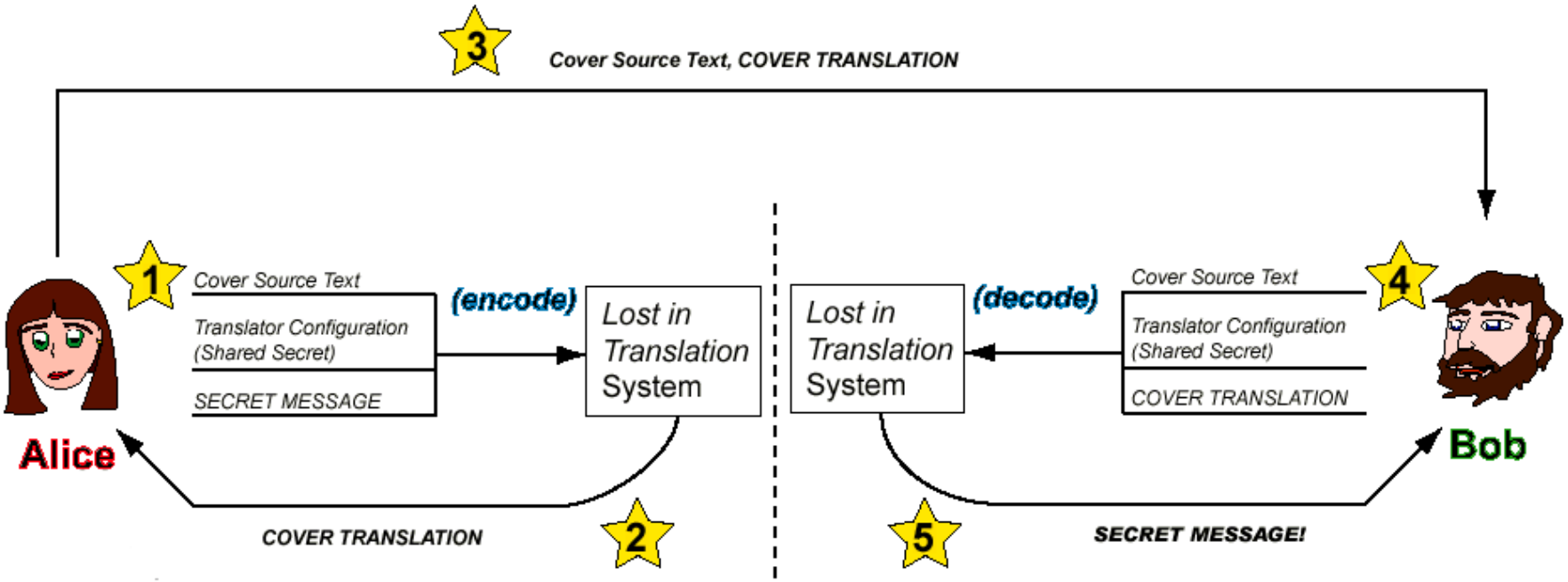
- Far from perfect
- Most systems are statistical engines – translate via pattern matching and sets of syntactic rules
- Context is usually completely neglected
- Translations often word-for-word, ignoring syntactic and semantic differences between source and target languages

Lost in Translation (LiT): a Translation-Based Steganographic System

We assume Alice and Bob have a shared secret in advance – in this case, it is the translation-system configuration.

To send a message, Alice first chooses a source text – it might be from a public text source. It does *not* have to be secret.

