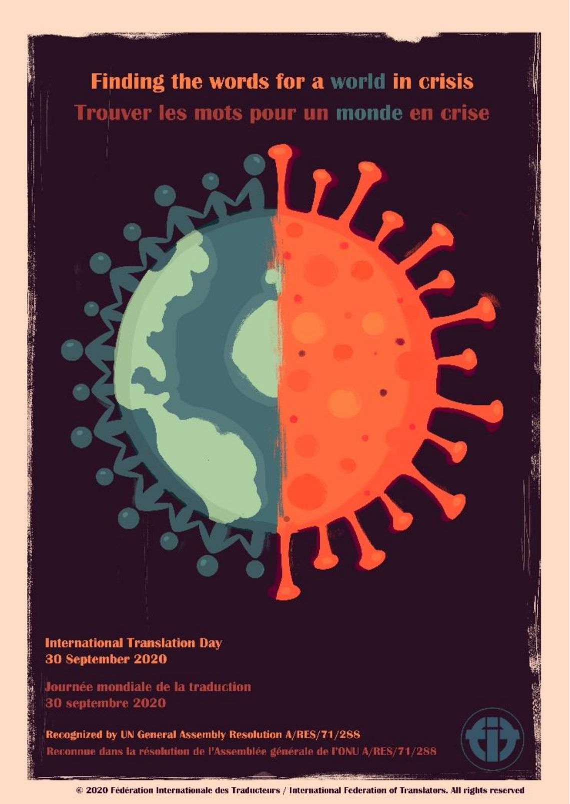
Using Contemporary US Government Data to Train Custom MT for COVID-19

Achim Ruopp

Polyglot Technology LLC

achim@polyglot.technology

Translation’s Role in the COVID-19 Crisis


Finding the words for a world in crisis
Trouver les mots pour un monde en crise
International Translation Day
30 September 2020
Journée mondiale de la traduction
30 septembre 2020
Recognized by UN General Assembly Resolution A/RES/71/288
Reconnue dans la résolution de l'Assemblée générale de l'ONU A/RES/71/288
© 2020 Fédération Internationale des Traducteurs / International Federation of Translators. All rights reserved

- 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

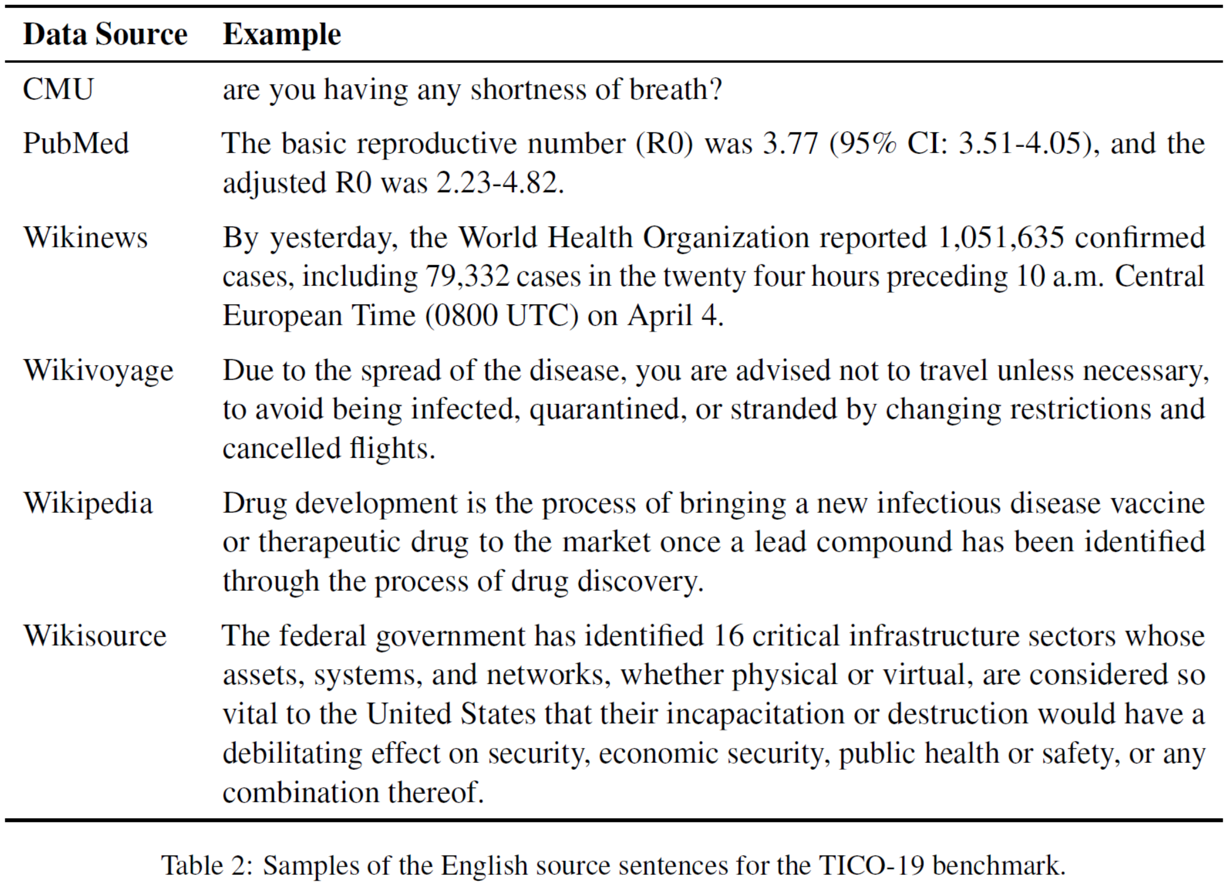
The Translation Community Coming Together

