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

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)

<table>
<thead>
<tr>
<th>Subset (no. of sentence pairs)</th>
<th>nt2017</th>
<th>nt2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: WMT18 minus ParaCrawl (6M)</td>
<td>32.0</td>
<td>39.1</td>
</tr>
<tr>
<td>Full WMT18 parallel dataset (42M)</td>
<td>21.3</td>
<td>26.2</td>
</tr>
<tr>
<td>Filtered corpus (5M)</td>
<td>31.4</td>
<td>38.7</td>
</tr>
<tr>
<td>Filtered corpus (7.5M)</td>
<td>33.7</td>
<td>41.5</td>
</tr>
<tr>
<td>Filtered corpus (10M)</td>
<td>34.5</td>
<td>42.2</td>
</tr>
<tr>
<td>Filtered corpus (15M)</td>
<td>34.3</td>
<td>42.2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sent. pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtered WMT18 corpus (incl. ParaCrawl)</td>
<td>10 M</td>
</tr>
<tr>
<td>Back-translations (News Crawl 2017)</td>
<td>20 M</td>
</tr>
</tbody>
</table>

* Oversampled 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)

<table>
<thead>
<tr>
<th>System</th>
<th>nt2017</th>
<th>nt2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (WMT18 minus ParaCrawl, 6M pairs)</td>
<td>32.0</td>
<td>39.1</td>
</tr>
<tr>
<td>Filtered corpus (including ParaCrawl, 10M pairs)</td>
<td>34.5</td>
<td>42.2</td>
</tr>
<tr>
<td>+ Synthetic data (2×10M + 20M pairs), 40K BPE</td>
<td>35.9</td>
<td>44.7</td>
</tr>
<tr>
<td>Ensemble (×4)</td>
<td>36.2</td>
<td>45.1</td>
</tr>
</tbody>
</table>

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. 767158 (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.