On the Importance of Subword Information for Morphological Tasks in Truly Low-Resource Languages

Yi Zhu[1] Benjamin Heinzerling[2,3] Ivan Vulić [1] Michael Strube[4] Roi Reichart[5] Anna Korhonen[1]

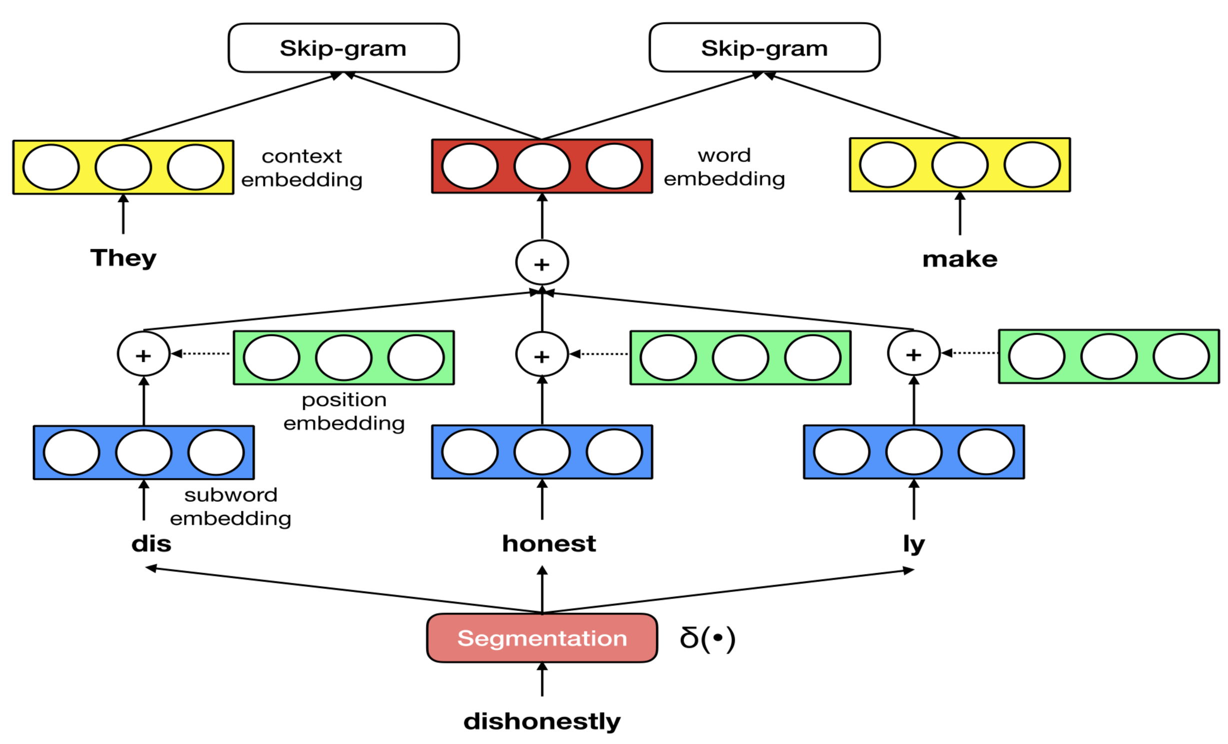
[1]Language Technology Lab, University of Cambridge [2]RIKEN AIP [3]Tohoku University [4]Heidelberg Institute for Theoretical Studies [5]Technion, IIT

Heidelberg Institute for Theoretical Studies

tl;dr

- Subwords are great, but how do they perform in low-resource settings?
- This work: thorough analysis of several subword methods
- Three morphological tasks, 16 languages
- Simulated and actual low-resource settings
- Simulate two types of scarcity: scarce embedding training data and scarce task data
- Scarcity of task data has a much larger impact
- No subword method best in all settings, but character n-gram often strongest, followed by BPE.

