熊本県立大学
Prefectural University of Kumamoto

Improving Polish to English Neural Machine Translation with Transfer Learning: Effects of Data Volume and Language Similarity

Multi3Generation Workshop 2023

Juuso Eronen, Michal Ptaszynski, Karol Nowakowski, Cheng Lin Chia, Fumito Masui

Background

熊本県立大学
Prefectural University of Kumamoto

- Most languages have insufficient resources for model training

- A handful of languages (especially English) are dominating NLP field

Cross-Lingual Transfer Learning

熊本県立大学
Prefectural University of Kumamoto

- Leverage knowledge from other languages
- From higher resource language
  - E.g. English, German
- Multilingual transformer models

Cross-Lingual Transfer Learning

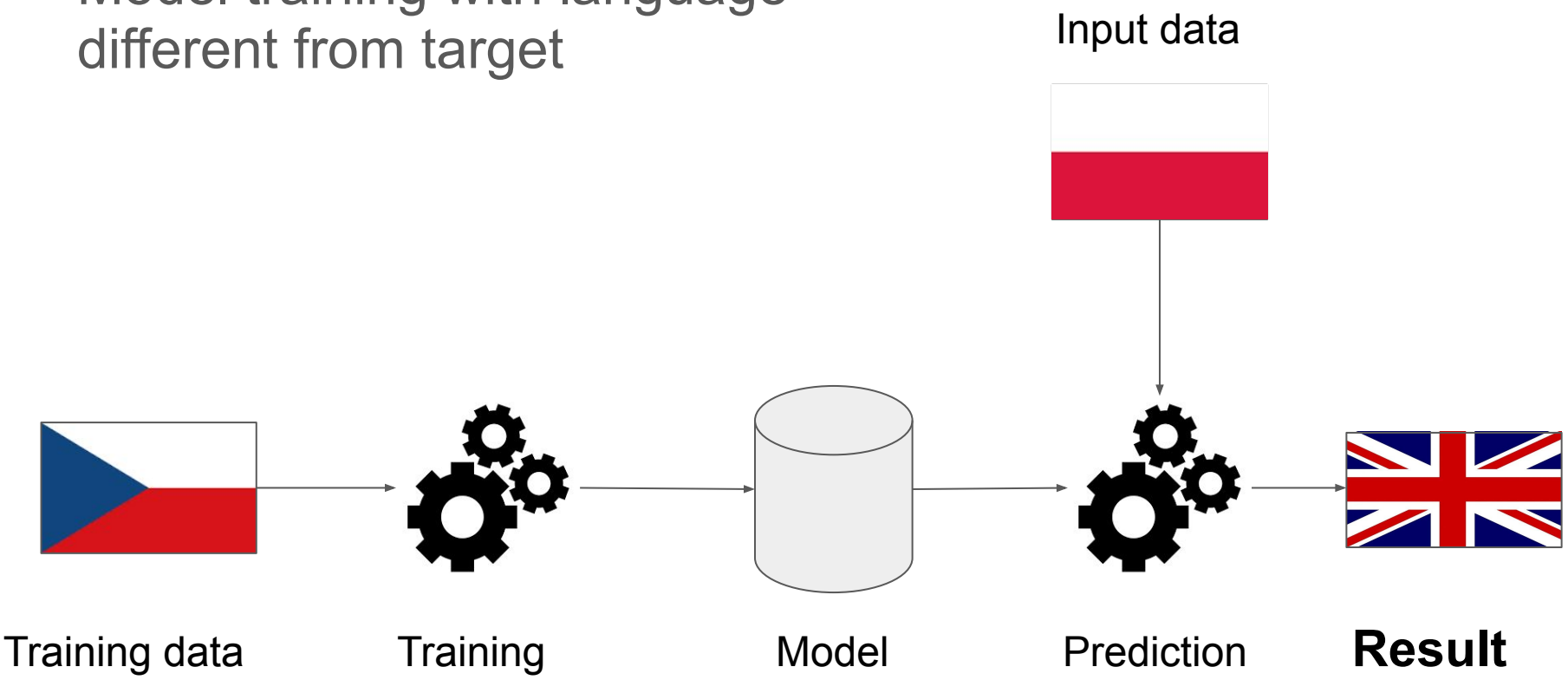
熊本県立大学
Prefectural University of Kumamoto

