Subword Segmental Machine Translation: Unifying Segmentation and Target Sentence Generation

Francois Meyer and Jan Buys, University of Cape Town

Findings of ACL 2023

Motivation

- Subword segmenters like BPE are preprocessing steps, completely separated from MT training.
- MT models are reliant on the subwords produced by segmenters - they cannot adjust or improve them.
- Poor subwords hurt performance, especially for low-resource languages that are morphologically complex.

In this paper we move subword segmentation from preprocessing into MT training. We propose SSMT - a model that simultaneously learns how to generate translations and how to segment translated sentences into subwords. Our model can learn subwords that optimise its MT training objective.

Training

For any specific subword segmentation $\mathbf{s}$ of the raw target sentence $\mathbf{y}$, we use the chain rule to compute

$$p(\mathbf{s}|\mathbf{x}) = \prod_{i=1}^{|\mathbf{s}|} p(s_i|\mathbf{s_{<i}}, \mathbf{x}).$$

We still segment the source sentence $\mathbf{x}$ with BPE, but the **segmentation of the target sentence is a latent variable** that is marginalised over as

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathcal{S}} p(\mathbf{s}|\mathbf{x}),$$

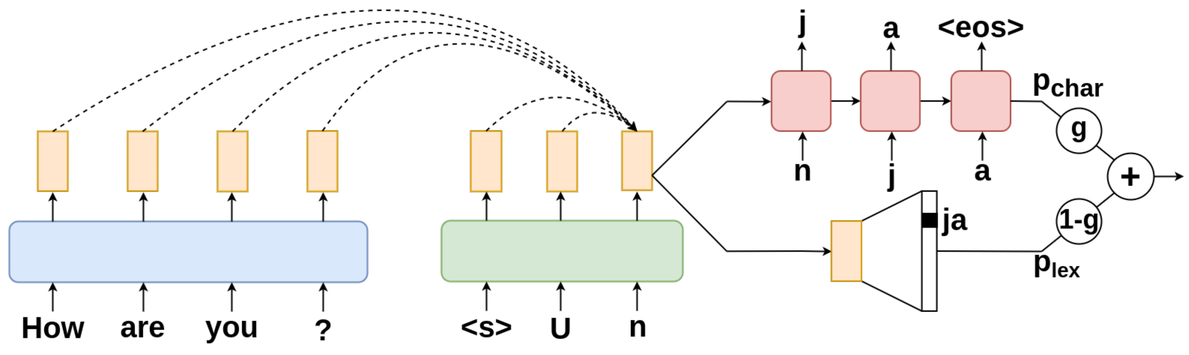
where $\mathcal{S}$ is the set of all possible subword segmentations of the target sentence. We marginalise efficiently using **dynamic programming**.

Contributions

1. Subword segmental machine translation (SSMT) - a model that jointly learns **target-side subword segmentation** and **machine translation**.
2. Dynamic decoding - a decoding algorithm that **reconsiders and adjusts the subword segmentation** of a translation as it is generated.
3. MT experiments across **6 morphologically diverse languages** and linguistically informed **analyses of the subwords** learned by SSMT.

Subword Segmental Machine Translation (SSMT)

The SSMT architecture is a standard Transformer encoder and a *subword segmental* Transformer decoder. At every character position in the target sentence, the probabilities of all possible subsequent subwords are computed.



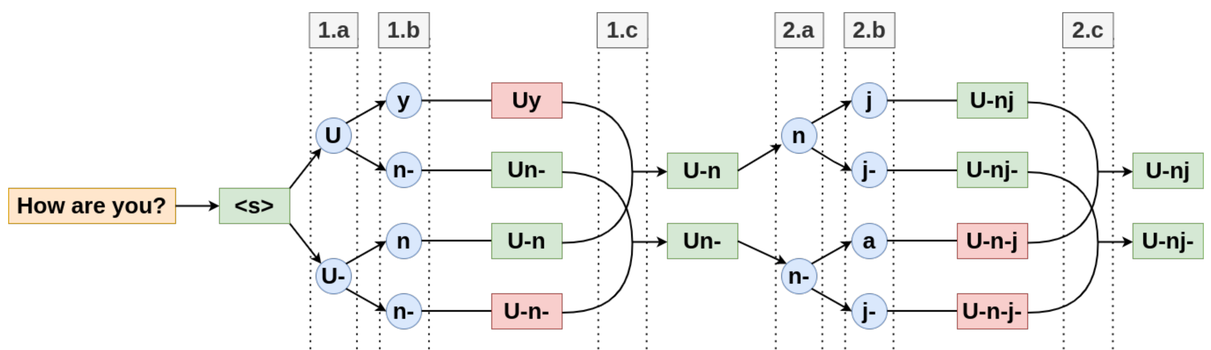
SSMT translates "How are you?" to the Zulu "Unjani?", computing the probability for subword "ja". A Transformer encoder-decoder encodes the BPE-segmented source sentence and character-level target sentence. A mixture between a character decoder and lexicon model produces the next subword probability:

$$p(s_i|\mathbf{s_{<i}}, \mathbf{x}) \approx p(s_i|\pi(\mathbf{s_{<i}}), \mathbf{x}) = g_j p_{\text{char}}(s_i|\mathbf{y_{<j}}, \mathbf{x}) + (1 - g_j) p_{\text{lex}}(s_i|\mathbf{y_{<j}}, \mathbf{x}),$$

where $\pi(\mathbf{s_{<i}})$ is a concatenation operator that converts the sequence $\mathbf{s_{<i}}$ into the raw unsegmented characters $\mathbf{y_{<j}}$ preceding subword $\mathbf{s_i}$.

