https://arxiv.org/abs/1910.04519

@CanonMedicalEDI @pschrempf

Language Transfer for Early Warning of Epidemics from Social Media

Mattias Appelgren[1], Patrick Schrempf[1,2], Matúš Falis[1], Satoshi Ikeda[1], Alison Q O'Neil[1,3]
[1]Canon Medical Research Europe Ltd. [2]University of St Andrews [3]University of Edinburgh

AI + HADR Workshop, NeurIPS 2019 (Vancouver, Canada)
13 December 2019

CANON MEDICAL SYSTEMS

University of St Andrews | FOUNDED 1413

THE UNIVERSITY of EDINBURGH

Made For life

Motivation

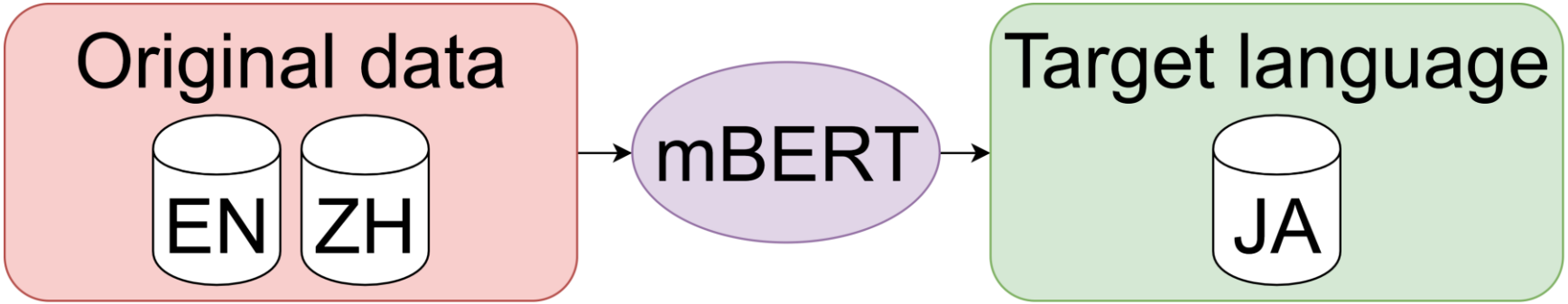
- Diseases spread across populations speaking different languages
- To track red flag medical symptoms in epidemics we need models that work across multiple languages
- It is difficult, time-consuming and expensive to construct training datasets in many languages

Data (MedWeb NTCIR-13 challenge [1])

Dataset	#Pseudo-Tweets	Mean #labels per example	Influenza	Diarrhoea	Hayfever	Cough	Headache	Fever	Runny nose	Cold	#Examples with no labels
Training	1,920	0.997	106	182	163	227	251	345	375	265	530
Test	640	0.933	24	64	46	80	77	93	123	90	195