- Few-shot learning
  - Only a handful of transfer target language samples
- Zero-shot learning
  - Does not require any labeled data in the transfer target language for training

Cross-Lingual Transfer Learning

熊本県立大学
Prefectural University of Kumamoto

Model training with language different from target



Research goal

熊本県立大学
Prefectural University of Kumamoto

Analyze impact of data volume on transfer learning in a machine translation task

Examine the influence of language relatedness on transfer learning in machine translation

Previous research

熊本県立大学
Prefectural University of Kumamoto

- Using large amounts of data from high-resource languages improves performance on low-resource languages [1] [2]
- The size of the used source corpus can be more important than the relatedness of the source and target languages [3]
- Transferring from multiple languages increases performance [4]

[1] Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Conference on Empirical Methods in Natural Language Processing.

[2] Koehn, Philipp and Rebecca Knowles. 2017. Six chal-lenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, August. Association for Computational Linguistics.

[3] Kocmi, Tom and Ondˇrej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 244–252, Brussels, Belgium, October. Association for Computational Linguistics

[4] Chen, Xilun, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3098–3112, Florence, Italy, July. Association for Computational Linguistics.

Previous research

熊本県立大学
Prefectural University of Kumamoto

- Transferring between more similar languages could yield higher scores [5] [6]
- Language similarity correlates with cross-lingual transfer efficacy [7]

[5] Anne Lauscher, et al. “From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers.” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499, Online, November 2020. Association for Computational Linguistics

[6] Gaikwad, Saurabh, et al. "Cross-lingual offensive language identification for low resource languages: The case of Marathi." arXiv preprint arXiv:2109.03552 (2021).

[7] Eronen, Juuso, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. Information Processing & Management, 60(3):103250.

Method

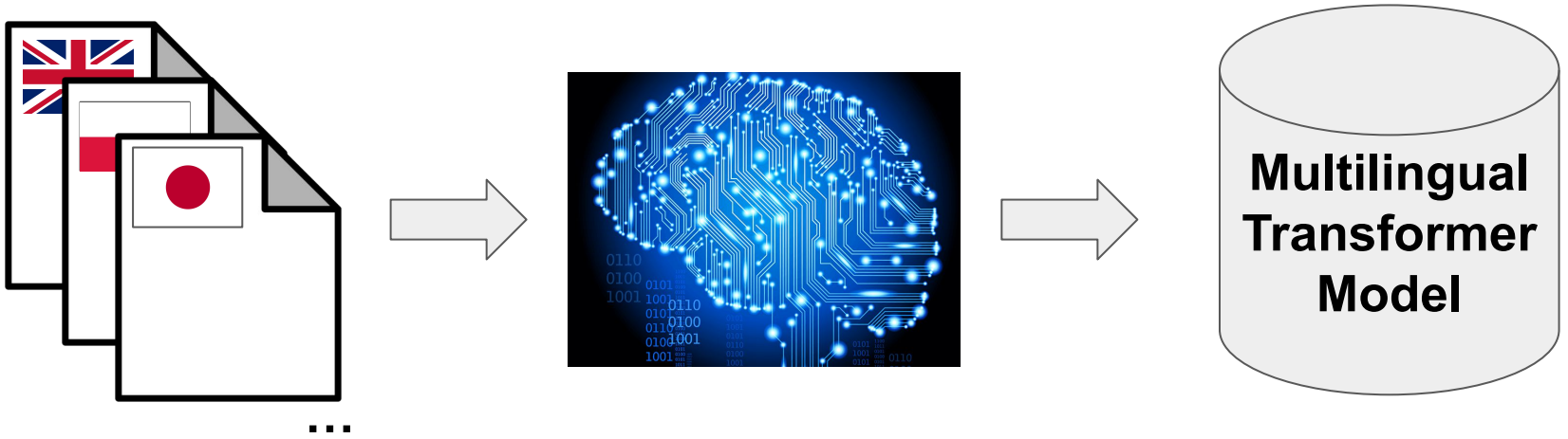
熊本県立大学
Prefectural University of Kumamoto

