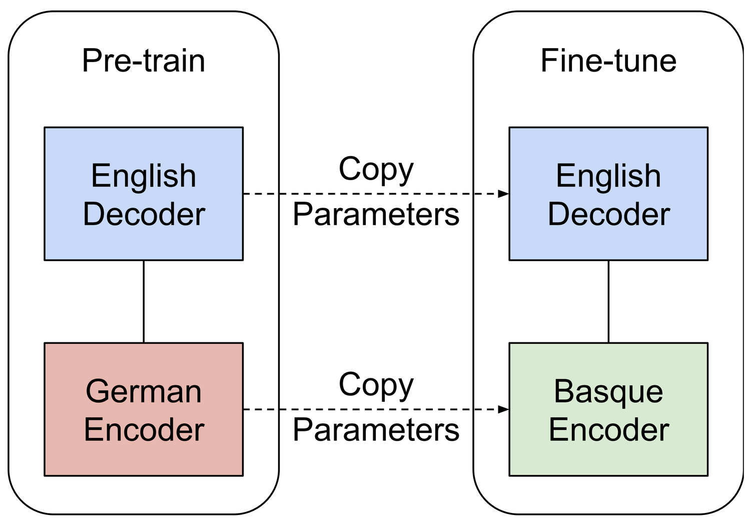
Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies

Yunsu Kim, **Yingbo Gao**, Hermann Ney

Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany

Cross-lingual Transfer Learning for NMT

High-resource language pair (*parent*) ⟶ Low-resource language pair (*child*)



How to mitigate language differences?

So far: shared vocabulary (joint BPE)

- Difficult to adapt to a new language (must be re-trained)

This work: **separate** vocabulary

- Cross-lingual word embedding
- Artificial noises
- Synthetic data from *parent*

⇒ Up to +5.1 Bleu[%] better transfer

Main Results

Datasets (xx⟶English)

xx	Family	#pairs	
German (de)	Germanic	10M	⟶ *parent*
Basque (eu)	Isolate	6K	*child* (distant from German)
Slovenian (sl)	Slavic	17K	
Belarusian (be)		5K	
Azerbaijani (az)	Turkic	6K	
Turkish (tr)		10K	

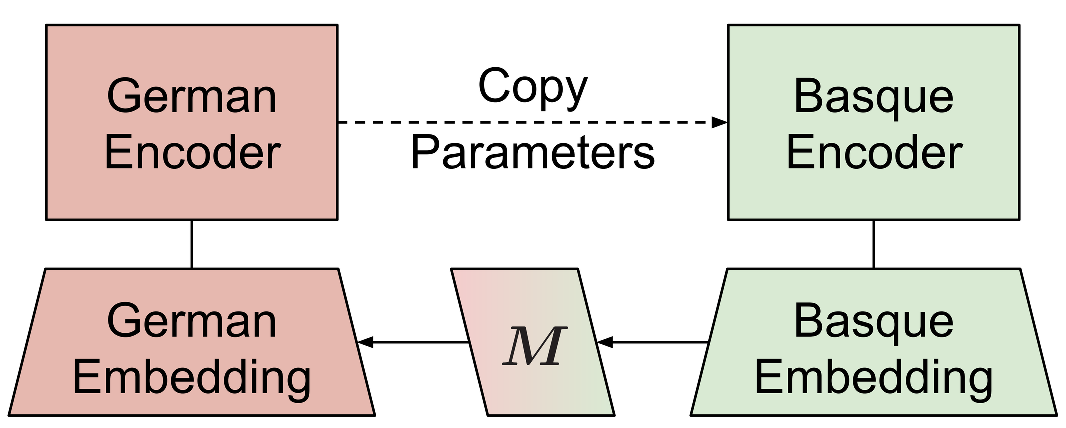
Results (Bleu [%])

System	eu-en	sl-en	be-en	az-en	tr-en
Transformer baseline (*child* only)	1.7	10.1	3.2	3.1	0.8
Multilingual (*parent* + *child*)	5.1	16.7	4.2	4.5	8.7
Transfer	4.9	19.2	8.9	5.3	7.4
+ Cross-lingual word embedding	7.4	20.6	12.2	7.4	9.4
+ Artificial noises	8.2	21.3	12.8	8.1	10.1
+ Synthetic data from *parent*	**9.7**	**22.1**	**14.0**	**9.0**	**11.3**

- Naive training of Transformer fails for low-resource language pairs
  - ▸ Multilingual/Transfer learning is helpful yet still limited
- Incremental improvements with our proposed methods
  - ▸ Up to +5.1 Bleu[%] over plain transfer
  - ▸ Up to +9.8 Bleu[%] over multilingual systems

Cross-lingual Word Embedding

Problem: Vocabulary mismatch between *parent* / *child* languages
Solution: Shared word embedding space



1. $\boldsymbol{E}_{\text{child}}$ = monolingual skip-gram, $\boldsymbol{E}_{\text{parent}}$ = pre-trained *parent* NMT
2. $\boldsymbol{M}$ = linear mapping $\boldsymbol{E}_{\text{child}} \rightarrow \boldsymbol{E}_{\text{parent}}$ (e.g. Muse)

$$M_i = \underset{M'}{\operatorname{argmin}} \sum_{(w,w')\in D_i} \|M' E_{\text{child}}(w) - E_{\text{parent}}(w')\|_2$$

D_0 = seed dictionary (obtained unsupervisedly)