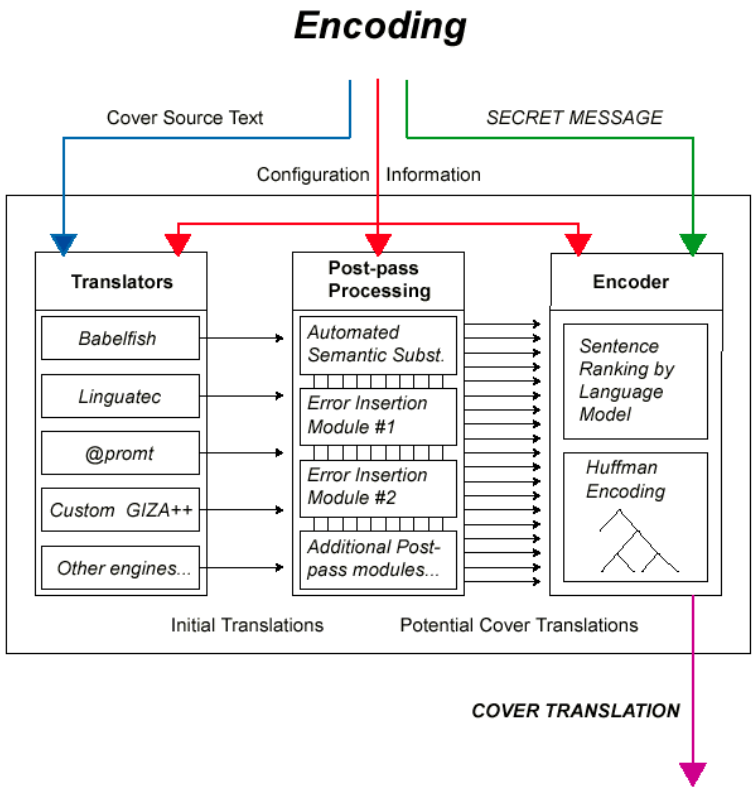
Protocol Overview



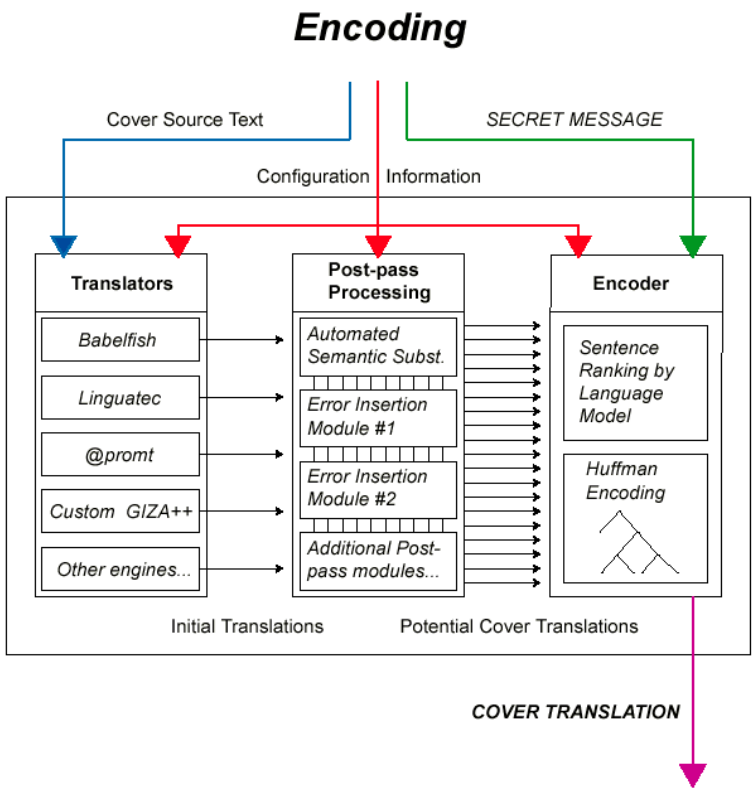
The System: Encoding

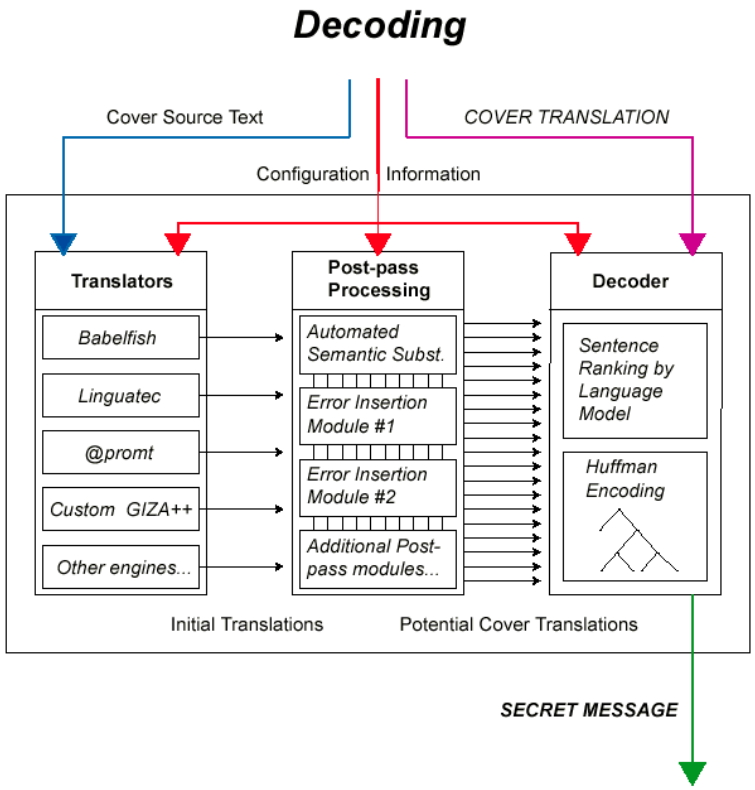
- Cover source text is run through several commercial and custom-generated translation engines
- Errors, semantic substitutions, and other modifications are made to these translations in a post-processing step – each modification is considered damage
- Each damaging action reduces the probability that a sentence looks like real translation – language model decides what modifications cause more damage
- Accumulated probabilities are used to build a Huffman tree – matching bit sequence from the secret message determines which translation sentence will be chosen

