From Characters to Words to in Between: Do We Capture Morphology?

Clara Vania and Adam Lopez
{c.vania@ed.ac.uk, alopez@inf.ed.ac.uk}

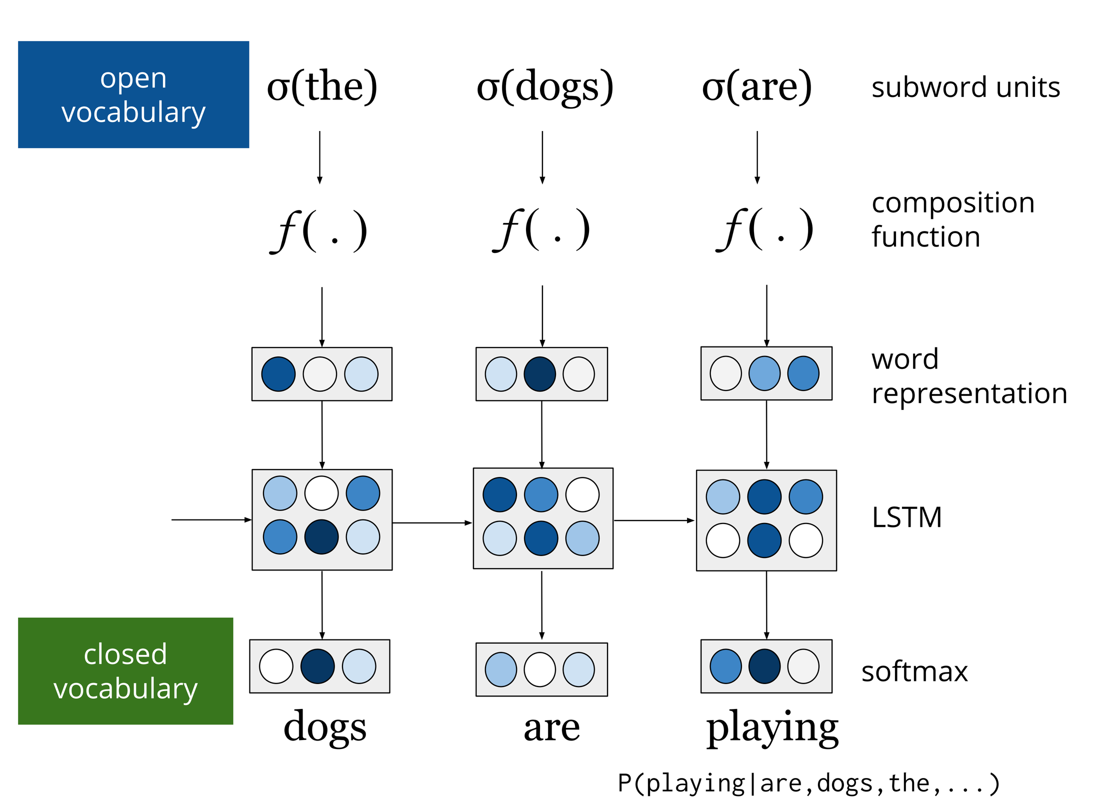
Institute for Language, Cognition, and Computation,
School of Informatics, University of Edinburgh

Motivation

Continuous word representations composed from subword representations have shown to be effective for learning the morphological regularities of words. But, some questions remain:

- Type of subword units: characters vs. morphemes?
- How to compose them: addition, bi-LSTM, or CNN?
- Do character-level models capture morphology in terms of predictive utility?
- How do they interact with languages of different morphological typologies?

Task: Language Modeling



Variable: Subword Units

Unit	Output of σ(*wants*)
Morfessor	want, s
BPE	w, ants
char-trigram	^wa, wan, ant, nts, ts$
character	w, a, n, t, s
analysis	want, +VB, +3rd, +SG, +PRS

Variable: Composition Function

- vector addition, bi-LSTM, CNN.

