BPE-knockout:

Pruning Pre-existing BPE Tokenisers with Backwards-compatible Morphological Semi-supervision

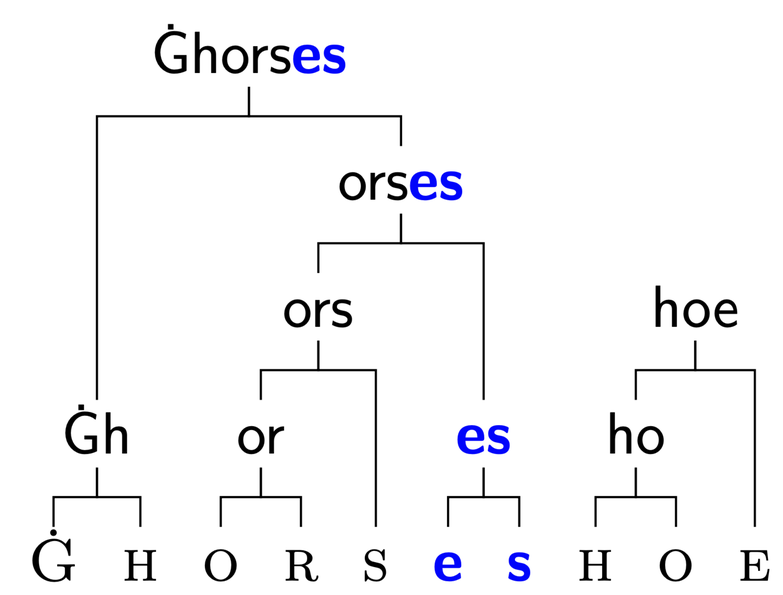
Thomas Bauwens & Pieter Delobelle
ThomasBauwens_ pieterdelobelle

DTAI Declarative Languages & Artificial Intelligence

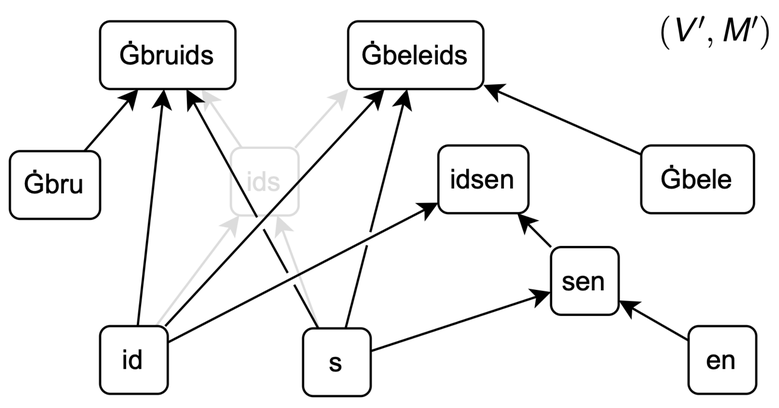
BPE tokens are not morphemes

	BPE	morphology
Sennrich et al. (2016)	Gesundheits|forsch|ungsin|stitute	Gesund|heit|s|forschung|s|institute
Bostrom et al. (2020)	comple|t|ely	complete|ly
	t|ric|y|cles	tri|cycle|s
	n|an|ote|chn|ology	nano|techno|logy
He et al. (2020)	vul|n|era|bility	vulner|ability
	t|ighter	tight|er
	emb|arked	embark|ed
	predic|table	predict|able
Vilar et al. (2021)	bewer|t|ungsin|stru|mente	bewert|ung|s|instrumente
	gefan|gen|genommen	ge|fangen|ge|nommen
	verbrau|ch|spru|ef|standard	verbrauch|s|pruef|standard
	haushalt|war|enab|teilung	haushalt|waren|abteilung
	not|arz|tau|tos	notarzt|auto|s

