A System for Worldwide COVID-19 Information Aggregation

Akiko Aizawa[3], Frederic Bergeron[1], Junjie Chen[5], Fei Cheng[1], Katsuhiko Hayashi[5], Kentaro Inui[4], Hiroyoshi Ito[6], Daisuke Kawahara[7], Masaru Kitsuregawa[3], Hirokazu Kiyomaru[1], Masaki Kobayashi[6], Takashi Kodama[1], Sadao Kurohashi[1], Qianying Liu[1], Masaki Matsubara[6], Yusuke Miyao[5], Atsuyuki Morishim[6], Yugo Murawaki[1], Kazumasa Omura[1], Haiyue Song[1], Eiichiro Sumita[2], Shinji Suzuki[8], Ribeka Tanaka[1], Yu Tanaka[1], Masashi Toyoda[8], Nobuhiro Ueda[1], Honai Ueoka[1], Masao Utiyama[2], Ying Zhong[6] (in alphabetical order)

[1]Kyoto University [2]NICT [3]NII [4]Tohoku University [5]The University of Tokyo [6]The University of Tsukuba
[7]Waseda University [8]Institute of Industrial Science, the University of Tokyo

The System

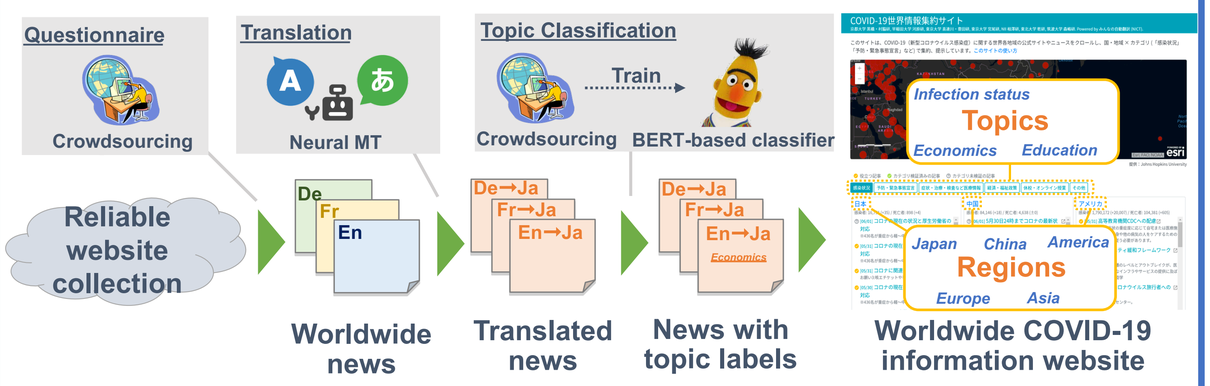
- People pay attention to COVID-19 news of various topics.
- The COVID-19 condition is very different among the countries.
- Getting first-hand information from other countries is essential.

A System for Worldwide COVID-19 Information Aggregation with Various Topics.

http://lotus.kuee.kyoto-u.ac.jp/NLPforCOVID-19

Challenges & Solutions

- The reliability of news sources.
- Translation quality to the local language.
- Topic classification for efficient searching.



- *Robust multilingual reliable website collection via crowdsourcing.*
- *A high-quality machine translation system.*
- *A BERT-based topic classifier*
- *And a user-friendly web interface.*

Reliable Website Collection

To avoid rumors and high-quality, reliable information, we use multiple crowdsourcing services and limit the workers' nationality, assuming that local citizens know the reliable websites in their country.

Website	Country	Primary	Reason	Topics
www.cdc.gov	US	True	The site is a government website, specifically the Center for Disease Control.	infection status prevention and emergency declaration symptoms, medical treatment and tests
www.covid19-yamanaka.com	Japan	False	Shinya Yamanaka is a famous medical researcher and his insights about COVID-19 are reliable.	prevention and emergency declaration
www.internazionale.it	Italy	False	This website collects and translates articles from news agencies and magazines from all over the world. Has up-to-date news, but also long-form analysis articles. Most of my deeper information comes from here.	infection status economics and welfare prevention and emergency declaration school and online classes
covid.saude.gov.br	Brazil	True	This site is the goverment web site.	infection status

Crowdworkers give trusted websites with reasons to choose it and what kind of information they can obtain from it.

