Character-level MT is good for noise robustness and not much else

Why don’t people use character-level machine translation?

Jindřich Libovický
libovicky@ufal.mff.cuni.cz

Helmut Schmid
schmid@cis.lmu.de

Alexander Fraser
fraser@cis.lmu.de

1. Extensive survey of research papers and WMT submissions.

- Research papers claim parity or superiority of char-level
  models over subwords
- Character-level model hardly ever used in competitive
  WMT setups (>90% submission use subwords)
- Char-level model 5-6x slower than subwords
  → standard WMT methods unfeasible

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

- Architecture exploration on small IWSLT data
  en ↔ {de, fr, ar}
- Various architectures for char processing
  - 1D Convolution + Max-pool
  - CANINE = local self-attention + 1D convolution
  - Charformer = based on n-gram averaging
  - Standard and vs fast novel 2-step decoder

  Winner: 1D convolution + Max pool + Vanilla decoder


- Use the best architecture from the small data
  experiments
- Use the same data as in used competitive WMT
  submissions (incl. back-translation)
- English → Czech
  - CzEng 2.0 dataset
  - 61M authentic sentences, 50M back-translated
- English → German
  - Data mix used in Edinburgh’s WMT21 submission
  - 66M authentic sentences, 52M back-translated

  Evaluation to assess often claimed advantages of character-level methods
  - Quality in news, IT, medical domain
    worse overall, consistent over domains
  - Gender evaluation dataset
    no clear advantage
  - Morphology using Morpheval benchmark
    German seems slightly better, no difference for Czech
  - Recall of novel forms and lemmas
    no difference between subwords and characters
  - Robustness towards source-side noise
    character-level clearly better