Morpheme boundaries disappear



Knockout: bypass tokens after blaming their merges

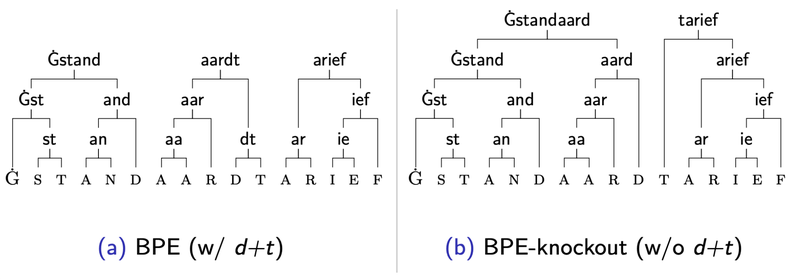


Blame merge when it contracts a morpheme boundary, then get: $R(m) = \frac{B(m)}{N(m)} \geq \frac{1}{2}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Char:	d	o	c	t	o	r	a	a	t	s	m	i	s	e	r	i	e
BPE:	d	o	c	t	o	r |	a	a	t |	s	m |	i	s	e |	r	i	e
Gold:	d	o	c	t	o	r |	a	a	t |	s |	m	i	s	e	r	i	e

Better boundaries

		$\|V\|$	word types Pr	word types Re	word types F_1	word tokens Pr	word tokens Re	word tokens F_1
English	BPE	40 000	43.0	49.6	46.1	40.7	4.0	7.3
	BPE-knockout	36 814	**53.2**	**75.1**	**62.3**	**84.5**	**59.2**	**69.6**
German	BPE	40 000	45.0	54.0	49.1	54.3	8.4	14.5
	BPE-knockout	35 919	**55.3**	**79.8**	**65.3**	**59.5**	**69.2**	**64.0**
Dutch	BPE	40 000	52.6	55.3	53.9	54.3	11.0	18.4
	BPE-knockout	35 525	**61.7**	**78.2**	**68.9**	**81.6**	**64.1**	**71.8**



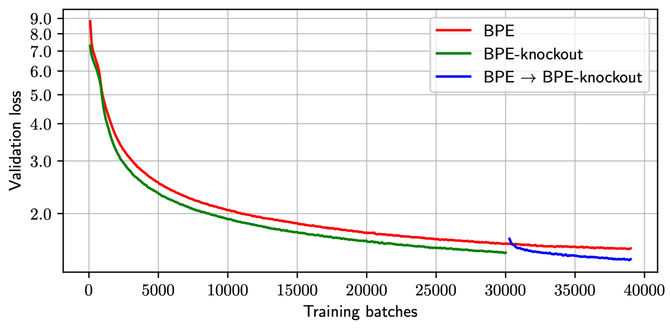
(a) BPE (w/ $d{+}t$)

(b) BPE-knockout (w/o $d{+}t$)

Switch at finetuning time = profit!

Pre-training RoBERTa

		PPPL 30k	PPPL +5k	PPPL +9k
Dutch	BPE	**3.88**	**3.74**	**3.66**
	BPE-knockout	4.67		
	BPE → BPE-knockout	*200.37*	4.62	4.35



Finetuning RoBERTa: token-level tasks benefit most

		Sequence-level SA 30k	SA +5k	SA +9k	NLI 30k	NLI +5k	NLI +9k	Token-level NER 30k	NER +5k	NER +9k	PoS 30k	PoS +5k	PoS +9k
Dutch	BPE	**82.10**	81.56	81.43	**83.16**	82.55	**83.53**	80.04	83.18	83.98	93.50	85.91	93.78
	BPE-knockout	82.01			82.98			**86.21**			91.19		
	BPE → BPE-knockout	81.34	**82.42**	**81.74**	81.74	**82.59**	83.14	80.58	**86.77**	**87.51**	**94.82**	**96.03**	**96.01**

Key takeaway

A BPE tokenizer can be improved after a model has been pre-trained on it, but more research is needed to exploit morphology in tokenizers.

KU LEUVEN LEUVEN.AI Artificial Intelligence Flanders fwo

Learn more at *pieter.ai/* ***bpe-knockout***