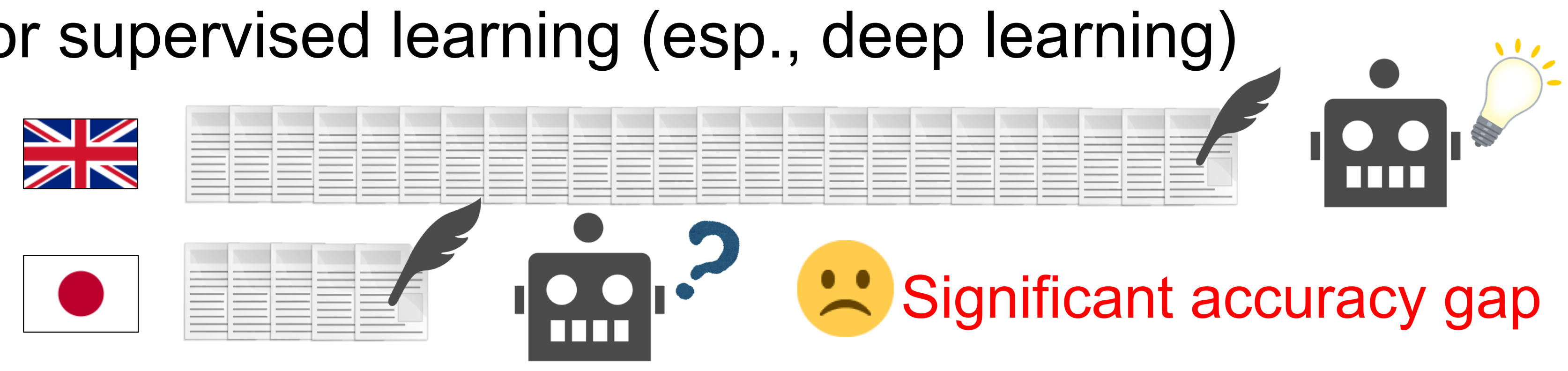
Unsupervised Cross-lingual Word Embeddings Based on Subword Alignment

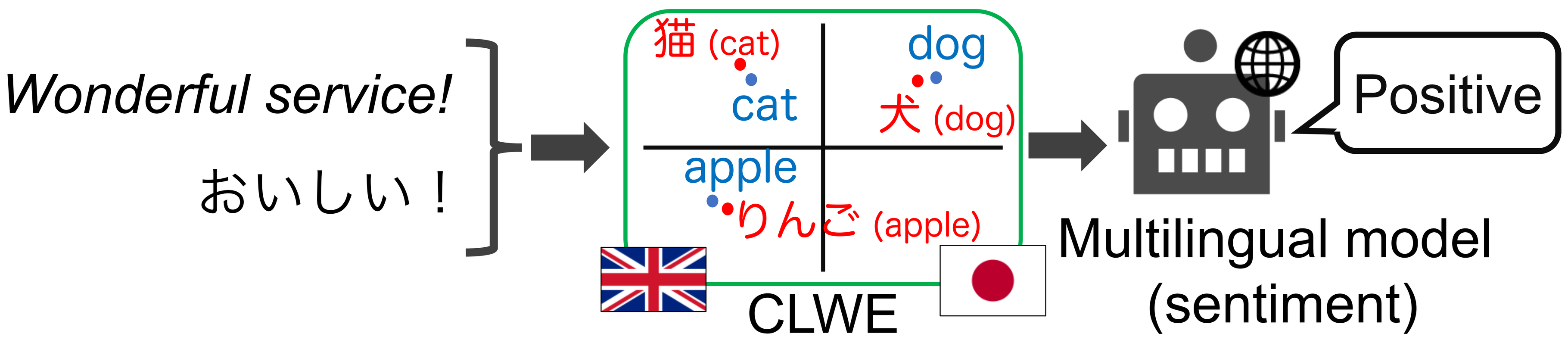
Jin Sakuma (The University of Tokyo), **Naoki Yoshinaga** (Institute of Industrial Science, The University of Tokyo)

Background

Only a few languages have sufficient resources for supervised learning (esp., deep learning)



