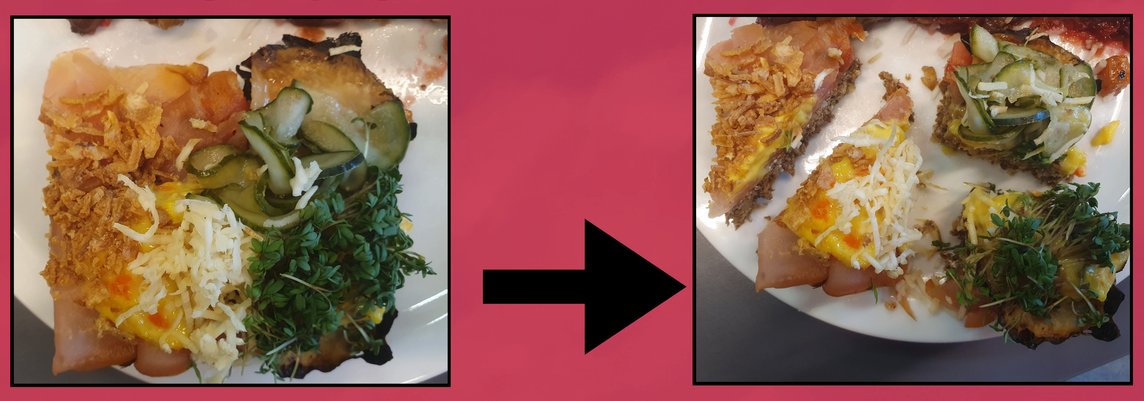
MorSeD: Morphological Segmentation of Danish and its Effect on Language Modeling

Rob van der Goot, Anette Jensen, Emil Allerslev Schledermann, Mikkel Wildner Kildeberg, Nicolaj Larsen, Mike Zhang, Elisa Bassignana

Morphological Segmentation

Converting words into morphemes (smallest meaning carrying units)



Smørrebrød → Smør-re-brød

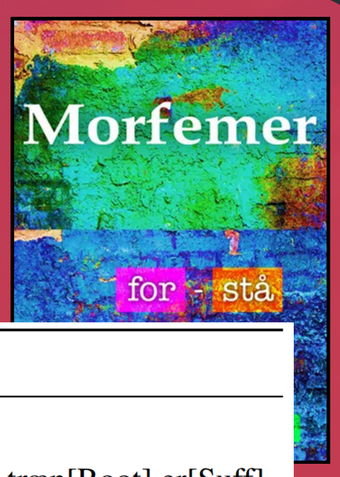
Why?

Language model use input units which are not linguistically motivated.

tænke -> t_æ_n_ke
luftpudefartøj -> lu_ftp_ude_f_art_ø_j

MorSeD

A Danish dataset annotated for morphological segmentation with the following categories:



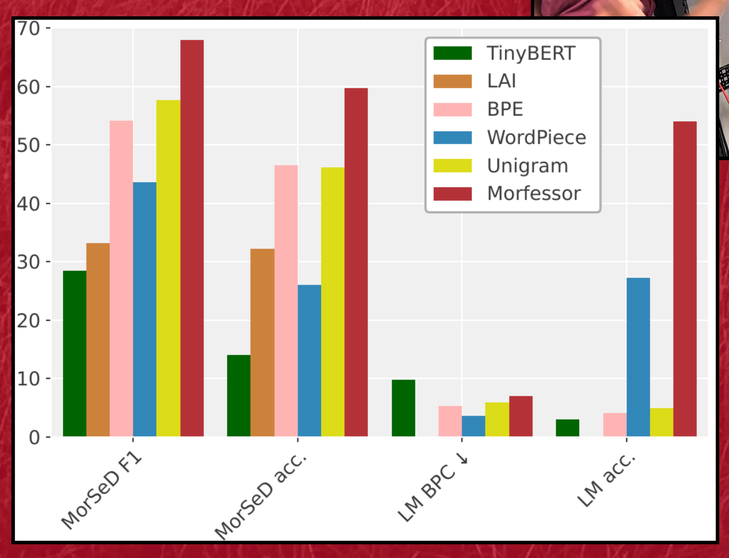
Word	Category	Split
Kranie	Root Morpheme	Kranie [Root]
Landstræner	Linking Morpheme	Land[Root] s[Link] træn[Root] er[Suff]
Lånte	Inflection	Lån[Root] te[Infl]
Bibringe	Prefix	Bi[Pref] bringe[Root]
Skoletaske	Compound	Skole [Root] taske [Root]
Venlig	Suffix	Ven[Root] lig[Suff]

Methods

Input:	måneraket	lærte
morsed:	måne-raket	lær-te
BPE:	må-ner-ak-et	lærte
WordPiece:	måne-rak-et	lærte
Unigram:	må-ne-rak-et	lærte
Morfessor:	måne-raket	lært-e
TinyBERT:	mane-rak-et	l-æ-rte

Input, gold, and output of segmentation algorithms, including popular subword segmentation algorithms, a morphological segmenter, and an English baseline.

Results



Scores for morphological segmentation (left), and language modeling (right).

Takeaways

- MorSeD: Wide coverage morphological segmentation dataset annotated by expert
- Subword algorithms do not correlate well with morphemes
- Language modeling (probably) improves when using units resembling morphemes

IT UNIVERSITY OF CPH

https://bitbucket.org/robvanderg/morsed