- 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
  - 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
    - 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, Rohinga 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 licensed
  - See website for links to many other COVID-19 related data projects (language data and beyond)

TICO-19 Translation Benchmark Diversity

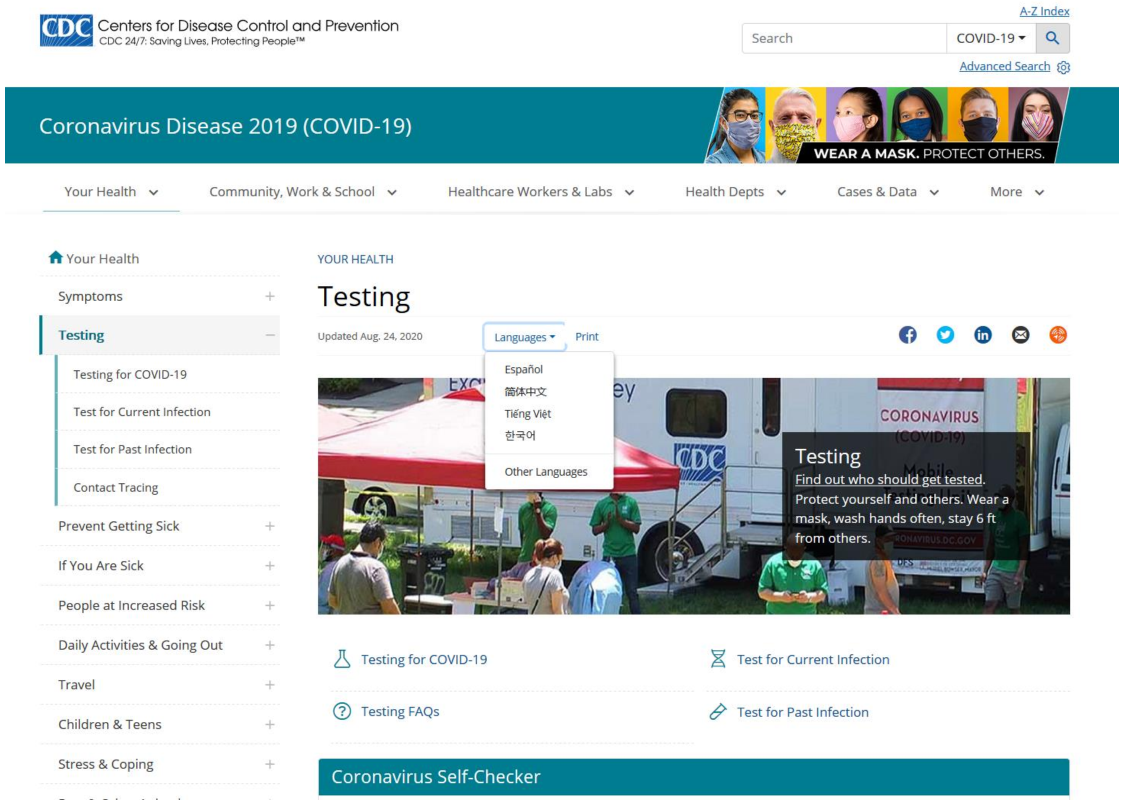


Data Source	Example
CMU	are you having any shortness of breath?
PubMed	The basic reproductive number (R0) was 3.77 (95% CI: 3.51-4.05), and the adjusted R0 was 2.23-4.82.
Wikinews	By yesterday, the World Health Organization reported 1,051,635 confirmed cases, including 79,332 cases in the twenty four hours preceding 10 a.m. Central European Time (0800 UTC) on April 4.
Wikivoyage	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.
Wikipedia	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.