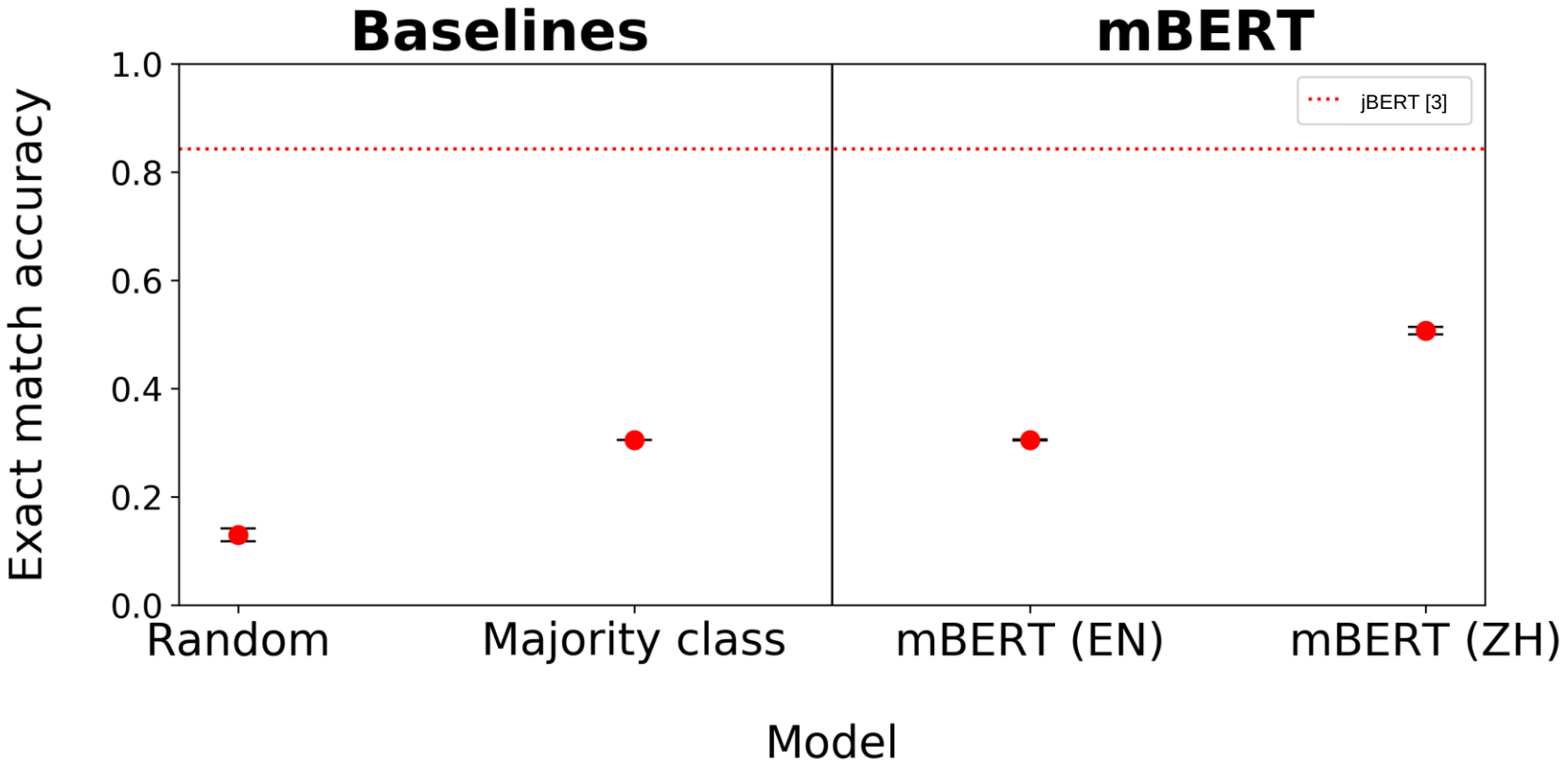
	Pseudo-tweet	Labels
(ja)	風邪を引くと全身がだるくなる。	Cold
(en)	The cold makes my whole body weak.	
(zh)	一感冒就身酸无力。	

