eTranslation's Submissions to the WMT 2019 News Task

Csaba Oravecz, Katina Bontcheva, László Tihanyi, Adrien Lardilleux and Andreas Eisele

DG Translation - DG CNECT, European Commission

Co-financed by the Connecting Europe Facility of the European Union

Introduction

eTranslation

- is a building block of the Connecting Europe Facility (CEF) to help European and national public administrations exchange information across language barriers in the EU
- provides access to machine translation between all 26 official languages of the EU and the EEA for translators and officials in EU and national authorities
- works with engines trained on Euramis translation memories, over 1 billion sentences in the 24 official EU languages → good coverage of language used in official EU documents, not so good on other domains

Aim for the workshop

- widen the scope of the service and improve the coverage in more general types of texts
- participate in WMT News Task to find best practices that guarantee the production of solid systems in a constrained resource environment

Language Pairs

- English→German, French→German, English→Lithuanian, Russian→English

Trainings

Resources

- no large scale computing facilities
- multi-GPU setups on 4 NVIDIA P100 with 16GB RAM
- Marian toolkit (also the core of the NMT framework in the eTranslation service)

Models

- mainly base transformer, tests with big transformer (T-Big) for Fr→De and En→Lt
- hyperparameters: default settings for base transformer in Marian, dynamic batching and tying all embeddings, CE early-stopping: 5
- T-Big: `-learn-rate .0002 -lr-warmup 8000 -lr-decay-inv-sqrt 8000 -transformer-dim-ffn 4096 -transformer-heads 16`
- 30k joint SentencePiece vocabulary

Data

Selection

- baseline models: all provided original parallel data (except UN corpus for Ru→En)
- back-translation and LMs: recent target language News Crawl
- fine-tuning: previous years' test sets

Filtering

- simple general filters:
  language identification, segment deduplication,
  segment < 110 tokens,
  $\frac{1}{3} < \frac{length(\text{source})}{length(\text{target})} < 3$,
  no segment without alphabetic character
- cross-entropy filtering: in ParaCrawl and CommonCrawl parallel data sets for high resource language pairs → 40% less data with no drop in BLEU

Pre/postprocessing

- $\emptyset$ → SentencePiece with raw text input/output

Table 1: Summary of data used in training.

	True parallel	Synthetic	Mono (LM)	Fine tune
En→De	24.08M	10.00M	117.48M	32k
Fr→De	6.38M	5.23M	—	13k
En→Lt	2.21M	2.48M	0.38M	—
Ru→En	5.90M	—	—	17.8k

Improving the baseline

Fr→De: in-domain synthetic data created using guided topic modelling techniques

- generating additional synthetic data with back-translation
- using the development data (where available) to fine-tune converged models with continued trainings
- building (small) ensembles from models trained from different seeds

English→German

System	Parallel data	2018	2019
M1 Baseline	6.8M	41.3	38.1
M2 M1+PC	24M	44.6	39.9
M3 M2+BT	34M	45.4	38.7
M4 M3 ens.	34M	46.0	40.1
M5 M4+LM	34M	46.3	40.3
M6 M5+FT	34M+32k	**47.8**	**42.4**

M1: only true parallel data
M2: + filtered v3 ParaCrawl
M3: + back-translated synthetic data
M4: ensemble of 3 M3 models
M5: language model added to the ensemble with weight 0.1
M6: each model in ensemble fine tuned on devset

French→German

selection from 2008–2014 test sets

System	Parallel data	Dev	2019
M1 Baseline	2.6M	20.8	26.1
M2 M1+PC	6.9M	22.4	29.4
M3 M2+BT	11.6	22.8	33.1
M4 M3+FT	11.6M+13k	23.8	32.4
M5 M3 ens.	11.6M	22.7	**33.5**
M6 M4 ens.	11.6M+13k	**24.3**	32.7

M1: only true parallel data
M2: + filtered v3 ParaCrawl
M3: + topic selected synthetic data
M4: fine tuned M3 model (decrease on 2019!)
M5: ensemble of 2 big transformers
M6: ensemble of fine tuned big transformers

English→Lithuanian

System	Parallel data	2019D	2019
M1 Baseline	0.84M	15.5	11.4
M2 M1+PC	2.2M	19.4	12.5
M3 M2+BT	4.7M	25.7	16.6
M4 M3+OS	5.9M	25.8	15.9
M5 M4+LM	5.9M	26.1	16.0
M6 M5 ens.	5.9M	**27.0**	**17.1**

M1: only true parallel data
M2: + full v3 ParaCrawl
M3: + synthetic data from back-translating all mono (-CC)
M4: 2× oversampled Rapid and in-domain back-translated
M5: (1, 0.1) ensemble of M4 and transformer type LM
M6: ensemble of T-Bigs and tLM

Russian→English

System	Parallel data	2018	2019
M1 Baseline	2.1M	27.3	32.4
M2 M1+PC	5.9M	29.5	35.9
M3 M2+FT	5.9M+17.8k	**32.9**	**37.4**

M1: only true parallel data (-UN)
M2: + filtered v3 ParaCrawl
M3: M2 fine tuned on devset

Results and conclusions

- En→De: average result in BLEU, second cluster in Direct Assessment
- En→Lt: average result in BLEU, fourth cluster in DA
- Fr→De: second best in BLEU, first cluster in DA
- Ru→En: 4-5th in BLEU, first cluster in DA
- good selection of training data, back-translation and fine-tuning are the most rewarding
- reasonable models can be produced using established techniques in very constrained conditions