Breaking Words Better: Linguistically Guided Segmentation for Morphologically rich Languages

Çağla Ece Azizoğlu, Aqil Ahmed Abdul Khaliq, Nyi Nyi Linn Htet

Check our code here

Why Morphology Matters

Some languages pack meaning into word endings. One token can encode what English needs multiple words to express.

Example: talossanikin (Finnish) → also in my house (English)

- Morphological stacking → many surface forms per lemma
- Data sparsity: rare inflections are hard to learn
- BPE is frequency-based and doesn't produce linguistically meaningful units
- Morpheme-aware units to improve alignment + generation

Method: Data & Model

Task: Finnish → English translation
Model: Transformer (fairseq IWSLT-style)
Segmentation: BPE vs Morfessor variants
Metrics: sacreBLEU, COMET

Dataset splits (data from WMT)

- Train: 496,334 sentence pairs
- Dev: 3000 sentence pairs
- Test: 3000 sentence pairs

FAIRSEQ

fairseq (NLP) is a PyTorch-based toolkit from FAIR for training state-of-the-art NLP sequence models like transformers for machine translation, summarization, and language modeling.

Pipeline

text → segment → train → evaluate

Segmentation Pipelines Compared

Same Transformer, different segmentation. Which side should be morphology-aware?

Pipeline A) Baseline

raw text → bpe → Transformer → English

Pipeline B) Morfessor source only

raw text → morfessor → bpe → Transformer → English

BPE is applied to control vocabulary size and handle rare or unseen words by splitting text into frequent subword units.

BPE

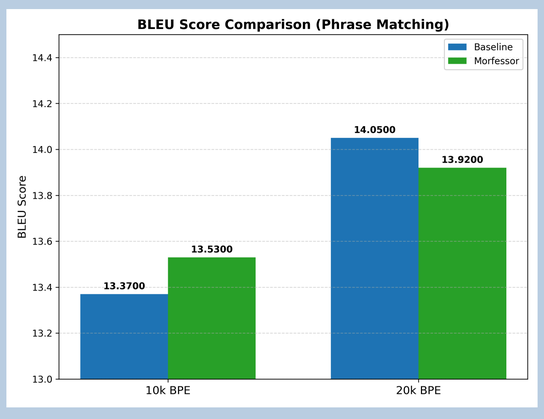
- BPE = subword tokenization
- splits into common chunks
- helps with rare words
- starts from chars
- merges most common pair repeatedly

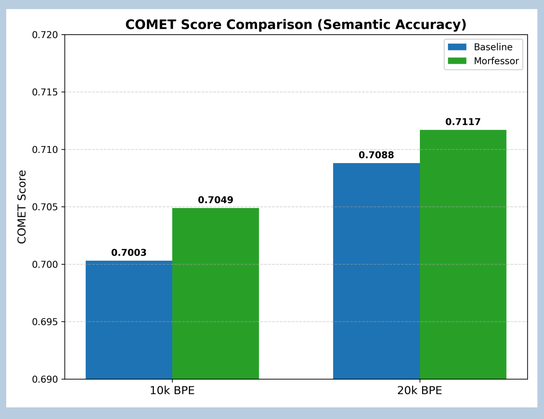
Morfessor

- Morfessor = unsupervised morphological segmentation
- splits words into morphemes
- targets linguistic structure
- unlike BPE (frequency-based)
- not mainly for compression

Morfessor is added to introduce linguistically motivated word structure, preserving meaningful units that BPE alone may miss.

Results





We compare two different vocabulary sizes: **10,000** & **20,000**

- **BLEU**: Measures n-gram overlap with reference translations; easy to compute and compare, but insensitive to meaning and fluency.
- **COMET**: Uses a neural model to assess semantic adequacy and fluency; aligns better with human evaluations, but is more complex.

Discussion & Takeaways

Morfessor helps by reducing sparsity and aligning subunits with meaning-bearing morphemes.

Key Takeaways

- We observe an improvement
  - in COMET scores
  - in BLEU scores for the 10K

Future work

- Tune Morfessor hyperparameters / vocab control
- Translating from English to Finnish
- Add human eval for inflection & aggreement errors
- Test Turkish / Hungarian /Estonian