Experiment – Zero-shot language transfer

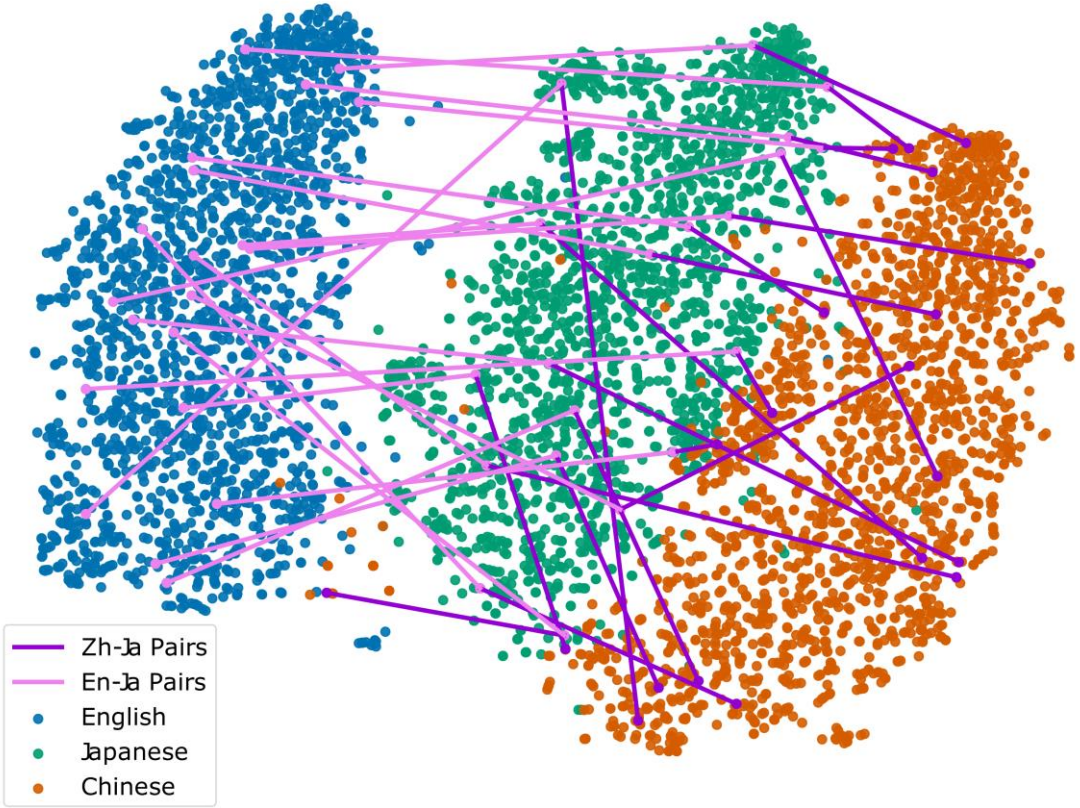


Multilingual BERT (mBERT) [2]

