Using Contemporary US Government Data to Train Custom MT for COVID-19

Achim Ruopp
Polyglot Technology LLC
achim@polyglot.technology
Translation’s Role in the COVID-19 Crisis

• Gretchen McCulloch’s article “Covid-19 Is History’s Biggest Translation Challenge” in WIRED
  • Communicating health information is an essential factor in addressing this crisis
  • Familiar issue that MT currently only supports 100+ languages
    • Languages with millions of speakers are unsupported
    • Long tail of thousands of human languages are unsupported/endangered
  • Some issues with register in the high resource languages
    • Japanese translation of “Wash your hands” in tone of a parent instructing a child
  • People want to gist information in their languages – accurate MT can help to address disinformation

• Translators without Borders did great work in previous health crises like Ebola, but scope of COVID-19 is unprecedented
The Translation Community Coming Together

- **TAUS Corona Virus Corpora**
  - English↔French/German/Italian/Spanish/Chinese/Russian
  - Translation data (mainly) selected from existing parallel corpora with a COVID-19 specific English query corpus
    - ~ 200k-900k segments
    - [Creative Commons Attribution-NonCommercial 4.0 license](https://creativecommons.org/licenses/by-nc/4.0/)
  - SYSTRAN built **custom COVID-19 MT systems** using the data
  - The MT broker Intento
    - did an extensive human evaluation/post-editing study with a test subset of the data [https://try.inten.to/mt-evaluation-covid-domain](https://try.inten.to/mt-evaluation-covid-domain)
    - is creating a custom routing for COVID-19 content in their platform
  - customized MT with the TAUS Corona Virus Corpora
    - For only 2 out of 7 language pairs did the custom MT systems outperform the stock engines
    - Possible explanation provided by Intento: medical domain is wide – data might require clustering
    - Alternative customization with just a bilingual glossary failed
The Translation Community Coming Together

• **Translation Initiative for COVID-19** aka TICO-19
  
  • **Partners**
  
  • academia: Carnegie Mellon University, Johns Hopkins University
  
  • industry: Amazon, Appen, Facebook, Google, Microsoft, Translated
  
  • non-profit: Translators without Borders
    
    • Strong track record communicating in previous crises (e.g. Ebola, Rohingya refugee crisis) and working with the non-profit organizations
    
    • Also runs COVID-19 Community Translation Program
  
  • **Data**
  
  • TICO-19 Translation Benchmark
    
    • 30 English documents with 3071 segments/69.7k words translated to 36 languages
    
    • English→Amharic, Arabic (Modern Standard), Bengali, Chinese (Simplified), Dari, Dinka, Farsi, French (European), Hausa, Hindi, Indonesian, Kanuri, Khmer (Central), Kinyarwanda, Kurdish Kurmanji, Kurdish Sorani, Lingala, Luganda, Malay, Marathi, Myanmar, Nepali, Nigerian Fulfulde, Nuer, Oromo, Pashto, Portuguese (Brazilian), Russian, Somali, Spanish (Latin American), Swahili, Congolese Swahili, Tagalog, Tamil, Tigrinya, Urdu, Zulu
    
    • COVID-19 specific translated terminologies (from Facebook and Google)
    
    • [Creative Commons CC0](https://creativecommons.org/publicdomain/zero/1.0/) licensed
  
  • See website for links to many other COVID-19 related data projects (language data and beyond)
## TICO-19 Translation Benchmark Diversity

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU</td>
<td>are you having any shortness of breath?</td>
</tr>
<tr>
<td>PubMed</td>
<td>The basic reproductive number (R0) was 3.77 (95% CI: 3.51-4.05), and the adjusted R0 was 2.23-4.82.</td>
</tr>
<tr>
<td>Wikinews</td>
<td>By yesterday, the World Health Organization reported 1,051,635 confirmed cases, including 79,332 cases in the twenty four hours preceding 10 a.m. Central European Time (0800 UTC) on April 4.</td>
</tr>
<tr>
<td>Wikivoyage</td>
<td>Due to the spread of the disease, you are advised not to travel unless necessary, to avoid being infected, quarantined, or stranded by changing restrictions and cancelled flights.</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Drug development is the process of bringing a new infectious disease vaccine or therapeutic drug to the market once a lead compound has been identified through the process of drug discovery.</td>
</tr>
<tr>
<td>Wikisource</td>
<td>The federal government has identified 16 critical infrastructure sectors whose assets, systems, and networks, whether physical or virtual, are considered so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, economic security, public health or safety, or any combination thereof.</td>
</tr>
</tbody>
</table>

Table 2: Samples of the English source sentences for the TICO-19 benchmark.