- mBART
- OPUS-100
- Polish to English translation task
  - **Different transfer languages, different shot levels**
- Evaluation: BLEU, METEOR

mBART

熊本県立大学
Prefectural University of Kumamoto

- Multilingual sequence-to-sequence model Based on BERT
- Developed by Facebook AI Research (FAIR)
- Achieved state-of-the-art performance on various machine translation benchmarks



https://publicdomainvectors.org/

https://www.telusinternational.com/articles/artificial-intelligence-ai-will-affect-customer-service

Dataset

熊本県立大学
Prefectural University of Kumamoto

- OPUS-100
  - English-centric parallel corpus
  - Common benchmark dataset for multilingual machine translation

- Transfer source:
  - Czech-English (100k samples)
  - Russian-English (100k samples)
  - German-English (100k samples)

- Transfer target and evaluation: Polish-English

Transfer learning configurations

熊本県立大学
Prefectural University of Kumamoto

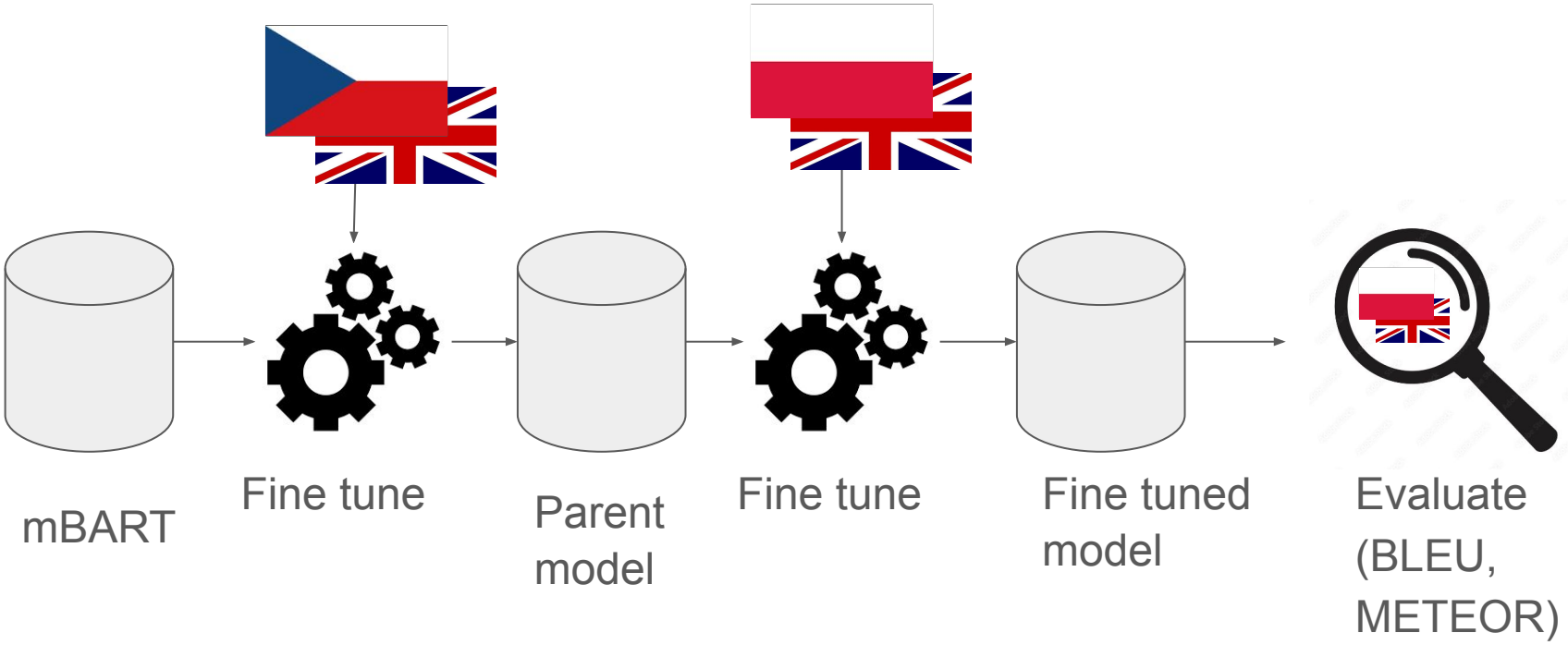
- Vanilla mBART

- “High-resource” parent models:
  - Czech-English (100k samples)
  - Russian-English (100k samples)
  - Slavic-English (200k samples, Czech + Russian)
  - German-English (100k samples)

- Each model fine-tuned with **0, 10, 100, 1k and 10k** of Polish-English data

Transfer learning configurations

熊本県立大学
Prefectural University of Kumamoto



Evaluation

熊本県立大学
Prefectural University of Kumamoto

- BLEU
  - Common metric
- METEOR
  - More advanced, shown to correlate well with human judgments

Results

熊本県立大学
Prefectural University of Kumamoto