Encoder


Encoding
Cover Source Text
SECRET MESSAGE
Configuration Information
Translators
Babelfish
Linguatec
@promt
Custom GIZA++
Other engines...
Post-pass Processing
Automated Semantic Subst.
Error Insertion Module #1
Error Insertion Module #2
Additional Post-pass modules...
Encoder
Sentence Ranking by Language Model
Huffman Encoding
Initial Translations
Potential Cover Translations
COVER TRANSLATION

Encoder and Decoder



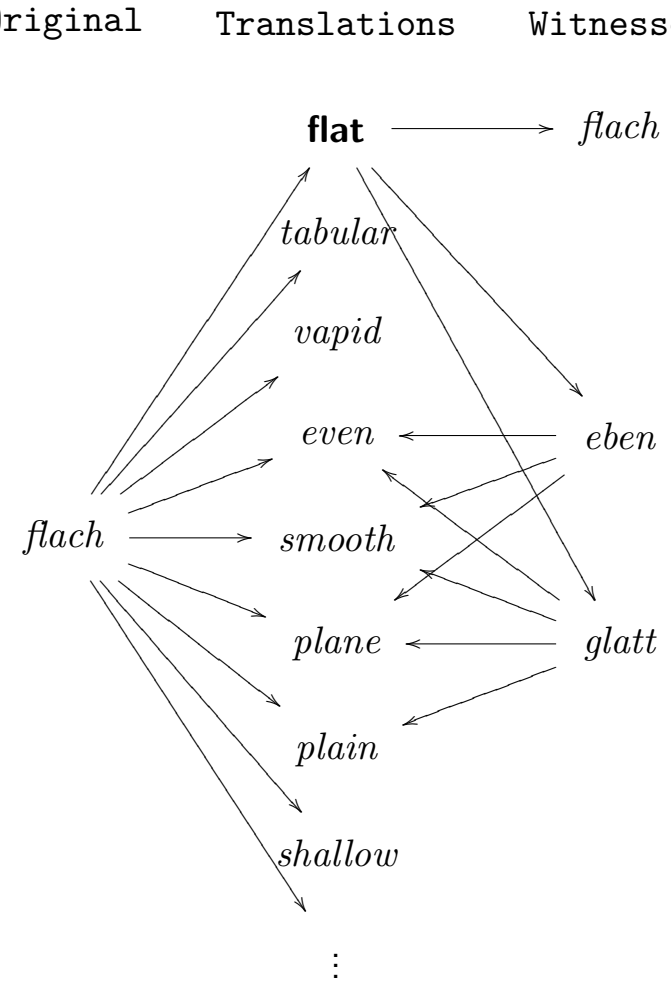


Post-pass Example: Error Insertion

Simple examples for errors when translating to English:

- Incorrect use of articles (definite/indefinite, incorrect omission/inclusion of articles)
- Prepositions are particularly tricky – because they have so many meanings, mapping them correctly is hard
- Leave less common words in their original language ("handzahm")

Post-pass Example: Semantic Substitution



About Post-Passes

- Which modules are run is determined by the (shared secret) system configuration
- New modules can be created and plugged in by the user
- This is where error insertion, error correction, semantic substitution, and any other transformation that mimics legitimate MT systems occur

Experimental Results: Translations

Original: "In dieser Zeit soll festgestellt werden, ob die Schüler die richtige Schule gewählt haben und ob sie ihren Fähigkeiten entspricht."