Opportunity: Centers For Disease Control and Prevention COVID-19 Website
Information for Medical Professionals is not Translated
CDC COVID-19 Site Updates

• Site frequently updated
• Crawled on June 24, 2020 and July 24, 2020
  • June 24 crawl yielded the most parallel data of the two
• Data represents translation practices of COVID-19 health info, but not ground truth about COVID-19 virus!
Data Statistics for CDC COVID-19 Parallel Data

• Non-deduplicated (in TMX with context)

<table>
<thead>
<tr>
<th>Language</th>
<th>Segments</th>
<th>Source words</th>
<th>Target words</th>
<th>Target characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish (US)</td>
<td>79,106</td>
<td>538,842</td>
<td>696,471</td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>79,757</td>
<td>550,066</td>
<td>895,573</td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>78,824</td>
<td>537,204</td>
<td>428,979</td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>70,423</td>
<td>508,297</td>
<td>2,958,795</td>
<td></td>
</tr>
</tbody>
</table>

• Deduplicated & shuffled (TSV)

<table>
<thead>
<tr>
<th>Language</th>
<th>Segments</th>
<th>Source words</th>
<th>Target words</th>
<th>Target characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish (US)</td>
<td>15,803</td>
<td>248,780</td>
<td>310,223</td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>15,849</td>
<td>249,006</td>
<td>380,113</td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>16,532</td>
<td>262,393</td>
<td>197,402</td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>11,911</td>
<td>254,876</td>
<td>1,413,993</td>
<td></td>
</tr>
</tbody>
</table>

• Volume in between TAUS Corona Virus Corpus and TICO-19
  • Test/validation/fine tuning data
  • Data with document context in TMX
    • Document context
      • original segment order
      • source-document property groups segments
      • Better: XLIFF (used in WMT)
    • Creation date
      • Better: webpage update date

• Custom crawling code based on Bitextor
  • Non-customized ParaCrawl/Bitextor contains only 5 English-Spanish segments from the CDC in the latest September 2020 release

• Additional medium resource languages/language variants covered
  • US-Spanish (= LatAm-Spanish?)
  • Vietnamese
  • Korean
CDC COVID-19 Parallel Data Licensing

• Content
  • Public domain
  • Disclaimer: Source: CDC; Reference to specific commercial products, manufacturers, companies, or trademarks does not constitute its endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention; The public domain material is available on the agency website https://www.cdc.gov/ for no charge.

• Database/database structure, i.e. TMX/TSV
  • Made available under the Open Data Commons Attribution License: http://opendatacommons.org/licenses/by/1.0
Google AutoML Translation Customized with CDC COVID-19 Parallel Data

- Significant BLEU score increases over already high baselines
  - English→Spanish +11.27
  - English→Vietnamese +4.1
  - Confirmed results with TER and BERTScore
- Enables increased productivity in post-editing scenario
- More appropriate raw machine translations of new or revised CDC COVID-19 content
Google AutoML Translation Customized with CDC COVID-19 Parallel Data

- Customized system performs worse with TICO-19 translation benchmark

Hypotheses
- Medical domain is wide
- COVID-19 is not that “novel” from the health information translation perspective
- Domain mismatch/overfitting to CDC data
  - Topic
  - Modality – TICO-19 corpora contain transcribed speech (CMU)
  - Register: Level of politeness – translator/project dependent
  - Intent – consistent
  - Style – translator/project dependent
  - Language variant
Larger Lessons – Future Research

• For high/medium resource languages
  • MT suppliers have now optimized Transformer-based NMT
    • Many ambiguities already resolved (especially intra-sentence ambiguities)
    • MT systems robust to variations in domain
  • Additional improvements for medium resource languages from transfer learning from other languages/massively multilingual systems
  • Research on using document-level context

⇒ It becomes harder and harder to beat the baseline models with custom MT!
Larger Lessons – Future Research

• Test/development data becomes ever more important - we can’t detect if we beat the baseline if we don’t specify what we expect!
  • Evaluation data is as crucial as evaluation measures
  • Development sets, e.g. for training data selection, cannot be source-only anymore
• Opens a great opportunity to include linguists – human-in-the-loop MT
  • For the post-editing use case some MT suppliers already build this into their workflow: Lilt, ModernMT, Unbabel
• MT suppliers need to improve guidance which data sets are sufficient/good – manual experimentation is tedious/expensive
Larger Lessons – Future Research

• Low resource languages still suffer from lack of language resources
  • Again coming into clear focus in the COVID-19 crisis – resource light approaches unlikely to help
  • Investment needed – public/private?
Other Parallel Corpora from Polyglot Technology LLC

• Healthcare.gov
  • Healthcare/health insurance content
  • English→Spanish
  • Blog article from May 2019

• US Department of State news releases/announcements
  • English→Arabic, Spanish, Farsi, French, Hindi, Indonesian, Portuguese, Russian, Urdu, Vietnamese, Chinese

• Custom Crawling