Translation source: Polish	0 shot		10 shot		100 shot		1k shot		10k shot	
Transfer source:	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
N/A	--	--	0.45	0.05	0.01	0.01	10.43	0.33	15.42	0.36
Czech	11.61	0.35	14.3	0.41	13.41	0.37	14.35	0.42	17.17	0.41
Russian	0.42	0.11	3.16	0.26	4.86	0.31	16.44	0.41	19.42	0.44
Slavic	8.33	0.27	11.94	0.36	10.87	0.35	16.44	0.41	18.18	0.43
German	0.12	0.05	0.56	0.07	3.72	0.29	16.82	0.42	19.35	0.44

Results

熊本県立大学
Prefectural University of Kumamoto

Translation source: Polish	0 shot		10 shot		100 shot		1k shot		10k shot	
Transfer source:	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
N/A	--	--	0.45	0.05	0.01	0.01	10.43	0.33	15.42	0.36
Czech	11.61	0.35	14.3	0.41	13.41	0.37	14.35	0.42	17.17	0.41
Russian	0.42	0.11	3.16	0.26	4.86	0.31	16.44	0.41	19.42	0.44
Slavic	8.33	0.27	11.94	0.36	10.87	0.35	16.44	0.41	18.18	0.43
German	0.12	0.05	0.56	0.07	3.72	0.29	16.82	0.42	19.35	0.44

- Despite including the same Czech data and additional Russian data, the Slavic model shows performs worse than the Czech model

Results

熊本県立大学
Prefectural University of Kumamoto

Translation source: Polish	0 shot		10 shot		100 shot		1k shot		10k shot	
Transfer source:	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
N/A	--	--	0.45	0.05	0.01	0.01	10.43	0.33	15.42	0.36
Czech	11.61	0.35	14.3	0.41	13.41	0.37	14.35	0.42	17.17	0.41
Russian	0.42	0.11	3.16	0.26	4.86	0.31	16.44	0.41	19.42	0.44
Slavic	8.33	0.27	11.94	0.36	10.87	0.35	16.44	0.41	18.18	0.43
German	0.12	0.05	0.56	0.07	3.72	0.29	16.82	0.42	19.35	0.44

