- Pretrain with any high-resource language pair even **unrelated**.
- **Iterate** NMT backtranslation.

CUNI Submission for Low-Resource Languages in WMT News 2019

Tom Kocmi and Ondřej Bojar
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

Transfer Learning

Transfer learning means to:

- Train any high-resource language pair
- Continue training ("fine-tune") on the language pair of interest

We follow transfer learning by two rounds of iterated backtranslation.

Backtranslation with Cleaning

After backtranslation, we:

- Remove sentence pairs with repetitive patterns.
- Remove sentences automatically identified as a different language.

Synthetic from Scratch

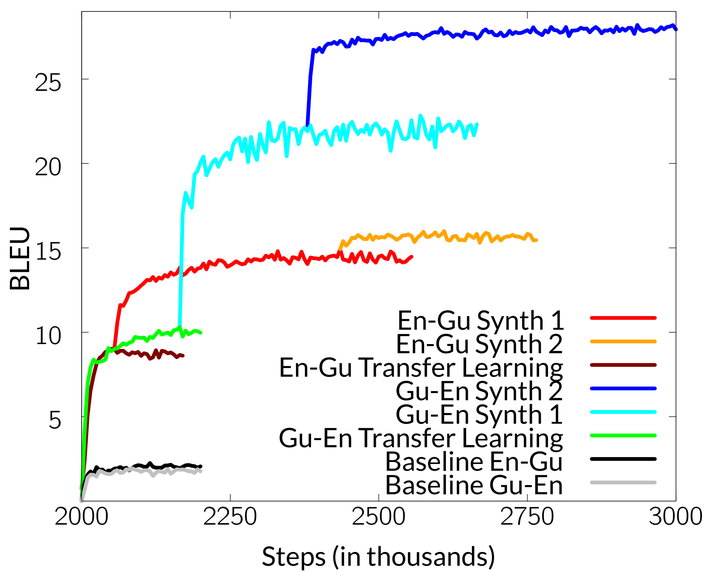
Train second round of backtranslation from scratch instead of fine-tuning the previous model.

Training dataset	cased	uncased
Auth (baseline)	1.8	2.2
Synth only	**16.9**	**18.7**
Auth:Synth 20:1	16.8	18.4
Auth:Synth 40:1	16.3	17.8
Auth:Synth 80:1	15.2	16.8
Submitted model	16.2	17.9

Final Behavior

The English-Gujarati has English-Czech parent model.
The English-Kazakh has English-Russian parent model.

Training dataset	EN→GU	GU→EN	EN→KK	KK→EN
Authentic (baseline)	2.0	1.8	0.5	4.2
Parent dataset	0.7	0.1	0.7	0.6
Transfer learning	① 9.1	9.2	6.2	① 14.4
Synth generated by ①	-	② 14.2	② 8.3	-
Synth generated by ②	③ 13.4	-	-	17.3
Synth generated by ③	-	④ 16.2	-	-
Synth generated by ④	13.7	-	-	-
Averaging + beam 8	14.3	17.4	8.7	18.5

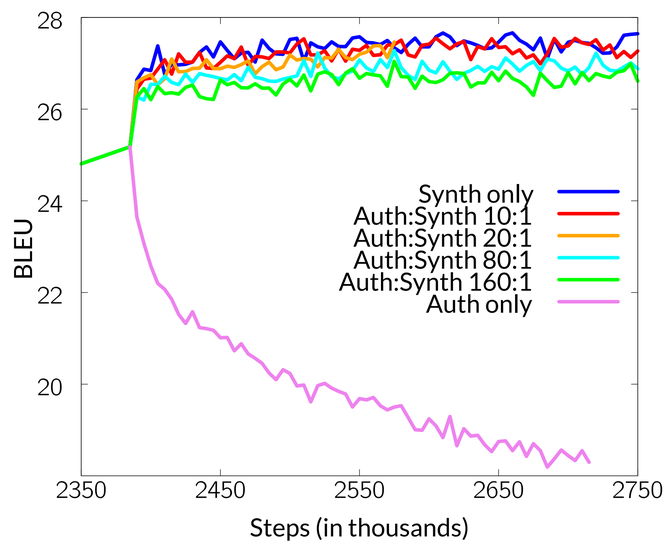


Training data

Language pair	Parallel	Monolingual
Russian-English	13.5M	-
Czech-English	57.4M	-
Kazakh-English	220k	13.2M–15.4M
Gujarati-English	173k	4.2M–15.4M

Mixing Synth-Auth at Various Ratios

For low-resource languages, do not mix the authentic and synthetic data. Training on synthetic is the best.



The full paper

This study was supported in parts by the grants SVV 260 453 of the Charles University, 18-24210S of the Czech Science Foundation and 825303 (Bergamot) of the European Union. This work has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071).