Why don’t people use character-level MT?

Jindřich Libovický¹, Helmut Schmid², Alexander Fraser²
¹ Charles University, ² LMU Munich

May 22–27, 2022
Outline

1. Extensive survey of research papers and WMT submissions.

2. Explore both existing and new character-level architectures.

Motivation

Subwords are sort of ugly

_The_ _c at _s le eps _on _a _m at .

Wishful thinking: what we could get from the character-level

• Simpler processing pipelines
• Learn better segmentation

• Noise robustness
• Generalize towards morphology and domain-specific vocab
Character-level MT in time

<table>
<thead>
<tr>
<th></th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subwords</td>
<td>92%</td>
<td>93%</td>
<td>97%</td>
</tr>
<tr>
<td>Morphological Words</td>
<td>4%</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Words</td>
<td>2%</td>
<td>3%</td>
<td>—</td>
</tr>
<tr>
<td>Character</td>
<td>2%</td>
<td>2%</td>
<td>—</td>
</tr>
</tbody>
</table>

- Research papers often report parity or outperforming subwords
- The results of research papers got never confirmed in the competitive WMT setup
- Suspected reasons:
  - not better quality, 5–6× slower
State-shrinking and Two-step decoding

Explore various architectures on small data
Convolutional encoder w/ downsampling + vanilla decoder

Why don’t people use character-level MT?
Previous work makes optimistic conclusions based on small and old datasets... …let’s do it properly

- CzEng 2.0 corpus
- 61M authentic parallel sentences
  50M back-translated

- Data mix Edinburgh used for WMT’21 submission
- 66M authentic parallel sentence
  52M back-translated

...data almost comparable to best WMT submissions
*(tagged back-translation, Transformer BIG architecture, FairSeq)*
Extensive evaluation

Character-level methods often motivated by morphological generalization and noise robustness.

- Quality in News, IT and medical domain
- Gender dataset
- Morpheval: Specific morphological phenomena
- Recall of novel forms and lemmas (in news)
- Quality under sampled noise

Characters are better in noise robustness
Why don’t people use character-level MT?

Summary

• Research in character-level MT is not used in practice
• Machine translation benefits from word-like units
• The best character-level architecture: convolutions + downsampling
• The only advantage of character-level: noise robustness

https://ufal.mff.cuni.cz/jindrich-libovicky