Findings of ACL 2022

Why don't people use character-level MT?

Jindřich Libovický[1], Helmut Schmid[2], Alexander Fraser[2]
[1]Charles University, [2] LMU Munich

📅 May 22–27, 2022

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

CC BY NC SA
unless otherwise stated

Outline

1. Extensive survey of research papers and WMT submissions.

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

3. Systematic evaluation of WMT-scale models.

Motivation

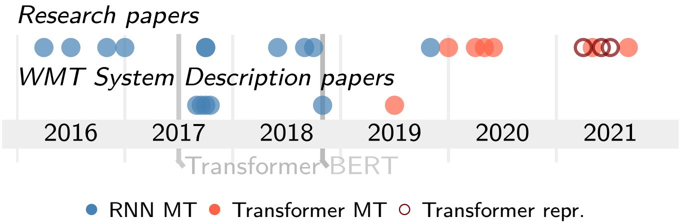
Subwords are sort of ugly

`_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

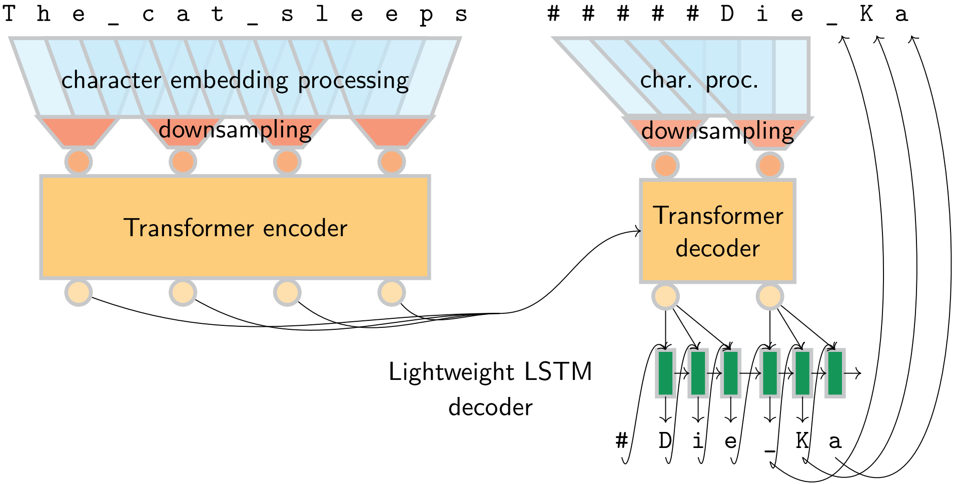


	2018	2019	2020
Subwords	92%	93%	97%
Morphological	4%	2%	3%
Words	2%	3%	—
Character	2%	2%	—

- 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

Competitive data setup

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

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`