Google: "In this time it is to be determined whether the pupils selected the correct school and whether it corresponds to its abilities."

Linguatec: "Whether the pupils have chosen the right school and whether it corresponds to its abilities shall be found out at this time."

LiT: "In this time it is toward be determined whether pupils selected a correct school and whether it corresponds toward its abilities." (8 bits hidden)

Experimental Results: Translations

Original: "Der marokkanische Film "Windhorse" erzählt die geschichte zweier, unterschiedlichen Generationen angehörender Männer, die durch Marokko reisen. Auf dem Weg suchen sie nach dem Einzigen, was ihnen wichtig ist: dem Sinn des Lebens."

Google: "The Moroccan film "Windhorse" tells the history of two, different generations of belonging men, who travel by Morocco. On the way they look for the none one, which is important to them: the sense of the life."

Linguatec: "The Moroccan film "Windhorse" tells the story of men belonging to two, different generations who travel through Morocco. They are looking for the only one which is important to them on the way: the meaning of the life."

LiT: "The Moroccan film "Windhorse" tells story from men belonging by two, different generations who travel through Morocco. They are looking for the only one which is important to them on the way: the sense of a life."

The sentence above hides the message "lit" (24 bits).

Advantages

- LiT hides within the limits of MT, as MT models change, so can our system
- The generation problem is avoided by mimicking the results of an imperfect transformation, *not* correct, human-produced text
- Secret key (implementation, training corpora and configuration) allows for many encoders
- Cover text can be public and obtained from public sources

Disadvantages

- low bitrate ($\log_2 n$ bits per sentence for n translations)
- need to transmit both source text (or a reference to it) and translation

Increasing the Bitrate

The bitrate can be increased by:

- implementing more MT systems
- creating new corpora to train existing MT implementations
- performing additional, plausible modifications (pre- and post-passes) to the translation system in order to obtain additional variants

Experimental Results: Bitrate

- Bitrate is for a prototype
  - Limited dictionaries
  - No build-in knowledge about grammar or semantics
  - Few translation engines
- Low information density of text $\Rightarrow$ compression
- Highest bitrate achieved: 0.0082/0.022

Use for Watermarking (1/2)

- Read mark from marked copy only
  - Original text is not available
  - No reference translation is available
- LSB(Keyed Hash(sentences)) = mark bit
  - Modify until equal to mark bit
  - Different sentences for every mark bit

Use for Watermarking (2/2)

- Which sentences?
- Key directly selects mark bits' locations
  - Simple
  - Fragile
- More robust: Use of "marker" sentences
  - Mark bit is in sentences that follow marker
  - Secret ranking of sentences
  - Lowest-ranked are markers

Attacks

An adversary could attack the protocol by:

- spotting obvious inconsistencies:
  - same sentence translated in two ways
  - certain mistakes made inconsistently ("foots")
- constructing some new statistical model for languages that *all* translation systems obey, except for the steganographic encoder.

White-box Security

- Given such a new statistical model, it is *easy* to modify the steganographic encoder to become model-aware (i.e. produce sentences consistent with the model)
- Creating new models is equivalent to improving (statistical) machine translation.
- Attacking the protocol becomes an arms race in terms of understanding (machine) translation. Given equal knowledge, the defender wins.

> "Of course, the quality of the model influences the security of the steganographic algorithm – if an attacker possesses a better model (...) he is able to distinguish between stego images and steganographically unused data." – E. Franz and A. Schneidewind

Avoiding Transmission of the Original

- Receiver and sender agree on small constant h.
- Receiver computes keyed hash of translation, lowest h bits say how many bits of message are in rest of hash.
- Encoding is purely statistical and unlikely to fail if h small and number of available translations t large:

$$\left(1 - \frac{1}{2^h} \cdot \sum_{i=0}^{2^h-1} \frac{1}{2^i}\right)^t . \tag{1}$$

- Use FEC to correct encoding errors.

Conclusions

- Translation-based steganography is a promising new approach for text steganography.
- The bit-rate that can be achieved is lower than that of systems operating on binary data.
- Statistical attacks can be defeated if the underlying statistical language model is made public.
- Machine translation is not dead.

Copyright