Subword-informed Word Representations



Subword Methods

Word	*dishonestly*
`morf`	(*dishonest*, *ly*)
`charn`	(*dis*, *ish*, *sho*, ..., *tly*, *dish*, *isho*, *shon*, ..., *stly*, *disho*, *ishon*, *shone*, ..., *estly*, *dishon*, *ishone*, *shones*, ..., *nestly*)
`bpe1e3`	(*d*, *ish*, *on*, *est*, *ly*)
`bpe1e4`	(*dish*, *on*, *est*, *ly*)
`bpe1e5`	(*dishonest*, *ly*)

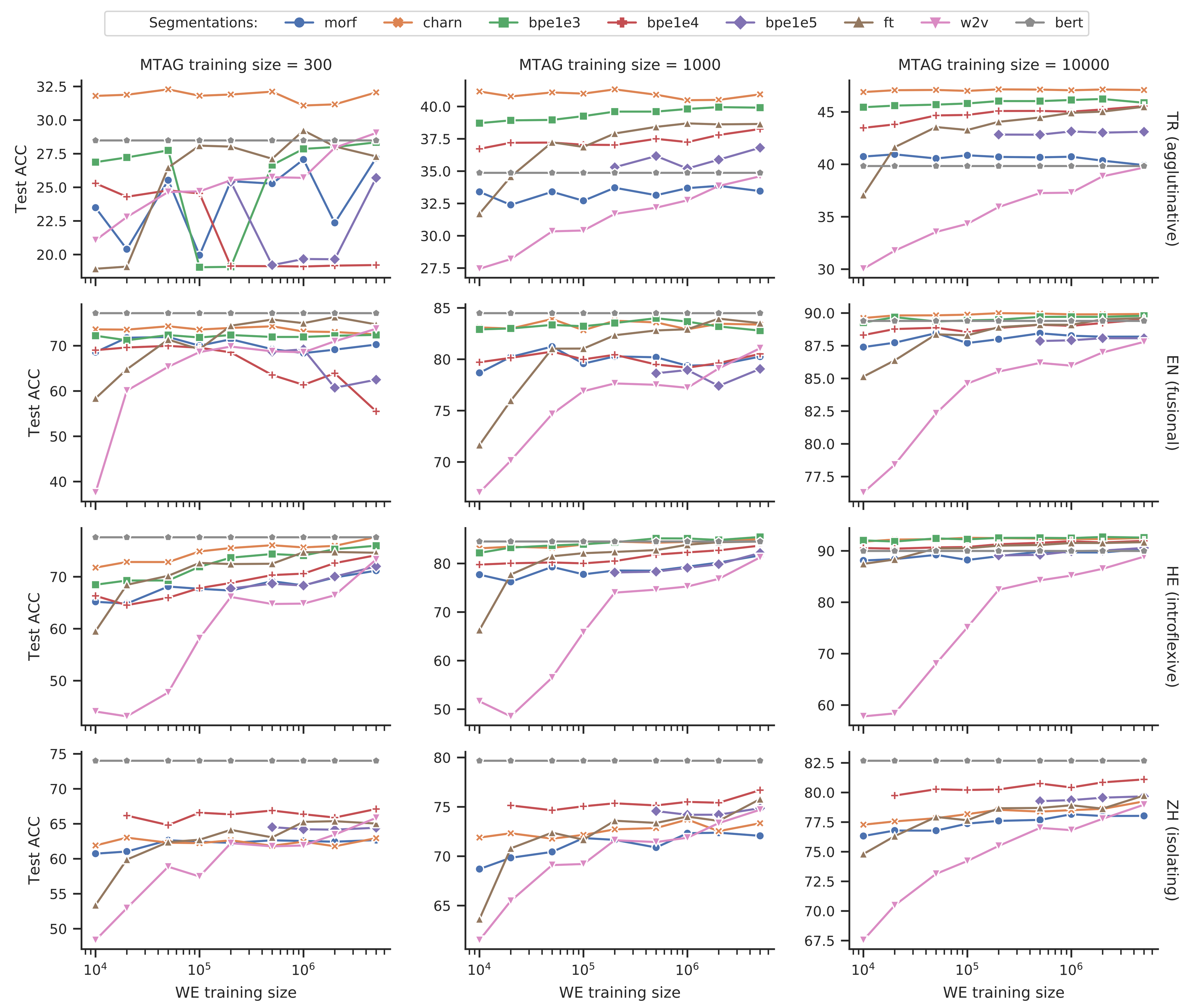
Three Morphological Tasks

- Fine-grained entity typing (FGET): *Lincolnshire* → /location/county
  Data: Wikidata + Freebase
- Morphological tagging (MTAG): *her* → {Gen=Fem, Num=Sing, Per=3, Poss=Yes, PronType=Prs}
  Data: Universal Dependencies
- Named entity recognition (NER): *Barack Obama* (→ person) *was born in Hawaii* (→ location).
  Data: WikiAnn