- Despite including the same Czech data and additional Russian data, the Slavic model shows performs worse than the Czech model

Results

熊本県立大学
Prefectural University of Kumamoto

Translation source: Polish	0 shot		10 shot		100 shot		1k shot		10k shot	
Transfer source:	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
N/A	--	--	0.45	0.05	0.01	0.01	10.43	0.33	15.42	0.36
Czech	11.61	0.35	14.3	0.41	13.41	0.37	14.35	0.42	17.17	0.41
Russian	0.42	0.11	3.16	0.26	4.86	0.31	16.44	0.41	19.42	0.44
Slavic	8.33	0.27	11.94	0.36	10.87	0.35	16.44	0.41	18.18	0.43
German	0.12	0.05	0.56	0.07	3.72	0.29	16.82	0.42	19.35	0.44

- The performance of both Russian and German are also rising and catching up to Czech and Slavic

Results

熊本県立大学
Prefectural University of Kumamoto

Translation source: Polish	0 shot		10 shot		100 shot		1k shot		10k shot	
Transfer source:	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
N/A	--	--	0.45	0.05	0.01	0.01	10.43	0.33	15.42	0.36
Czech	11.61	0.35	14.3	0.41	13.41	0.37	14.35	0.42	17.17	0.41
Russian	0.42	0.11	3.16	0.26	4.86	0.31	16.44	0.41	19.42	0.44
Slavic	8.33	0.27	11.94	0.36	10.87	0.35	16.44	0.41	18.18	0.43
German	0.12	0.05	0.56	0.07	3.72	0.29	16.82	0.42	19.35	0.44

- Equal performance between transfer languages from 1k shot
- Using only Polish without any transfer learning starts to produce comparable results from 10k shot

Results

熊本県立大学
Prefectural University of Kumamoto

Translation source: Polish	0 shot		10 shot		100 shot		1k shot		10k shot	
Transfer source:	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
N/A	--	--	0.45	0.05	0.01	0.01	10.43	0.33	15.42	0.36
Czech	11.61	0.35	14.3	0.41	13.41	0.37	14.35	0.42	17.17	0.41
Russian	0.42	0.11	3.16	0.26	4.86	0.31	16.44	0.41	19.42	0.44
Slavic	8.33	0.27	11.94	0.36	10.87	0.35	16.44	0.41	18.18	0.43
German	0.12	0.05	0.56	0.07	3.72	0.29	16.82	0.42	19.35	0.44

- Zero-shot with Czech outperforms 1k Polish

Effect of data volume

熊本県立大学
Prefectural University of Kumamoto

- Increasing the amount of transfer target language data (Polish) improves performance

- Surprisingly, increasing the amount of transfer source language data did not increase the performance.
  - The slavic model has 2x more data but performs worse than just Czech

Effect of language similarity

熊本県立大学
Prefectural University of Kumamoto

- Importance of similarity in zero- and few-shot settings

  → Low-resource scenarios
- Seems to diminish as the amount of transfer target language data increases
  - 1k, 10k samples: performance almost equal across languages

- Comparably high zero-shot results when the transfer source language is of high similarity (Czech) with Translation source language (Polish)

Impact

熊本県立大学
Prefectural University of Kumamoto

- Transfer learning can provide a temporary solution to the lack of data to enable service

- **Can be enhanced with the use of similar languages**

Conclusions

熊本県立大学
Prefectural University of Kumamoto

- Additional transfer data does not necessarily result in higher performance

- **Importance of language similarity in low resource scenarios**

Limitations

熊本県立大学
Prefectural University of Kumamoto

- Only Polish-English task
- Limited amount of languages for transfer learning
- Use of only a single corpus

Future Research

熊本県立大学
Prefectural University of Kumamoto

- Using other language pairs
- Confirmation with other datasets and NLP tasks
- Use of multiple transfer languages

熊本県立大学
Prefectural University of Kumamoto

Thank you for listening