Dynamic Decoding

Beam search cannot be directly applied to SSMT, because SSMT has 2 vocabularies in its mixture model. We propose dynamic decoding, a text generation algorithm that computes next-character probabilities based on the mixture model and continually adjusts its preferred subword segmentations during generation.



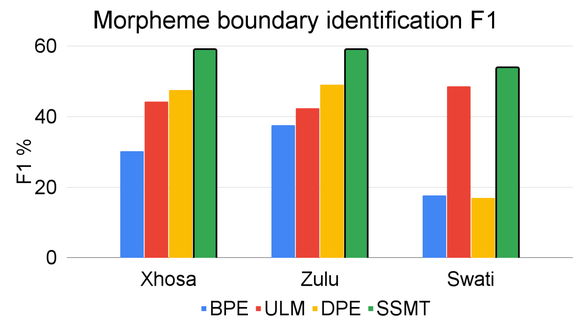
Ablation Study

Model variant	chrF
2 models: segment + translate	
+BPEvocab –char (DPE)	23.3
+lexicon –char (SSMT –char)	23.7
+lexicon +char (SSMT)	23.1
1 model with dynamic decoding	
+lexicon –char (SSMT –char)	26.2
+lexicon +char (SSMT)	**26.4**

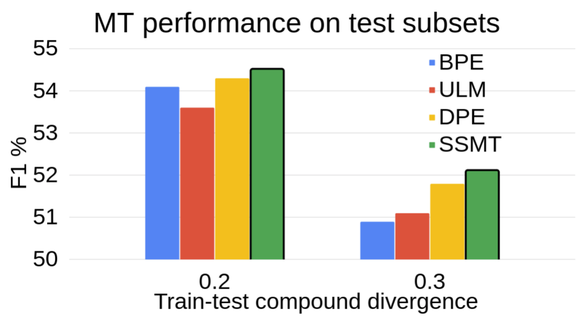
Different model settings for English → Swati. Unifying segmentation and generation is the key factor in SSMT's performance gains.

Subword Analysis

To what extent does SSMT discover morphemes? We extract subwords learned by SSMT and compare them to annotated morphemes. SSMT subwords are closer to morphemes than baseline segmenters.



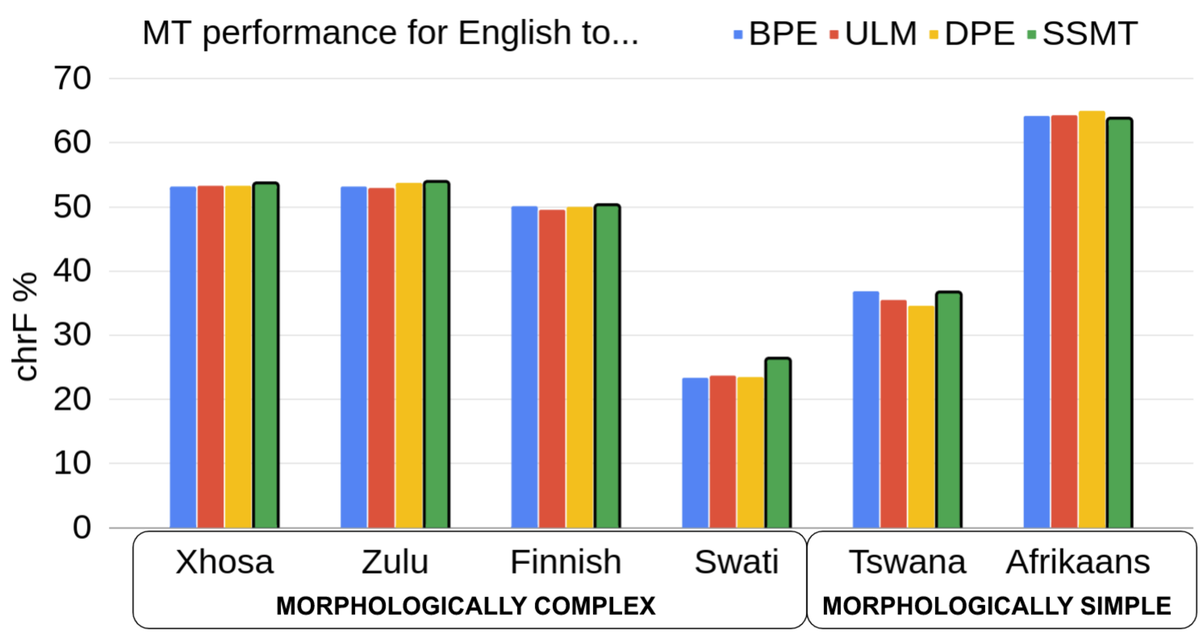
Morphological compositional generalisation We extract subsets of the English-Zulu test set and control the degree to which they contain previously unseen morpheme combinations. SSMT proves more robust on the more challenging subset.



Main Findings

- Improvements for conjunctively written agglutinative languages, but not for disjunctively written or analytic languages.
- SSMT's biggest gains are for the extremely low-resource case of English to Swati translation.
- The subwords learned by SSMT are closer to actual morphemes than BPE, ULM, and DPE subwords.
- SSMT generalises better than baselines to previously unseen combinations of morphemes.

Machine Translation Results



Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number: 129850). Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: `hpc.uct.ac.za`. Francois Meyer is supported by the Hasso Plattner Institute for Digital Engineering, through the HPI Research School at the University of Cape Town.

UCT NLP

http://www.janmbuys.com/uctnlp

Authors:
– MYRFRA008@myuct.ac.za,
– jbuys@cs.uct.ac.za