The MLLP-UPV Supervised Machine Translation Systems for WMT19 News Translation Task

Javier Iranzo-Sánchez   Gonçal V. Garcés Díaz-Munío   Jorge Civera   Alfons Juan

www.mllp.upv.es

MLLP Machine Learning and Language Processing   VRAIN Valencian Research Institute for Artificial Intelligence   Universitat Politècnica de València

Introduction

- Neural Machine Translation (NMT) systems created for the WMT19 News Translation shared task (DE↔EN, DE↔FR)
- Transformer architecture (2017): state of the art, quick training
- Techniques applied:
  - Multi-GPU & Half-Precision to improve and speed up training
  - Corpus filtering applied on bigger, noisier ParaCrawl corpus
  - Data augmentation: Back-translations from monoling. corpora
  - Domain adaptation through fine-tuning on in-domain data

Experimental setup

Corpus filtering

- Goals: to take out the noise, to perform some domain adaptation
- Two approaches were compared:
  - **LM-based filtering**: Two 9-gram character-based LMs, one for target and one for source. Sort sentence pairs by perplexity combination ($\sqrt{s_1 \cdot s_2}$); take the n lowest-scored pairs.
  - **Dual Conditional Cross-Entropy filtering**: Sent. pairs sorted by product of partial scores: language id (*lang*), dual conditional cross-entropy (*adq*), and cross-entropy difference (*dom*).
- Cross-Entropy provided the best results
- Applied on the bigger, noisier ParaCrawl corpus

Data setup

Languages	Sentence pairs (M): WMT19 bilingual	Filtered ParaCrawl	Back-trans
DE → EN	5.5	10.0	44.0
EN → DE	5.5	10.0	18.0
DE → FR	2.5	1.0	10.0
FR → DE	2.5	1.0	18.0

- Back-translation model: Transformer Base baseline (1 GPU)
- WMT19+Filtered oversampled for 1:1 ratio with back-trans
- DE↔EN: We tested adding noise to the source side of the back-translations, jointly with no oversampling

Model configuration

- Standard Transformer “base” and “big” configurations
- Vocabulary: 40K joint BPE
- Gradient accumulation & Half-Precision training
- Software used: Fairseq NMT toolkit

System evaluation

- Fine-tuning (after training converges) on a small in-domain subset (DE↔EN: newstest08–16; DE↔FR: half of the dev data)
- Inference with checkpoint averaging (8 last checkpoints)

Languages	System	BLEU nt2018 (test)	BLEU nt2019 (hidden test)
DE → EN	Big, 8GPU	47.6	37.7
	+ fine-tuned	47.8	39.4
	+ noise	48.0	40.2
	+ fine-tuned	47.9	40.1
EN → DE	Big, 8GPU	45.7	39.4
	+ fine-tuned	48.1	41.7
DE → FR	Big, 4GPU	33.3	34.4
	+ fine-tuned	33.5	34.5
FR → DE	Big, 4GPU	24.9	26.9
	+ fine-tuned	25.4	27.5

WMT19 official results

System	Rank: Human evaluation	Auto eval. (BLEU)
DE → EN	1 / 3	6 / 11
EN → DE	2 / 4	10 / 17
DE → FR	1 / 4	3 / 6
FR → DE	2 / 3	4 / 5

Conclusions

- Transformer performance is highly dependent on batch size (multi-GPU systems and gradient accumulation are very helpful).
- Domain adaptation through fine-tuning provides improvements.
- Adding noise helps to take advantage of more back-translations. Further study is needed on effects of noise jointly with fine-tuning.

Acknowledgments

*The research leading to these results has received funding from the **European Union's Horizon 2020** research and innovation programme under grant agreement no. 761758 (X5gon); the **Government of Spain**'s research project Multisub, ref. RTI2018-094879-B-I00 (MCIU/AEI/FEDER, EU); and the **Universitat Politècnica de València**'s PAID-01-17 R&D support programme.*