Wikisource	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.

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

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., . . . Tur, S. (2020). TICO-19: the Translation Initiative for COvid-19. arXiv, 2007.01788v2.

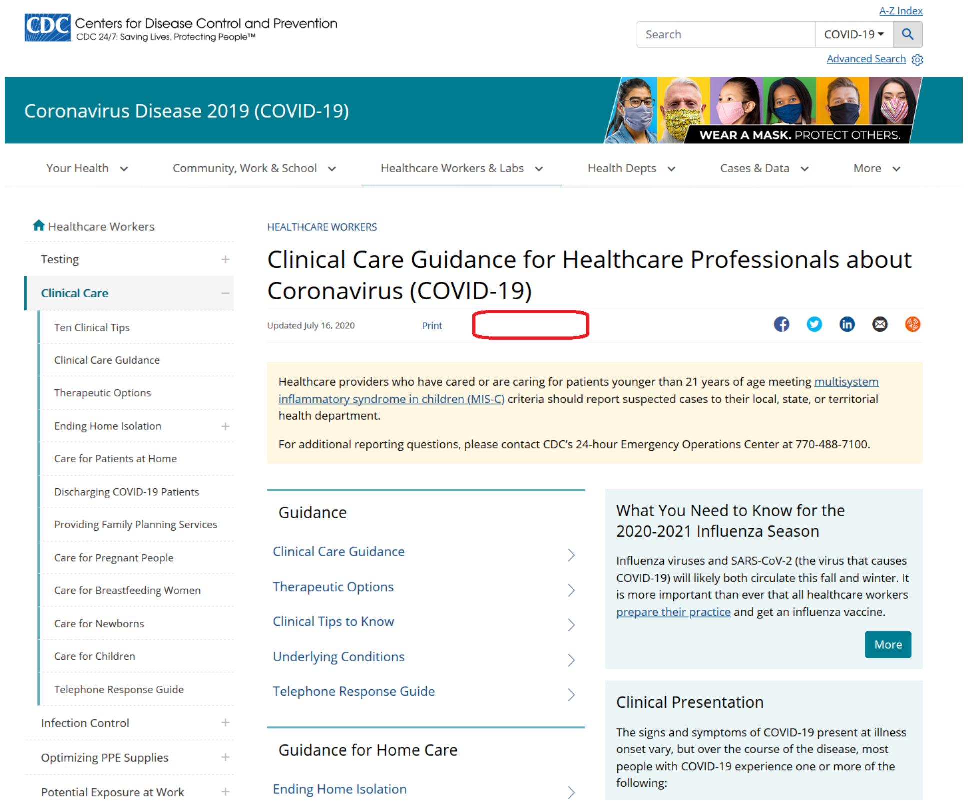
Opportunity: Centers For Disease Control and Prevention COVID-19 Website



Testing

Find out who should get tested. Protect yourself and others. Wear a mask, wash hands often, stay 6 ft from others.

Information for Medical Professionals is not Translated



HEALTHCARE WORKERS

Clinical Care Guidance for Healthcare Professionals about Coronavirus (COVID-19)

Updated July 16, 2020 Print

Healthcare providers who have cared or are caring for patients younger than 21 years of age meeting multisystem inflammatory syndrome in children (MIS-C) criteria should report suspected cases to their local, state, or territorial health department.