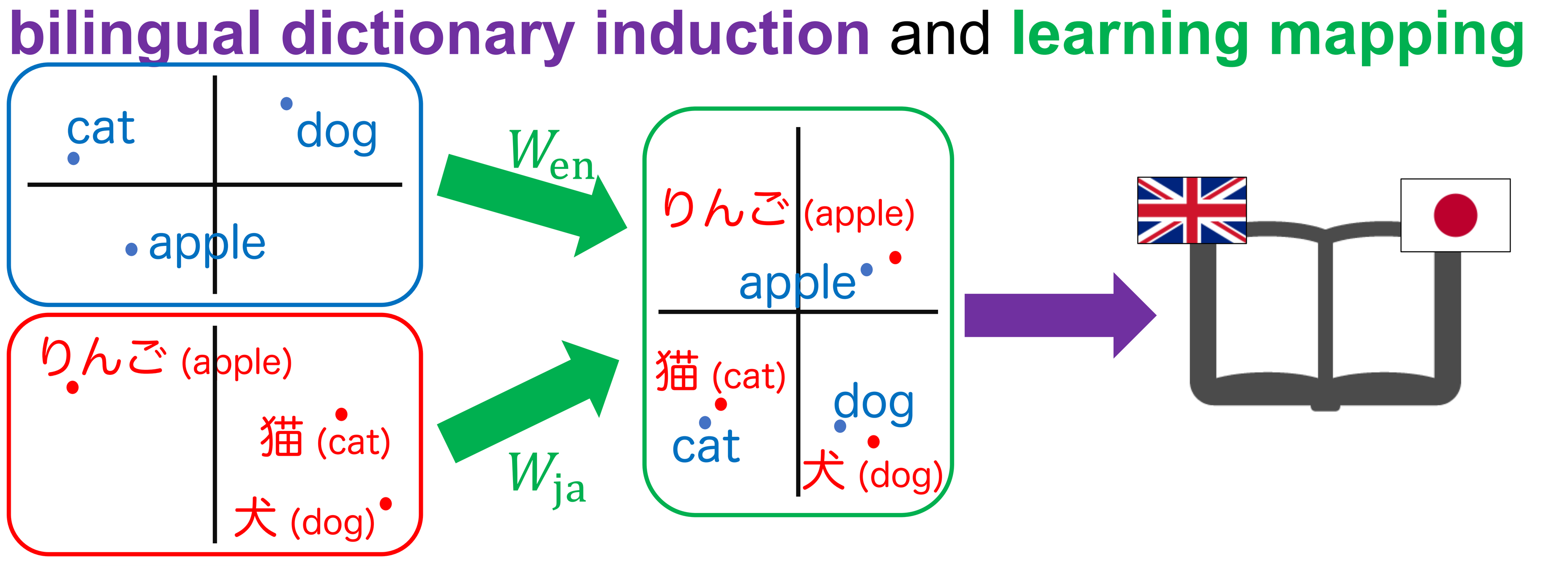
Multilingual models utilize resources across languages by taking cross-lingual word embeddings (**CLWE**) as input



Need high-quality CLWS for resource-rich (English) and resource-poor languages

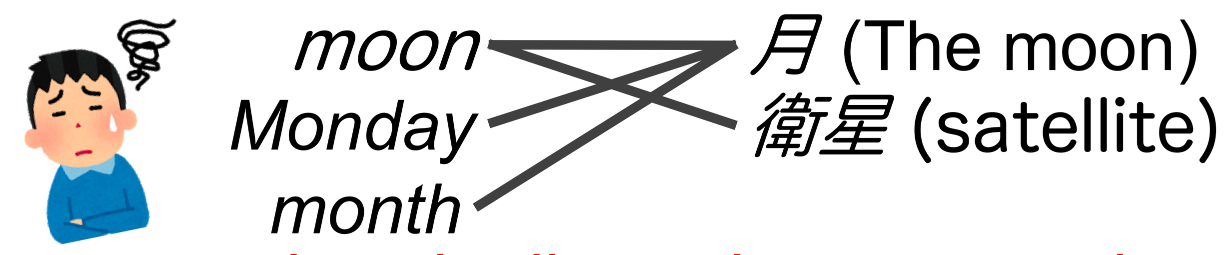
Existing Method [Artetxe+ 2018b]

