CUNI System for the WMT19 Robustness Task

Jindřich Helcl, Jindřich Libovický, Martin Popel

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University

The Robustness Task

- Based on MTNT dataset: translated Reddit posts
- Informal language, smileys, typos

Our Method

- Transformer BIG model, with large batch size of 16.8k subwords
- All WMT15 parallel data +



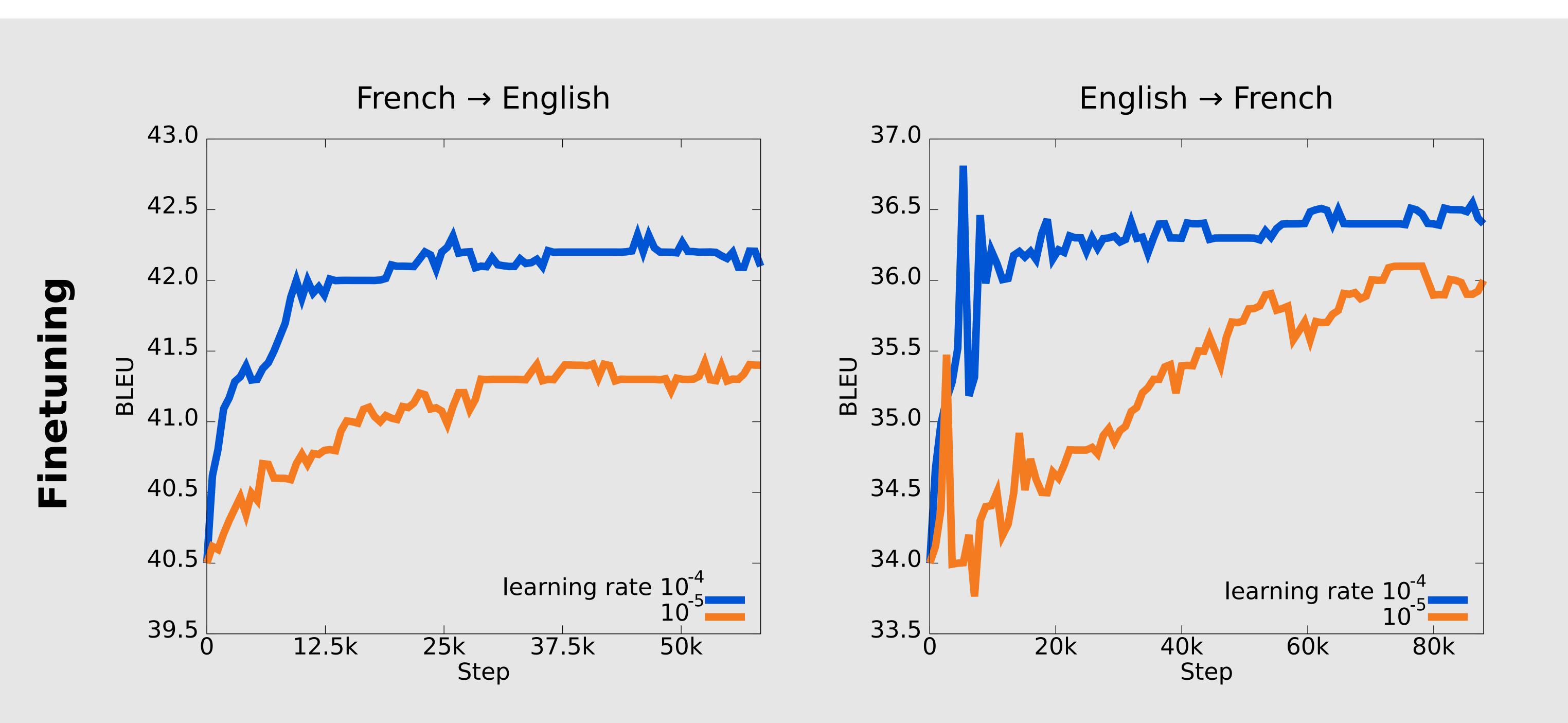
• Only 36k sentence pairs for training

en: Added a small AIC 😘 😘 Sounds perfect 😂 J'ai ajouté un petit AIC 😘 😘 😪 Ça sonne bien 😂 fr:

「_(ツ)_/ en: fr: __(ツ)_/_

French WMT15 News Crawl monolingual

- Three iterations of back-translation
- Finetuning on MTNT data with SGD and small learning rate



Results BLEU scores for English ↔ French Translation

	French → English				English → French					en-fr	fr-en
	WMT14	WMT15	MTNT	blind	WMT14	WMT15	MTNT	blind	Naver Labs Europe	41.4	47.9
MTNT baseline	28.9	30.8	23.3	25.6	33.5	33.0	21.8	22.1	ours	38.5	44.8
	20.5	50.0		23.0	55.5	55.0	21.0		Baidu & Oregon State Uni.	36.4	43.6
+ finetuning			30.3				29.7		Johns Hopkins Uni.	38.5	40.2
CUNI Transformer	42.9	39.6	39.9	42.6	43.6	41.6	34.0	37.0	Frauenhofer FOKUS-VISCOM	24.2	29.9
+ finetuning	41.5	40.9	42.1	44.8	43.5	41.6	36.6	38.5	MTNT Baseline	22.1	25.6

This research has been supported by the from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825303 (Bergamot), Czech Science Foundtion grant No. 19-26934X (NEUREM3), and Charles University grant No. 976518, and has been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071).

The poster was presented on August 2, 2019 at the Fourth Conference on Machine Translation (WMT19) in Florence, Italy.