For additional reporting questions, please contact CDC's 24-hour Emergency Operations Center at 770-488-7100.

Guidance

- Clinical Care Guidance
- Therapeutic Options
- Clinical Tips to Know
- Underlying Conditions
- Telephone Response Guide

Guidance for Home Care

- Ending Home Isolation

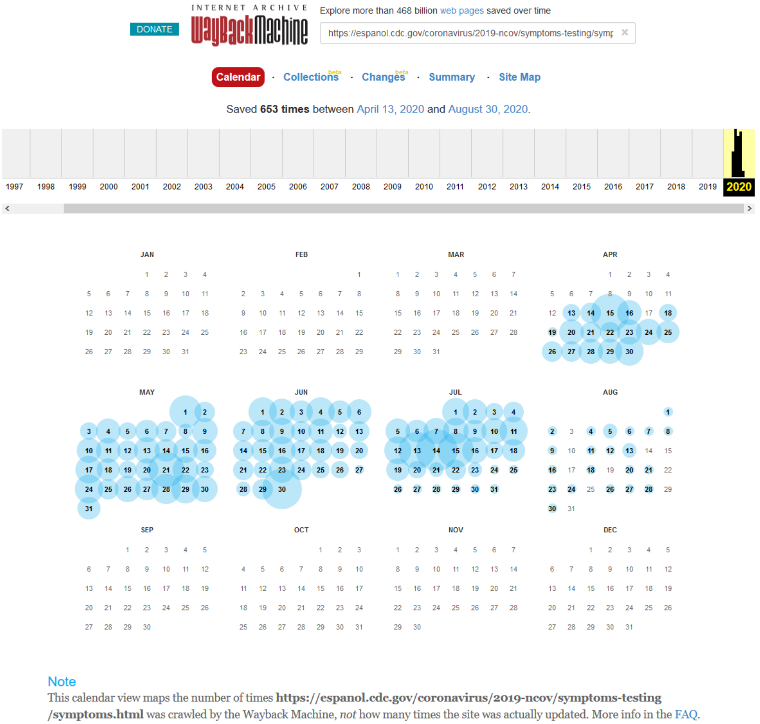
What You Need to Know for the 2020-2021 Influenza Season

Influenza viruses and SARS-CoV-2 (the virus that causes COVID-19) will likely both circulate this fall and winter. It is more important than ever that all healthcare workers prepare their practice and get an influenza vaccine.

Clinical Presentation

The signs and symptoms of COVID-19 present at illness onset vary, but over the course of the disease, most people with COVID-19 experience one or more of the following:

CDC COVID-19 Site Updates


DONATE
INTERNET ARCHIVE
WaybackMachine
https://espanol.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symp
Calendar · Collections · Changes · Summary · Site Map
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Explore more than 468 billion web pages saved over time

Saved **653 times** between April 13, 2020 and August 30, 2020.

Note

This calendar view maps the number of times **https://espanol.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

- 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)

English→	Segments	Source words	Target words	Target characters
Spanish (US)	79,106	538,842	696,471	
Vietnamese	79,757	550,066	895,573	
Korean	78,824	537,204	428,979	
Chinese	70,423	508,297		2,958,795

- Deduplicated & shuffled (TSV)

