Lost in Translation
Characterizing Automated Censorship in Online Translation Services

Samuel Ruo, Jeffrey Knockel (presenting), and Zoë Reichert
Hi, how’s it going!!! ;)

こんにちは、元気ですか？

こんにちは、元気ですか？
## Chinese translation platforms analyzed

<table>
<thead>
<tr>
<th>Service</th>
<th>Total</th>
<th>% CN</th>
<th>% US</th>
<th>% TW</th>
<th>% HK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba</td>
<td>100.2K</td>
<td>85.3</td>
<td>6.50</td>
<td>1.35</td>
<td>1.28</td>
</tr>
<tr>
<td>Baidu</td>
<td>52.9M</td>
<td>91.1</td>
<td>1.52</td>
<td>1.80</td>
<td>1.56</td>
</tr>
<tr>
<td>Tencent</td>
<td>791.4K</td>
<td>89.3</td>
<td>2.85</td>
<td>3.35</td>
<td>N/A</td>
</tr>
<tr>
<td>Youdao</td>
<td>7.2M</td>
<td>86.2</td>
<td>2.55</td>
<td>1.96</td>
<td>2.39</td>
</tr>
</tbody>
</table>
Microsoft’s Bing Translate

(1) 蔡英文是台湾总统。(2) 习近平是中国国家主席。(3) 唐纳德特朗普是美国总统。
Modeling censorship rules

(ordered, keyword, combinations)

Covers:

- singletons
- .*common.*regular.*expressions.*
- {unordered, keyword, combinations}

ordered + keyword + combination
Automated algorithm

Source: “President Xi Jinping gave a motivating speech despite the bad weather.”

Censorship rule: Xi + Jinping + bad

Censorship oracle: censored (❌) or not (✅)?

❌ President Xi Jinping gave a motivating speech despite the bad weather
Binary search for end of component

✅ President Xi Jinping gave an motivating speech despite the bad weather.

❌ President Xi Jinping gave an motivating speech despite the bad weather.

❌ President Xi Jinping gave an motivating speech despite the bad weather.

❌ President Xi Jinping gave an motivating speech despite the bad weather.
Linear search for beginning (typically faster than binary)

✅ President Xi Jinping gave a motivating speech despite the bad
✅ President Xi Jinping gave a motivating speech despite the bad
❌ President Xi Jinping gave a motivating speech despite the bad

✅ bad (← incomplete rule)
Binary search for end of component (#2)

❌ PresidentXiJinpinggaveamo%o bad
✔️ PresidentXiJ%o bad
❌ PresidentXiJinping%o bad
✔️ PresidentXiJinp%o bad
✔️ PresidentXiJinpi%o bad
✔️ PresidentXiJinpin%o bad (← bisected end)
Linear search for beginning (#2)

✅ PresidentXiJinpin%o%ng%o%bad
✅ PresidentXiJinpin%o%ing%o%bad
✅ PresidentXiJinpin%o%ping%o%bad
✅ PresidentXiJinpin%o%nping%o%bad
✅ PresidentXiJinpin%o%inping%o%bad
❌ PresidentXiJinpin%o%Jinping%o%bad (← found beginning)

✅ Jinping%o%bad (← incomplete rule)
Binary search for end of component (#3)

✓ President Jinping bad
✗ President Xi Jinping bad
✓ President X Jinping bad
✗ President Xi Jinping bad
Linear scan for beginning (#3)

✅ PresidentXi Jinping bad
❌ PresidentXi Jinping bad
❌ Xi Jinping bad (← complete rule 🎉🎉🎉️)
“President Xi Jinping gave a motivating speech despite the bad weather.”

Final output:

Xi + Jinping + bad
Testing keywords one-by-one

- Slow!
- Limited to simple rules!
- INACCURATE!
Our algorithm

- Faster!
- Discover complicated rules!
- ACCURATE!
Weibo has apparently censored Canadian PM Justin Trudeau’s name. Search his name and you’d only see posts from official accounts (Blue V).

And his nickname 小土豆 (little potato) too. So is English word “potato”.
(1) 蔡英文是台湾总统。 (2) 习近平是中国国家主席。 (3) 唐纳德特朗普是美国总统。

(1) Tsai Ing wen is the President of Taiwan. (2) Donald Trump is the President of the United States.
【网络民议】洞庭湖大堤决口合龙，救灾现场变表演舞台

2024年7月9日

大堤决口合龙一刻，现场上演了魔幻一幕：有大量手持红旗的救援人员向推土机聚拢，并不断挥舞旗帜，将夜晚的救灾现场搞出了舞台效果。这一红旗招展的画面引发了大量网民的吐槽，称之为“丧事喜办”的又一经典场景。
505,903 general rules
<table>
<thead>
<tr>
<th>Service</th>
<th>Detection mechanism</th>
<th># (people)</th>
<th># (general)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba</td>
<td>Keyword combination</td>
<td>344</td>
<td>N/A</td>
</tr>
<tr>
<td>Baidu</td>
<td>Keyword</td>
<td>3</td>
<td>131</td>
</tr>
<tr>
<td>Bing</td>
<td>Regular expression</td>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td>Tencent</td>
<td>Keyword combination</td>
<td>4</td>
<td>2,452</td>
</tr>
<tr>
<td>Youdao</td>
<td>Multiple types</td>
<td>45</td>
<td>9,414</td>
</tr>
</tbody>
</table>
Youdao’s ML rules…

胡主 + 胡錦 + 錦濤 (Lord Hu + Hu Jin + Jintao, an unintuitive rule targeting Hu Jintao criticism)

他操操你操操蛋操朝鮮操比毛操 (not translatable, but contains the characters 操 “fuck” and 毛 “Mao [Zedong]”)

螺 + 螺 + 螺 + 螺 + 螺 + 螺 + 蝦 + D + 哑 + 大 (screw + screw + screw + screw + screw + screw + screw + [a homophone of Xi] + D + [a homophone of 大] + 大)
Practical implications

- Microsoft Bing had less censorship
- Although most extreme method, also most obvious method of censorship
- Bing censored fewest categories
- Bing being better — far from a foregone conclusion!
<table>
<thead>
<tr>
<th>Service</th>
<th>Total</th>
<th>% CN</th>
<th>% US</th>
<th>% TW</th>
<th>% HK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba</td>
<td>100.2K</td>
<td>85.3</td>
<td>6.50</td>
<td>1.35</td>
<td>1.28</td>
</tr>
<tr>
<td>Baidu</td>
<td>52.9M</td>
<td>91.1</td>
<td>1.52</td>
<td>1.80</td>
<td>1.56</td>
</tr>
<tr>
<td>Tencent</td>
<td>791.4K</td>
<td>89.3</td>
<td>2.85</td>
<td>3.35</td>
<td>N/A</td>
</tr>
<tr>
<td>Youdao</td>
<td>7.2M</td>
<td>86.2</td>
<td>2.55</td>
<td>1.96</td>
<td>2.39</td>
</tr>
</tbody>
</table>
Questions?

All discovered censorship rules:

https://github.com/citizenlab/chat-censorship/tree/master/translator