Training efficient and open source ML models for private translation in Firefox

And how you can help.

Marco Castelluccio <marco@mozilla.com>
FOSDEM 2024
BACKGROUND
• ~64% of the World Wide Web is in English

• Only ~26% of users are native English speakers
24 official languages in Europe 🇪🇺
93 languages in EU 🇪🇺 spoken by more than 100k people
Why privacy-friendly translations

- Current cloud-based services do not offer the **privacy guarantees** that we like to offer in Firefox
Privacy-friendly translations

- **European Union-funded project** to investigate client-side private translation capabilities
Languages

Released

- Bulgarian, Dutch, English, French, German, Italian, Polish, Portuguese, Spanish

In development

- Catalan, Czech, Estonian, Finnish, Greek, Hungarian, Icelandic, Lithuanian, Maltese, Norwegian Bokmål, Norwegian Nynorsk, Persian (Farsi), Russian, Slovenian, Turkish, Ukrainian
INFERENCES
Model download

- Firefox will download and update the model from our servers - just for the languages that you need
Run models locally

- Use **machine learning on the client side** to translate locally
- All the data remains on the system
WASM engine with SIMD optimizations

- NMT engine compiled to **WebAssembly**

- Using **SIMD instructions** via **gemmology**
  - 10x perf improvement
TRAINING
Open Datasets

An overview of the OPUS collection

1,210 CORPORA

45,945,946,108 TOTAL SENTENCE PAIRS

744 LANGUAGES AVAILABLE

THIS MAP DISPLAYS 10 CORPORA, WHICH MAKE UP A TOTAL 93.40% OF THE ENTIRE OPUS COLLECTION

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentences</th>
<th>% of OPUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLLB</td>
<td>13B</td>
<td>28.31</td>
</tr>
<tr>
<td>CCMatrix</td>
<td>11B</td>
<td>23.64</td>
</tr>
<tr>
<td>OpenSubtitles</td>
<td>8.5B</td>
<td>18.53</td>
</tr>
<tr>
<td>MultiCCAAligned</td>
<td>2.2B</td>
<td>4.87840</td>
</tr>
<tr>
<td>ParaCrawl</td>
<td>1.5B</td>
<td>3.26229</td>
</tr>
<tr>
<td>DGT</td>
<td>1.1B</td>
<td>2.37845</td>
</tr>
<tr>
<td>XLEnt</td>
<td>883M</td>
<td>1.92148</td>
</tr>
<tr>
<td>MultiParaCrawl</td>
<td>789M</td>
<td>1.71653</td>
</tr>
<tr>
<td>LinguaTools-WikiTitles</td>
<td>487M</td>
<td>1.06082</td>
</tr>
<tr>
<td>CCAigned</td>
<td>439M</td>
<td>0.95442</td>
</tr>
</tbody>
</table>

https://opus.nlpl.eu/
Data cleaning and augmentation

- **Rule-based data cleaning** + ML-based detection of bad sentence pairs (bicleaner-ai)

- **Back-translation** for data augmentation from monolingual data
  - Translate target to source
  - Add pair “noisy translated source” -> “good target”
Model training and compression

- **Training** of a large model

- Compression using **knowledge distillation** into a smaller one

- **Quantization** (float32 -> int8) for further compression and perf improvement

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>Total number of parameters</th>
<th>Dataset decoding time on 1 CPU core</th>
<th>Quality, BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>798Mb</td>
<td>192.75M</td>
<td>631s</td>
<td>52.5</td>
</tr>
<tr>
<td>Student quantized</td>
<td>17Mb</td>
<td>15.7M</td>
<td>17.9s</td>
<td>50.7</td>
</tr>
</tbody>
</table>
https://mozilla.github.io/firefox-translations-training/pipeline-steps.html
NEXT STEPS
Improving quality

### Fix numbers

The shirt costs $25.73

La camisa cuesta 24,39€.

### Fix capitalization

![SOMETIMES THINGS ARE IN ALL CAPS](image)

### The web is messy!

Our training data needs sto ebee messsyy 😞

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Model</th>
<th>Test Dataset</th>
<th>Augmentation of test dataset</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>No OpusTrainer augmentation</td>
<td>teacher</td>
<td>sacrebleu_wmt19</td>
<td>all upper case</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>teacher</td>
<td>sacrebleu_wmt19</td>
<td>all title case</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>teacher</td>
<td>sacrebleu_wmt19</td>
<td>no</td>
<td>31.5</td>
</tr>
<tr>
<td>Upper case and title case augmentation with OpusTrainer</td>
<td>teacher</td>
<td>sacrebleu_wmt19</td>
<td>all title case</td>
<td>31.8</td>
</tr>
<tr>
<td></td>
<td>teacher</td>
<td>sacrebleu_wmt19</td>
<td>all upper case</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>teacher</td>
<td>sacrebleu_wmt19</td>
<td>no</td>
<td>31.6</td>
</tr>
<tr>
<td></td>
<td>student</td>
<td>sacrebleu_wmt19</td>
<td>no</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td>student</td>
<td>sacrebleu_wmt19</td>
<td>all upper case</td>
<td>28.9</td>
</tr>
<tr>
<td></td>
<td>student</td>
<td>sacrebleu_wmt19</td>
<td>all title case</td>
<td>29.8</td>
</tr>
</tbody>
</table>

Supporting more languages

Support Chinese translations

Status: Trending idea

Add Chinese as a supported language for translations in Firefox.

https://github.com/mozilla/firefox-translations-training/issues/369
Android

https://bugzilla.mozilla.org/show_bug.cgi?id=1820240
Settings improvements

https://bugzilla.mozilla.org/show_bug.cgi?id=1869015
https://bugzilla.mozilla.org/show_bug.cgi?id=1855907
Help us!

- **Know any dataset** that we could use?
- **Want to contribute code** for a Firefox feature?
- **Interested in adding support for your language?**