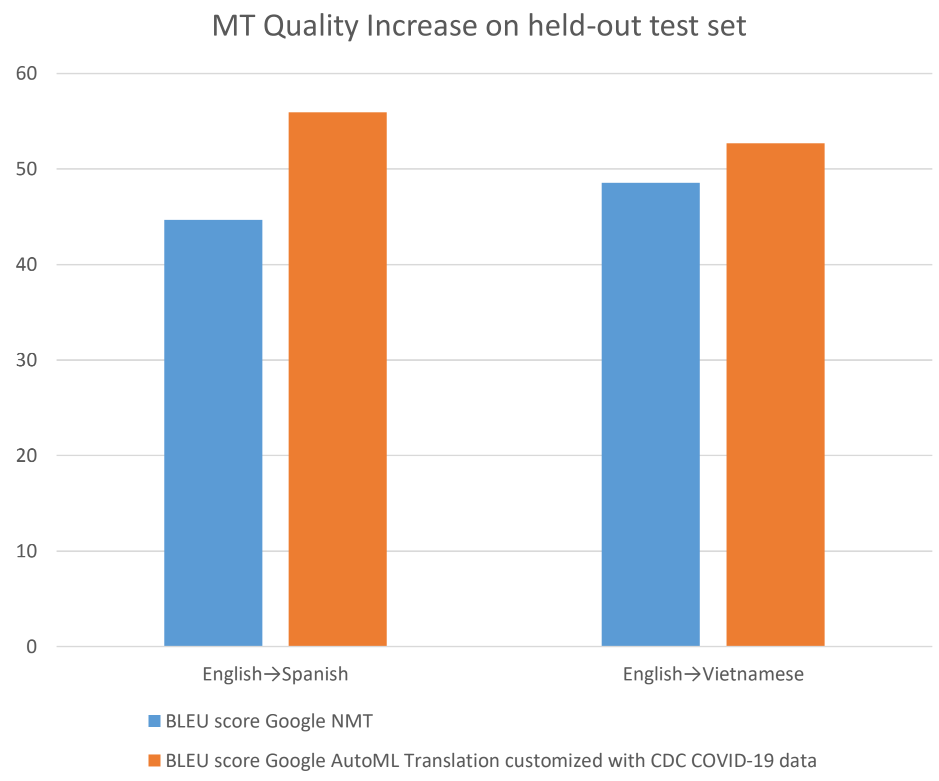
English→	Segments	Source words	Target words	Target characters
Spanish (US)	15,803	248,780	310,223	
Vietnamese	15,849	249,006	380,113	
Korean	16,532	262,393	197,402	
Chinese	11,911	254,876		1,413,993

- 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

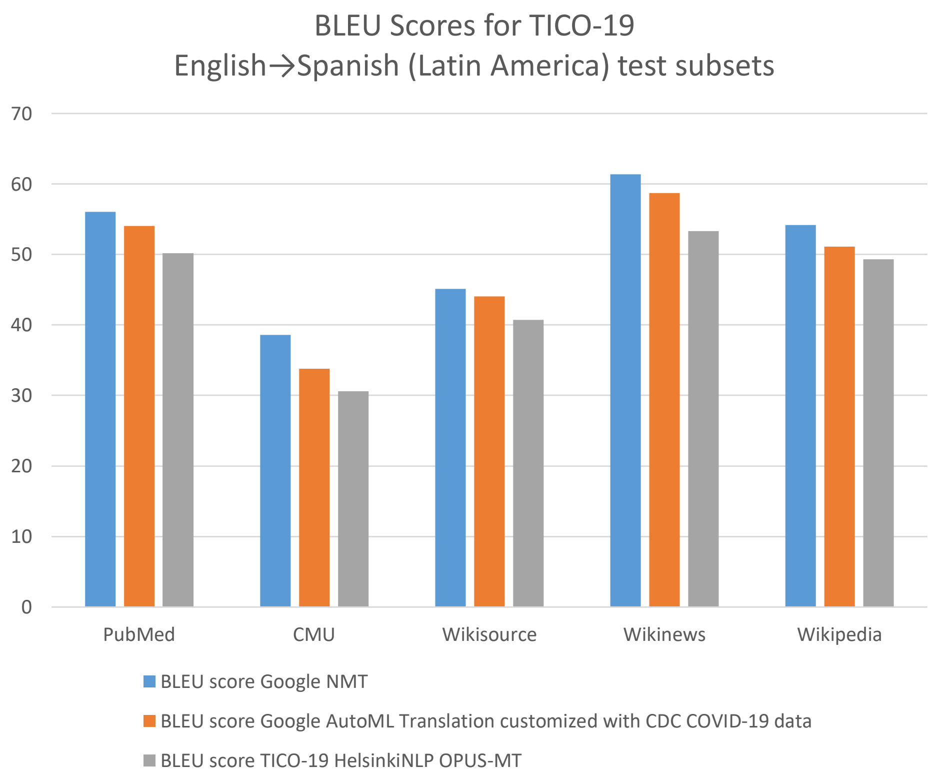
- Content
  - Public domain
  - Disclaimer: Source: CDC; Reference to specific commercial products, manufacturers, companies, or trademarks does not constitute its endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention; 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