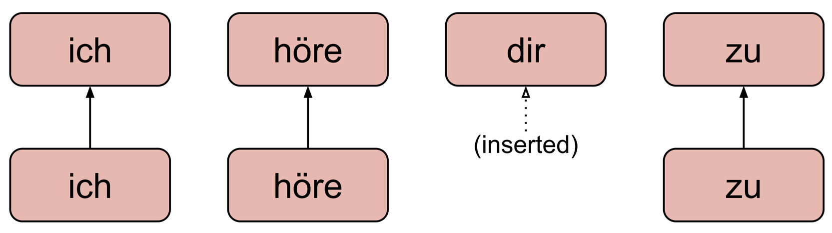
$$D_i = \{(w, w') \mid w' = \underset{w^*}{\operatorname{argmin}} \|M_{i-1} E_{\text{child}}(w) - E_{\text{parent}}(w^*)\|_2\}$$

3. *parent* model + mapped *child* embedding ($\boldsymbol{M}\boldsymbol{E}_{\text{child}}$) ⇒ fine-tuning
   - ▸ *child* vocabulary in *parent* embedding space
   - ▸ Applicable to languages with different alphabets

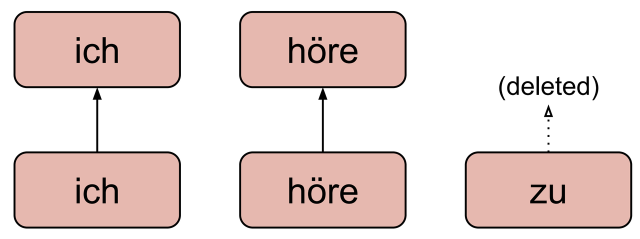
Artificial Noises

Problem: Word order difference between *parent* / *child* languages
Solution: Pre-train syntax-agnostic *parent* encoder with noisy input

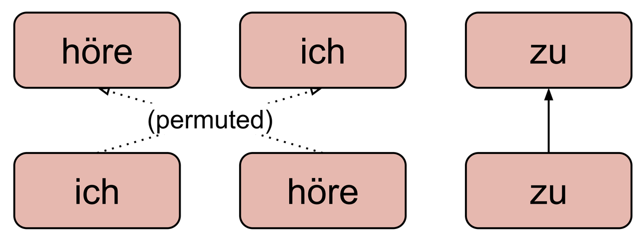
Insertion (10%)



Deletion (10%)



Permutation (100%, max. dist. 3)

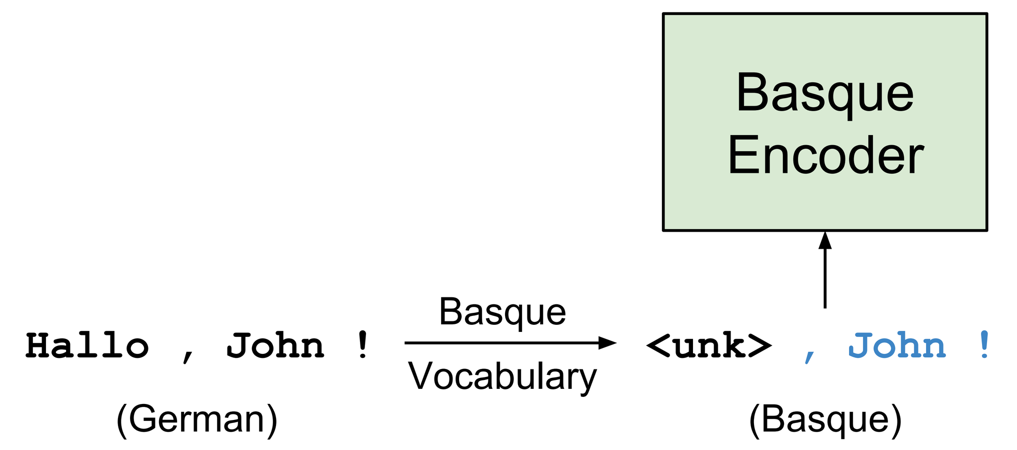


- Randomized for each batch
  - ▸ *parent* learns various syntax
  - ▸ General monotonicity is retained
- Easier to adapt to a new language

Synthetic Data from Parent

Problem: Back-translation does not work (poor English⟶xx model)
Solution: Reuse *parent* training data and adjust to *child* vocabulary

▸ Keep shared tokens and map the rest to <unk>



- Shared tokens: English named entities, digits, punctuations, etc.
  - ▸ Synergy with cross-lingual word embedding
  - ▸ Basic sentence structure is retained
- Prevent abrupt changes of training data in fine-tuning
- Avoid overfitting to small *child* data

Ablation Studies

Vocabulary size (xx-en)

BPE merges (xx)	Bleu [%] sl-en	Bleu [%] be-en
10k	**21.0**	11.2
20k	20.6	**12.2**
50k	20.2	10.9
70k	20.0	10.9

▸ Transfer on a small vocabulary

Pre-trained embeddings (az-en)

Embedding	Bleu [%]
None	5.3
Monolingual	6.3
Cross-lingual (az-en)	7.1
Cross-lingual (az-de)	**7.4**

▸ Cross-lingually map *parent*-child

Synthetic data generation (eu-en)

Synthetic data	Bleu [%]
None	8.2
Back-translation	8.3
Empty source	8.2
Copied target	8.9
parent data	**9.7**

▸ Reuse *parent* data

Freeze decoder params (sl-en)

Frozen parameters	Bleu [%]
None	21.0
Target embedding	21.4
+ Target self-att.	**22.1**
+ Encoder-decoder att.	21.8
+ Feedforward sublayer	21.3
+ Output layer	21.9

▸ Freeze target-specific parameters

Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1, project "CoreTec"). The GPU cluster used for the experiments was partially funded by DFG Grant INST 222/1168-1. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

Paper & Code

Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany   Email: <surname>@i6.informatik.rwth-aachen.de   WWW: http://www.hltpr.rwth-aachen.de