



Restoration of Homoglyph Attack with BERT (3)



AutoML Lab

Hanyong Lee

2022. 2. 15.

Restoration of Homoglyph Attack with BERT (3)

✓ To Do List

- Last Week
 1. Related studies
 2. Setting the baseline
- This Week
 1. Making Translation Dictionary for baseline algorithms
 2. Preprocessing the dataset

Restoration of Homoglyph Attack with BERT (3)

1. Translation Dictionary (SimChar)

- Making Translation Dictionary based on **SimChar database**
- Reference: *H. Suzuki, D. Chiba, Y. Yoneya, T. Mori, and S. Goto, "ShamFinder: An Automated Framework for Detecting IDN Homographs," Proceedings of the 19th ACM Internet Measurement Conference (IMC 2019), pp. 449--462, October 2019*
- SimChar Dictionary creation algorithm

```
D := empty dict object
for char in SimChar_DB:
    if char is non-ASCII:
         $\Omega \stackrel{\text{def}}{=} \text{A set of homoglyphs of char (similar\_char array)}$ 
        sort  $\Omega$  by delta in ascending order
        for  $\omega$  in  $\Omega$ :
            if  $\omega$  is ascii and  $\omega$  is alphabet:
                D[char] :=  $\omega$ 
                break
```

Restoration of Homoglyph Attack with BERT (3)

1. Translation Dictionary – Dictionary File Sample

- {"à": "a", "á": "a", "â": "a", "ã": "a", "ä": "a", "å": "a", "ç": "c", "è": "e", "é": "e", "ê": "e", "ë": "e", "ì": "i", "í": "i", "î": "i", "ï": "i", "ñ": "n", "ò": "o", "ó": "o", "ô": "o", "õ": "o", "ö": "o", "ù": "u", "ú": "u", "û": "u", "ü": "u", "ý": "y", "þ": "b", "ÿ": "y", "ā": "a", "ą": "a", "ć": "c", "ĉ": "c", "ċ": "c", "č": "c", "ď": "d", "đ": "d", "ē": "e", "è": "e", "ę": "e", "ě": "e", "ĝ": "g", "ğ": "g", "ġ": "g", "ĥ": "h", "ĥ": "h", "ĩ": "i", "ī": "i", "ĵ": "j", "ĳ": "j", "Ĺ": "l", "Ľ": "l", "ŀ": "l", "ł": "l", "ń": "n", "ņ": "n", "ň": "n", "ŋ": "n", "ō": "o", "ó": "o", "rí": "r", "ŕ": "r", "ř": "r", "ś": "s", "ŝ": "s", "ș": "s", "š": "s", "ţ": "t", "ť": "t", "ţ": "t", "ũ": "u", "ū": "u", "û": "u", "ú": "u", "ų": "u", "w": "w", "y": "y", "z": "z", "ż": "z", "ž": "z", "b": "b", "b": "b", "d": "d", "f": "f", "k": "k", "l": "l", "n": "n", "t": "t", "t": "t", "a": "a", "i": "i", "o": "o", "u": "u", "o": "o", "ā": "a", "g": "g", "g": "g", "k": "k", "o": "o", "j": "j", "g": "g", "n": "n", "a": "a", "e": "e", "o": "o", "r": "r", "u": "u", "s": "s", "t": "t", "d": "d", "z": "z", "a": "a", "e": "e", "o": "o", ...

Restoration of Homoglyph Attack with BERT (3)

2. Translation Dictionary (OCR)

- Making Translation Dictionary using **OCR**
- Reference: *Sawabe, Yuta, et al. "Detection method of homoglyph internationalized domain names with OCR." Journal of Information Processing 27 (2019): 536-544.*
- Unicode Character Image Dataset information
 - 128 pixels × 128 pixels images
 - Using *GoNotoCurrent.ttf* TrueType font
 - <https://github.com/satbyy/go-noto-universal>
 - Unsupported Unicode characters by the font are excluded.
 - Character groups that are distinct from alphabets are excluded.