Learn CLWE in an unsupervised manner by iterating **bilingual dictionary induction** and **learning mapping**



Problem

Ambiguous word correspondence in dictionary



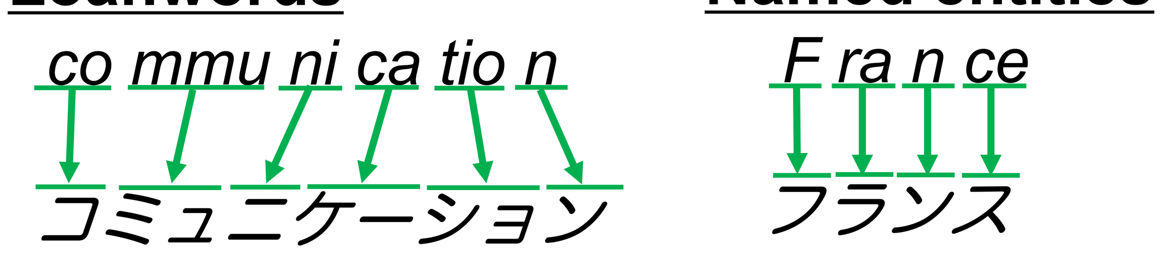
more serious in distant language pairs such as English (resource-rich) and Japanese (resource-poor)