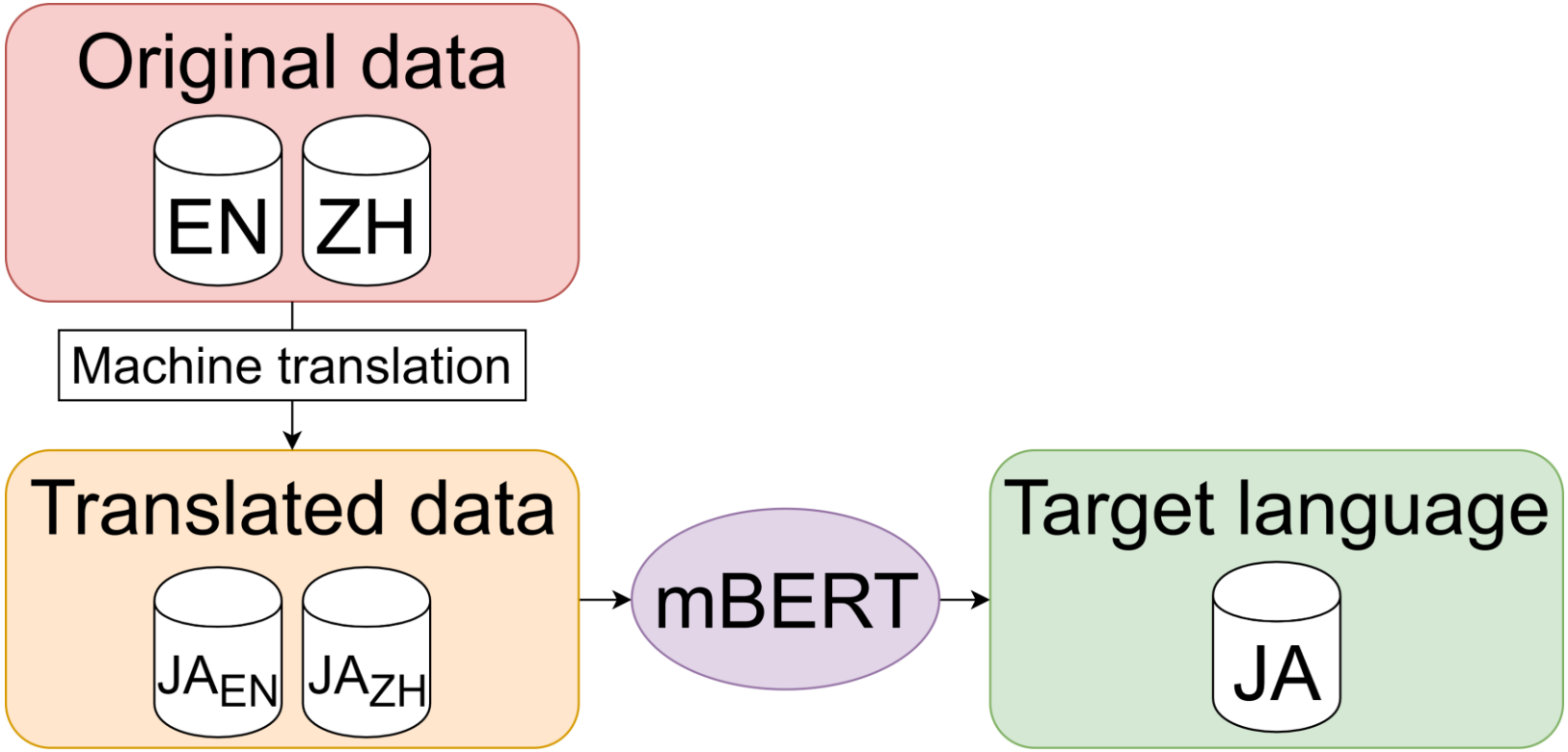
Results – Zero-shot language transfer



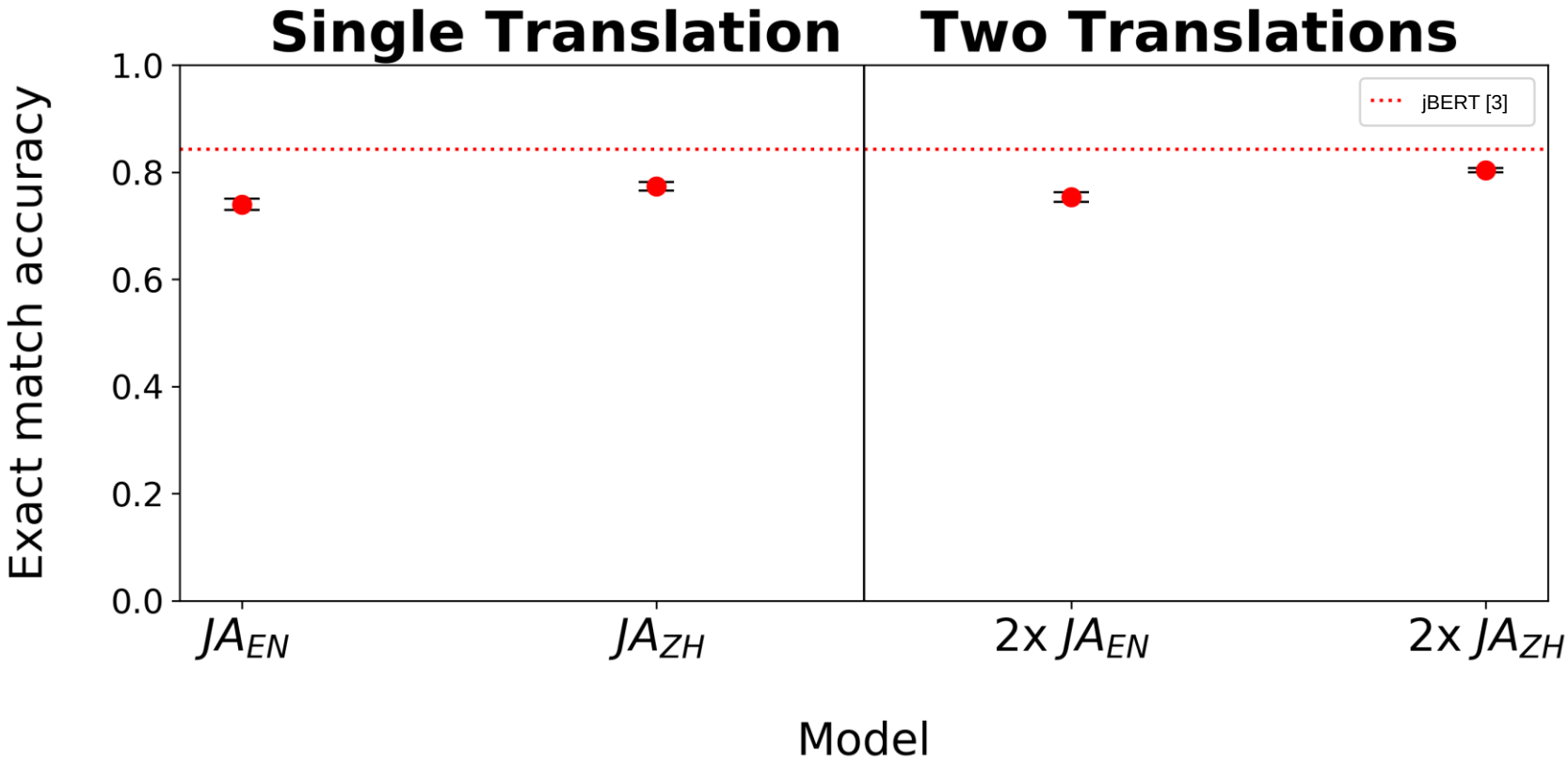
Analysis – Zero-shot language transfer



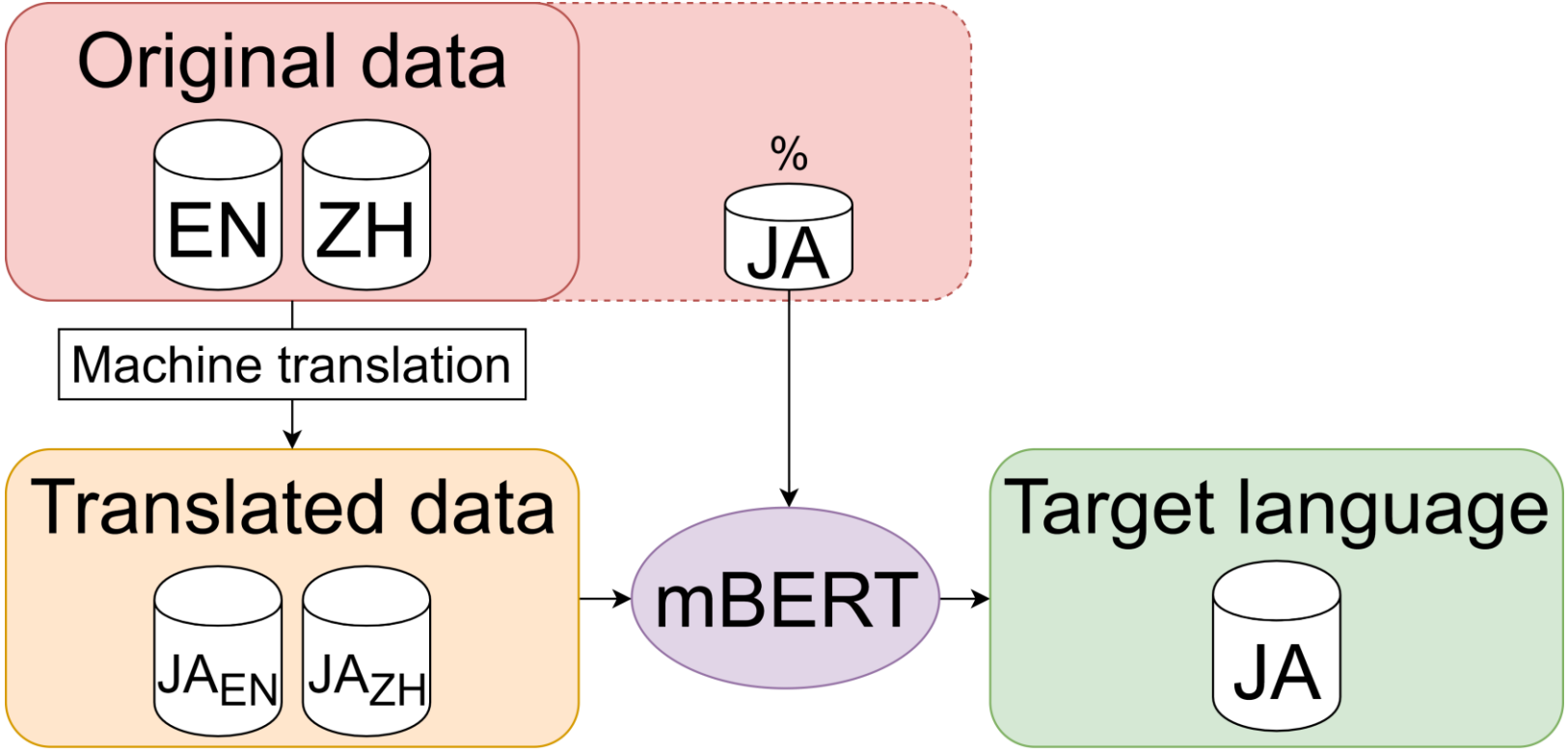
Experiment – Machine translated data



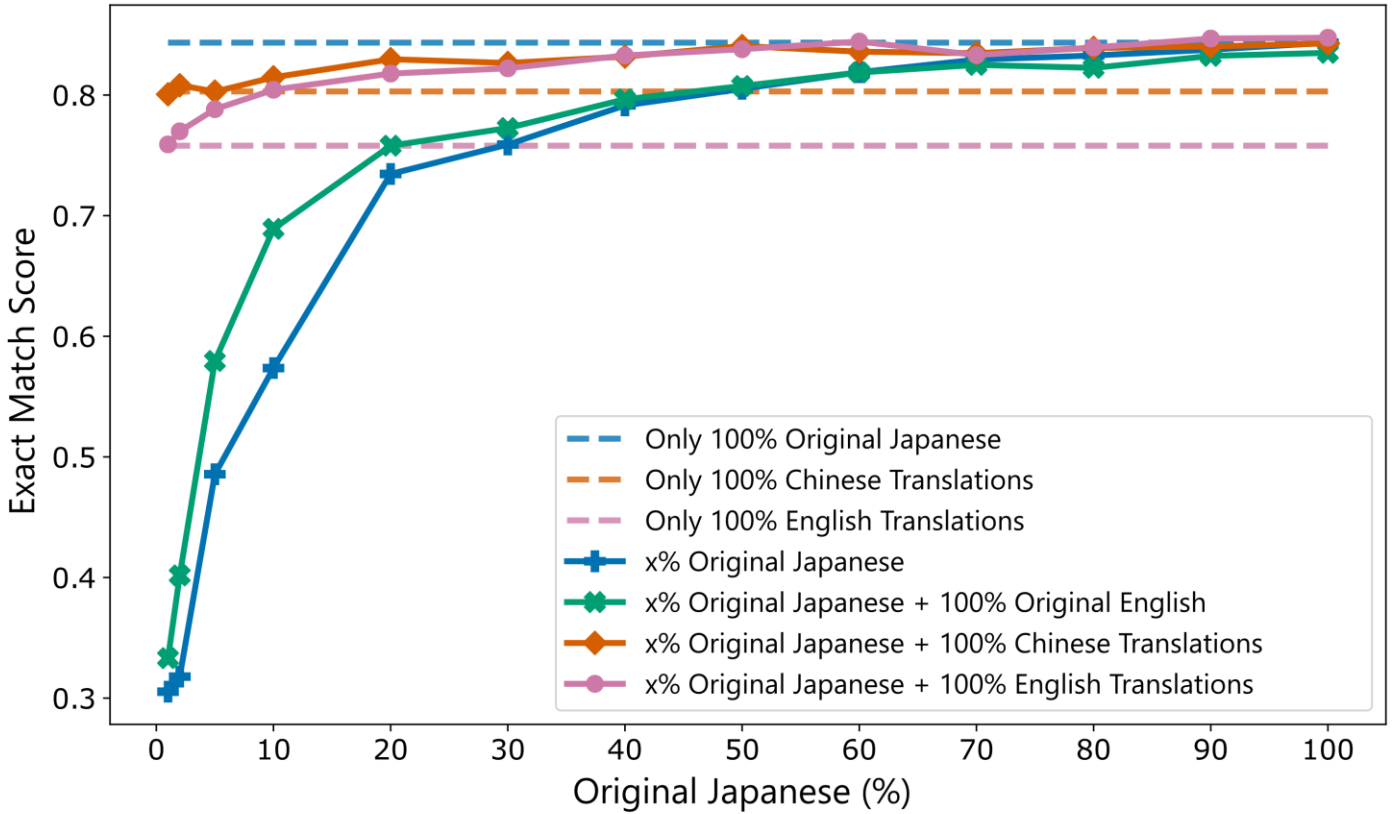
Results – Machine translated data



Experiment – Mixing translated data



Results – Mixing translated data



(Mis)translations

風邪を引くと全身がだるくなる。

The cold makes my whole body weak.

“Cold” has 2 meanings: cold (temperature) and cold (illness).

Both make sense in this context – but one does not match the label!

Conclusions