- *Totally 908 questionnaire results from 8 countries with totally 550 websites.*
- *Rumors are rampant in this era. The reliable websites dataset can help people to protect themselves.*

Country \ #	Questionnaire	Reliable sites
India	122	67
US	106	77
Italy	104	68
Japan	102	49
Spain	126	90
France	127	71
Germany	106	61
Brazil	115	67
Total	908	550

Statistics of the number of questionnaires and reliable websites collected from each country.

Crawl, Filter and Translation for Information Localization

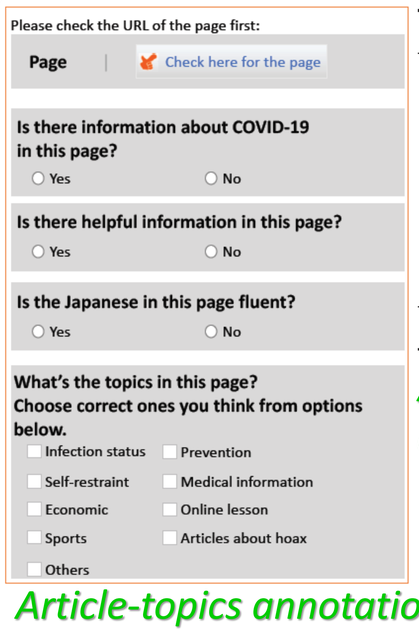
- Crawl COVID-19 related articles from 35 reliable websites.
- Keyword filtering for most related articles.
- Multilingual machine translation to Japanese by Tex-Tra*.

Country	Website	Mentioned times
United States	www.cdc.gov/coronavirus/2019-ncov www.usa.gov/coronavirus www.nytimes.com/news-event/coronavirus	14 6 4
Japan	hazard.yahoo.co.jp/article/20200207 www.mhlw.go.jp/... corona.feedal.com	17 13 6
Italy	www.salute.gov.it/nuovocoronavirus www.salute.gov.it/portale/home.html www.worldometers.info/coronavirus	11 4 3
France	www.gouvernement.fr/info-coronavirus www.who.int/fr/emergencies/diseases/novel-coronavirus-2019 www.lemonde.fr/coronavirus-2019-ncov/	28 7 6
Spain	www.usa.gov/coronavirus www.mscbs.gob.es/profesionales/... covid19.gob.es	9 7 4
Germany	www.rki.de/DE/Home/homepage_node.html www.bundesgesundheitsministerium.de/coronavirus.html interaktiv.morgenpost.de/corona-virus-karte-infektionen-deutschland-weltweit	7 6 5

Top 3 most mentioned reliable websites of each country

*https://mt-auto-minhon-mlt.ucri.jgn-x.jp

Topic-classification



Article-topics annotation through crowdsourcing

Country	Article with topic label
France	72K
America	11K
Japan	9K
China	10K
International	13K
Spain	1K
India	4K
Germany	2K
Total	122K

Article-topics dataset of different languages

Task	Keyword-based model Precision	Recall	F-score	BERT-based model Precision	Recall	F-score
Is about COVID-19	0.36	1.00	0.54	0.82	0.87	**0.84**
Topic: Infection status	0.09	0.53	0.16	0.43	0.81	**0.56**
Topic: Prevention	0.05	0.73	0.10	0.19	0.73	**0.30**
Topic: Medical information	0.17	0.70	0.27	0.27	0.91	**0.41**
Topic: Economic	0.06	0.36	0.10	0.14	0.84	**0.24**
Topic: Education	0.06	1.00	**0.11**	0.05	0.60	0.09
Topic: Art and Sport	0.06	0.41	0.10	0.08	0.94	**0.14**
Topic: Others	0.52	0.07	0.13	0.87	0.79	**0.83**

Topic-classification result of BERT-based model and keyword-based model.

- *Totally 122K article-topics pairs through crowdsourcing.*
- *BERT-based model outperforms the keyword-based model in most tasks.*
- *BERT-based topic-classifier is then applied to generate topic labels for other articles.*

System database

Country	Raw(↑/day)	Translated	With topics
France	774K(8K)	74K	9K
US	69K(730)	15K	2K
Japan	25K(260)	5K	2K
Europe	50K(510)	2K	50
China	38K(400)	3K	342
Int.	45K(470)	3K	263
Korea	16K(170)	260	71
Spain	4K(40)	370	36
India	14K(150)	860	66
Germany	16K(170)	8K	6K
Total	1.05M(11K)	110K	18K

Crawled, translated and labeled articles.

- *Totally 1.05M pages with 110K of them translated and 18K with topic labels. Still growing.*