16 Languages: Embedding and Task Data (Tokens)

	Agglutinative						Fusional						Introflexive		Isolating	
	BM	BXR	MYV	TE	TR	ZU	EN	FO	GA	GOT	MT	RUE	AM	HE	YO	ZH
EMB	40K	372K	207K	5M	5M	69K	5M	1.6M	4.4M	18K	1.5M	282K	659K	5M	542K	5M
FGET	29K	760	740	13K	60K	36K	60K	30K	56K	289	2.7K	1.5K	2.2K	60K	15K	60K
NER	345	2.4K	2.1K	9.9K	167K	425	8.9M	4.0K	7.6K	475	1.9K	1.6K	1.0K	107K	3.4K	–
MTAG	–	–	–	1.1K	3.7K	–	24K	–	–	3.4K	1.1K	–	–	5.2K	–	4.0K
BERT				✓	✓		✓		✓					✓	✓	✓

Simulated Low-resource Setting



Only MTAG results shown. Check paper for NER and fine-grained entity typing results!

Truly Low-resource Setting

		Agglutinative					Fusional					Intro	Isolat
		BM	BXR	MYV	TE	ZU	FO	GA	GOT	MT	RUE	AM	YO
FGET	`morf`	52.43	52.47	79.11	57.79	53.00	54.43	50.77	29.90	49.48	50.38	41.82	83.43
	`charn`	**56.09**	57.33	81.69	58.83	**56.34**	**58.44**	**52.62**	**34.02**	**54.46**	**58.59**	45.65	84.85
	`bpe1e3`	53.61	51.30	81.13	58.73	55.41	56.04	50.74	31.55	52.79	55.57	**47.99**	**85.22**
	`bpe1e4`	54.20	53.81	**81.93**	**59.24**	55.67	56.67	51.47	26.39	52.15	54.81	47.05	84.42
	`bpe1e5`	-	53.80	80.00	58.13	-	56.31	51.52	-	51.52	52.52	44.74	83.39
	`ft`	51.91	**57.96**	81.05	57.79	52.62	53.74	49.67	31.96	53.95	53.64	44.80	83.71
	`w2v`	52.28	42.19	76.86	56.99	52.95	53.07	49.07	24.53	46.61	47.36	36.81	82.56
	`bert`	-	-	-	49.20	-	-	47.09	-	-	-	-	81.76
NER	`morf`	73.29	76.58	83.40	77.01	65.22	84.29	86.94	59.49	74.37	81.87	66.67	90.01
	`charn`	**83.02**	**81.59**	**93.22**	**88.23**	**74.47**	**91.08**	88.95	**84.99**	**83.56**	**88.70**	**72.92**	94.68
	`bpe1e3`	77.22	79.33	89.00	85.82	71.91	89.73	89.18	81.03	81.63	85.30	70.84	92.35
	`bpe1e4`	76.43	79.73	89.00	85.44	65.22	89.25	88.48	70.59	80.26	86.39	64.07	92.47
	`bpe1e5`	-	80.65	89.36	84.02	-	88.66	**89.48**	-	81.64	86.12	68.95	93.07
	`ft`	73.29	79.81	88.57	86.88	58.16	89.48	89.18	58.16	81.64	83.54	68.29	92.58
	`w2v`	69.57	79.66	87.50	82.97	62.37	87.81	87.99	58.56	79.43	84.21	61.37	89.57
	`bert`	-	-	-	82.31	-	-	88.45	-	-	-	-	**95.53**

Takeaways

- Scarcity of task data has a much larger impact than scarcity of embedding data.
- Subword-informed architectures are better than word-based methods in most cases, particularly in low resource settings (e.g., ZU, BM, GOT).
- When available, multilingual BERT performs well in MTAG and NER, but subword models are better in FGET. Gap becomes smaller or disappears with more embedding training data.
- No one-size-fits-all method, but character n-grams often strongest, followed by BPE.

This work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909) and the Klaus Tschira Foundation, Heidelberg, Germany.