- Choice of source language impacts the performance, with Chinese-Japanese being a better language pair than English-Japanese

- Training on machine translated data shows promise, especially when used with small amounts of target language data

Mattias Appelgren

Patrick Schrempf

Matúš Falis

Satoshi Ikeda

Alison O'Neil

University of St Andrews | FOUNDED 1413

Canon
CANON MEDICAL SYSTEMS

THE UNIVERSITY of EDINBURGH

Canon

https://research.eu.medical.canon/

References

[1] Aramaki et al., NTCIR-13 MedWeb (Medical Natural Language Processing for Web Document) Task, http://mednlp.jp/medweb/NTCIR-13/

[2] Multilingual BERT (mBERT), https://github.com/google-research/bert/blob/master/multilingual.md

[3] Kikuta et al., BERT pre-trained model trained on Japanese Wikipedia articles.

Made For life

For over 100 years, the Canon Medical Systems `Made for Life' philosophy prevails as our ongoing commitment to humanity - generations of inherited passion creates a legacy of medical innovation and service that continues to evolve as we do. By engaging the brilliant minds of many, we continue to set the benchmark, because we believe quality of life should be a given, not the exception.

University of St Andrews | FOUNDED 1413

Canon
CANON MEDICAL SYSTEMS

THE UNIVERSITY of EDINBURGH