Restoration of Homoglyph Attack with BERT (3)

2. Translation Dictionary (OCR) – Excluded Unicode Character Groups

- 0 BMP
 - Basic Latin
 - CJK Unified Ideographs Extension A
 - CJK Unified Ideographs
 - Hangul Syllables
 - Hangul Jamo Extended-B
 - High Surrogates
 - High Private Use Surrogates
 - Low Surrogates
 - Private Use Area
 - CJK Compatibility Ideographs
 - Specials
- 1 SMP
 - Cuneiform
 - Egyptian Hieroglyphs
 - Egyptian Hieroglyph Format Controls
 - Anatolian Hieroglyphs
 - Bamum Supplement
 - Tangut
 - Tangut Components
 - Khitan Small Script
 - Tangut Supplement
 - Nushu
 - Mahjong Tiles
 - Domino Tiles
 - Playing Cards
 - (Empty part of) Enclosed Alphanumeric Supplement
 - Enclosed Ideographic Supplement
 - Miscellaneous Symbols and Pictographs
- **Emoticons**
 - Ornamental Dingbats
 - Transport and Map Symbols
 - Supplemental Arrows-C
 - Supplemental Symbols and Pictographs
 - Chess Symbols
 - Symbols and Pictographs Extended-A
- All characters of 2 SIP and 3 TIP
- 14 SSP
 - Variation Selectors Supplement
- All characters of 15 PUA-A and 15 PUA-A
- + **Unsupported characters by the font**

Restoration of Homoglyph Attack with BERT (3)

2. Translation Dictionary (OCR) – Dictionary File Sample

- {"i": "I", "¢": "C", "£": "f", "¤": "e", "¥": "X", "¦": "I", "§": "e", "¨": "a", "©": "K", "ª": "O", "±": "t", "²": "a", "µ": "U", "¶": "q", "·": "e", "°": "O", "¼": "Y", "½": "i", "¾": "H", "¿": "d", "À": "A", "Á": "A", "Â": "A", "Ã": "A", "Ä": "A", "Å": "A", "Æ": "A", "Ç": "C", "È": "E", "É": "E", "Ê": "E", "Ë": "E", "Ì": "I", "Í": "I", "Î": "I", "Ï": "i", "Ñ": "N", "Ò": "O", "Ó": "O", "Ô": "O", "Õ": "O", "Ö": "O", "×": "x", "Ù": "U", "Ú": "U", "Û": "O", "Ü": "V", "Ý": "Y", "Þ": "b", "ß": "K", "à": "a", "á": "a", "â": "a", "ã": "a", "ä": "a", "æ": "f", "ç": "C", "è": "e", "é": "e", "ê": "e", "ë": "a", "î": "T", "ï": "o", "ð": "O", "ñ": "n", "ò": "f", "ó": "O", "ù": "u", "ú": "U", "û": "d", "ü": "r", "ý": "y", "þ": "p", "ÿ": "y", "Ā": "A", "ā": "a", "Ă": "A", "ă": "a", "ą": "q", "Ć": "C", "ć": "C", "Ĉ": "C", "ĉ": "C", "Ċ": "C", "ċ": "C", "Č": "C", "č": "C", "Ď": "D", "ď": "d", "đ": "d", "Ē": "E", "ē": "e", "Ě": "E", "ě": "e", "Ė": "E", "ė": "e", "Ę": "E", "ę": "E", "ě": "e", "Ĝ": "G", "ĝ": "g", "Ğ": "G", "ğ": "g", "Ġ": "G", "ġ": "g", "Ĥ": "a", "ĥ": "h", ...

Restoration of Homoglyph Attack with BERT (3)

3. Creating Baseline Algorithm Evaluation Dataset

- Evaluate baseline algorithm with homoglyph/hand-annotated sentences pairs.
- The example of evaluate datasets: (Homoglyph characters are highlighted.)

you have less than 48 hours from
the moment you opened this email
(precisely two days).
Here is my Bitcoin wallet

homoglyph sentences
(input)

you have less than 48 hours from
the moment you opened this email
(precisely two days).
Here is my Bitcoin wallet

hand-annotated sentences
(ground truth)

Restoration of Homoglyph Attack with BERT (3)

4. Text Preprocessing

- Replacing space characters
 - e.g., U+00A0 (NO-BREAK SPACE), U+2002 (EN SPACE) – U+200A (HAIR SPACE), U+3000 (IDEOGRAPHIC SPACE)
- Removing nonprinting characters
 - e.g., U+200B (ZERO WIDTH SPACE), U+200C (ZERO WIDTH NON-JOINER), U+061C (ARABIC LETTER MARK), U+FEFF (ZERO WIDTH NO-BREAK SPACE), U+00AD(SOFT HYPHEN)
 - Control Characters (U+0000, U+0006 – U+0008, U+000B – U+001F, U+0080 – U+009F)
 - Some Control Characters(U+0001 – U+0005) are used as homoglyphs.
e.g., I give you 72 houus onuly to transfer the funds.
 - Private Use Area (U+E000 – U+F8FF)