Proposal

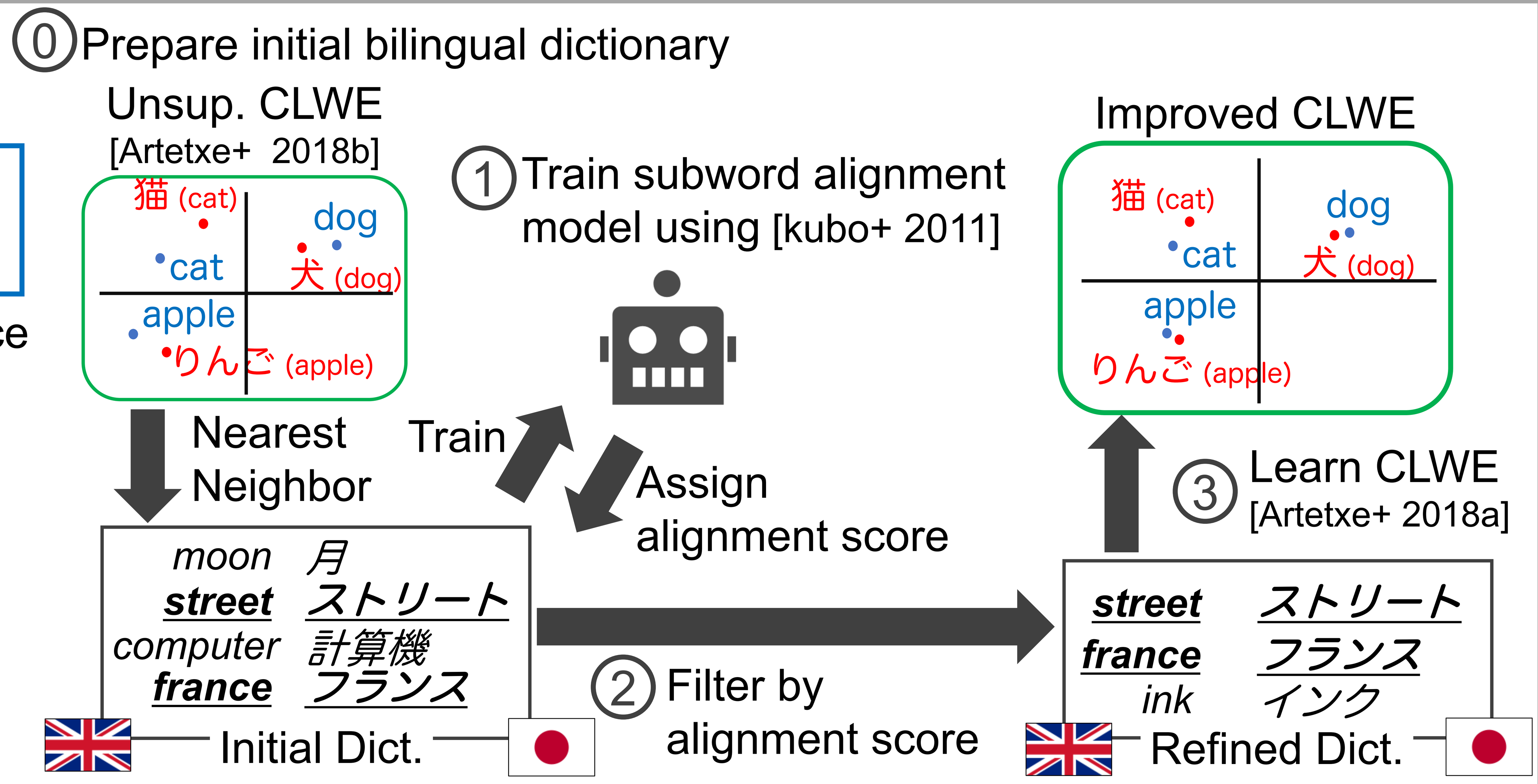
Idea

Exploit unambiguously translatable word pairs (e.g., loanwords, named entities)

- Assumption: words with the surface correspondence are likely to be unambiguously translatable



Filter an initial bilingual dictionary using a subword alignment model trained on it



Evaluation

Task: Bilingual dictionary induction

Predict the word translation from the source (English) to the target language

Settings

- Monolingual word embeddings:
  - fastText pretrained on Wikipedia[1]
  - fastText pretrained on Twitter corpora
- Bilingual dictionary: MUSE bilingual dictionary[2]
- Target languages: Japanese, Finnish (distant), Spanish, Italian (similar)

[Results (Top1 Accuracy)]

Results on Wikipedia embeddings

	Distant lang.		Similar lang.	
	en-ja	en-fi	en-es	en-it
[Artetxe+ 2018b] (unsupervised)	0.457	0.439	0.809	0.771
Proposed	**0.487***	**0.455***	0.809	**0.779**
[Artetxe+ 2018a] (supervised)	0.518	0.437	0.794	0.759
Proposed + MUSE dict. Join the MUSE dictionary with the refined dictionary in **Proposed** method	**0.521**	**0.477***	**0.803**	**0.769**

* statistically significant against baselines ($p < 0.05$)

Our method advanced the state-of-the-art for unsupervised and supervised CLWE

Results on Twitter embeddings

	Distant lang.		Similar lang.	
	en-ja	en-fi	en-es	en-it
[Artetxe+ 2018b]	**0.290***	0.783	0.522	0.439
Proposed	0.281	**0.791***	**0.553***	**0.443***

* statistically significant ($p < 0.05$)

Significant improvements on similar language pairs too

Possibly, Twitter embeddings have more ambiguity in translation

[1]https://fasttext.cc/docs/en/pretrained-vectors.html
[2]https://github.com/facebookresearch/MUSE

Analysis

Top-5 word pairs with highest subword alignment score

English	Finnish
croatia	kroatia
constantin	konstantin
israelis	israelin
india	intia
socrates	sokrates

English	Spanish
international	internacional
secretaries	secretarios
territories	territorios
mercenaries	mercenarios
initial	inicial

Subword alignment model successfully learns how words are imported across languages

Conclusion

Exploit subword alignment for CLWE for refining a bilingual dictionary used to induce CLWE

Improved quality of CLWE in distant language pairs

[Remaining Problem]

The accuracy for distant language pairs are still lower then similar languages

Possibly because:

- Difference in grammar
- Difference in word segmentation

Reference

Artetxe+ 2018a, Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformation, In AAAI 2018
Artetxe+ 2018b, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In ACL 2018
Kubo+ 2011, Unconstrained many to many alignment for automatic pronunciation annotation. In APSIPA 2011