THE MLLP-UPV GERMAN-ENGLISH MT SYSTEM FOR WMT18

J. Iranzo-Sánchez P. Baquero-Arnal G. V. Garcés Díaz-Munío A. Martínez-Villaronga J. Civera A. Juan

www.mllp.upv.es

MLLP | Machine Learning and Language Processing

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

1. INTRODUCTION

- Neural Machine Translation (NMT) system created for the WMT18 News Translation shared task (DE→EN)
- NMT outperformed Phrase-Based MT in WMT16 & WMT17
- Transformer architecture (2017): state of the art, quick training
- Corpus filtering has gained importance due to bigger, noisier corpus (ParaCrawl) in WMT18
- Data augmentation: Back-translations from monolingual corpora

2. SYSTEM DESCRIPTION

- Transformer architecture: "base" configuration (65M parameters)
  - 6 self-attentive layers (both in encoder and decoder)
  - Model dimension: 512 units
  - Feed-forward dimension: 2048 units
- Vocabulary: 40K joint BPE
- Training parameters:
  - Batch size: 3000 words
  - Adam optimization
  - Label smoothing
- Software used: Sockeye NMT framework

4. TRAINING DATA

Synthetic source sentences

- Trained EN→DE NMT system on 10M filtered WMT18 corpus
- Back-translated a 20M random subset of News Crawl 2017 (EN)

Final training data

Corpus	Sent. pairs
Filtered WMT18 corpus (incl. ParaCrawl)	10 M *
Back-translations (News Crawl 2017)	20 M

* Oversampled 2×

WMT18 OFFICIAL RESULTS (HUMAN EVALUATION)

German→English

	Ave. %	Ave. z	System
1	79.9	0.413	RWTH
	79.4	0.395	UCAM
	78.2	0.359	NTT
	77.3	0.346	ONLINE-B
	77.4	0.321	MLLP-UPV
	77.0	0.317	JHU
	76.9	0.315	UBIQUS-NMT
	76.7	0.310	ONLINE-Y
	75.7	0.268	ONLINE-A
	75.4	0.261	UEDIN
11	72.5	0.162	LMU-NMT
	72.2	0.149	NJUNMT-PRIVATE
13	65.2	−0.074	ONLINE-G
14	58.5	−0.296	ONLINE-F
15	45.4	−0.752	RWTH-UNSUPER
16	42.7	−0.835	LMU-UNSUP

3. CORPUS FILTERING

Language model–based approach

- Goals: to take out the noise, to perform some domain adaptation
- Two 9-gram character-based LMs, one for target and one for source
- Trained on a small in-domain dataset (newstest2014) with SRILM
- Sort by perplexity combination ($\sqrt{s_1 \cdot s_2}$); take n lowest-scored pairs
- We filter the whole corpus as one, without distinctions

Results on corpus filtering (BLEU)

Subset (no. of sentence pairs)	nt2017	nt2018
Baseline: WMT18 minus ParaCrawl (6M)	32.0	39.1
Full WMT18 parallel dataset (42M)	21.3	26.2
Filtered corpus (5M)	31.4	38.7
Filtered corpus (7.5M)	33.7	41.5
Filtered corpus (10M)	**34.5**	**42.2**
Filtered corpus (15M)	34.3	42.2

5. SYSTEM EVALUATION

Final system

- Baseline: WMT18 corpus without ParaCrawl, 20K BPE
- Improvements: corpus filtering, synthetic data, ensembling
- Ensemble: linear combination of 4 training runs
- Training time: about 120 hours (single GPU)

Evaluation and results (BLEU)

System	nt2017	nt2018
Baseline (WMT18 minus ParaCrawl, 6M pairs)	32.0	39.1
Filtered corpus (including ParaCrawl, 10M pairs)	34.5	42.2
+ Synthetic data (2×10M + 20M pairs), 40K BPE	35.9	44.7
Ensemble (×4)	**36.2**	**45.1**

6. CONCLUSIONS

- In the **1st rank** of WMT18 DE→EN News Translation official results
- A **competitive NMT system** with a short training time
- Based on Transformer architecture (a trend in WMT18 systems)
- Corpus filtering is key with larger, noisier corpora

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 TIN2015-68326-R (MINECO/FEDER) research project MORE, university collaboration grant programme 2017-2018, and faculty training scholarship FPU13/06241; the **Generalitat Valenciana**'s predoctoral research scholarship ACIF/2017/055; as well as the **Universitat Politècnica de València**'s PAID-01-17 R&D support programme.*