Variable: Language Typology

Concatenative

Agglutinative (Turkish)
oku-r-sa-m
read-AOR.COND.1SG
'If I read ...'

Fusional (English)
read-s
read-3SG.PRS
'reads'

Non-concatenative

Root&Pattern (Arabic)
k⟨a⟩t⟨a⟩b⟨a⟩
write-PST.3SG.M
'he wrote'

Reduplication (Indonesian)
anak~anak
child-PL
'children'

Perplexity Results

Typology	lang	word	character		char-trigram		BPE		Morfessor		%imp
			bi-lstm	CNN	add	bi-lstm	add	bi-lstm	add	bi-lstm	
Fusional	Czech	41.5	34.2	36.6	42.7	**33.6**	50.0	33.7	47.7	36.9	19.0
	English	46.4	43.5	44.7	45.4	**43.0**	47.5	43.3	49.7	49.7	7.4
	Russian	34.9	28.4	29.5	35.2	**27.7**	40.1	28.5	39.6	31.3	20.6
Agglutinative	Finnish	24.2	20.1	20.3	24.9	**18.6**	26.8	19.1	27.8	22.5	23.1
	Japanese	98.1	98.1	**91.6**	102.0	101.1	126.5	96.8	112.0	99.2	6.6
	Turkish	67.0	54.5	55.1	**50.1**	54.2	59.5	57.3	62.2	62.7	25.2
Root & Pattern	Arabic	48.2	42.0	43.2	50.9	**39.9**	50.9	42.8	52.9	45.5	17.3
	Hebrew	38.2	31.6	33.2	39.7	**30.4**	44.2	32.9	44.9	34.3	20.5
Reduplication	Indonesian	46.1	45.5	46.6	58.5	46.0	59.2	**43.4**	59.3	44.9	5.9
	Malaysian	54.7	53.0	**50.6**	68.5	50.7	69.0	51.2	68.2	52.5	7.5

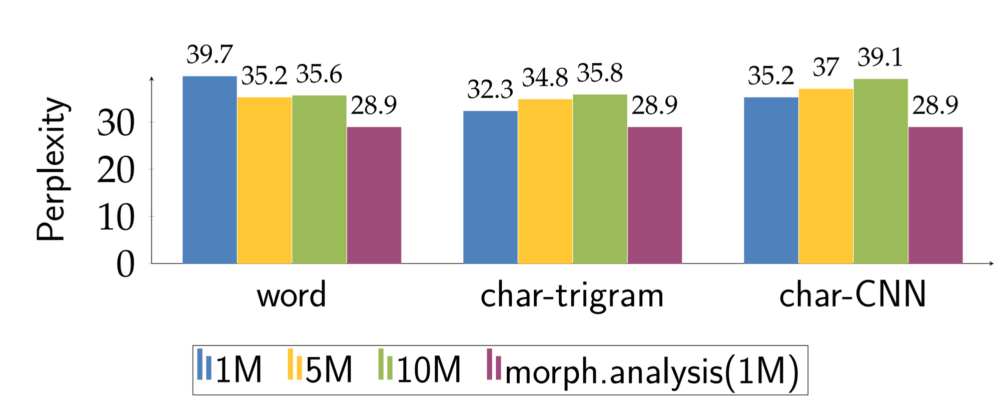
Qualitative Analysis

Subword Unit	In-Vocabulary	Rare	OOV
	including	*unconditional*	*uploading*
BPE bi-LSTM	called involve like creating	unintentional ungenerous unanimous unpalatable	upbeat uprising handling hand-colored
character trigram bi-LSTM	include includes undermining include	unconstitutional constitutional unimolecular medicinal	drifted affected conflicted convicted
character bi-LSTM	inclusion insularity includes include	undamaged unmyelinated unconditionally uncoordinated	musagète mutualism mutualists meursault

Do character-level models capture morphology in terms of predictive utility?

Language	Addition	bi-LSTM
Czech	51.8	**30.1**
Russian	41.8	**26.4**

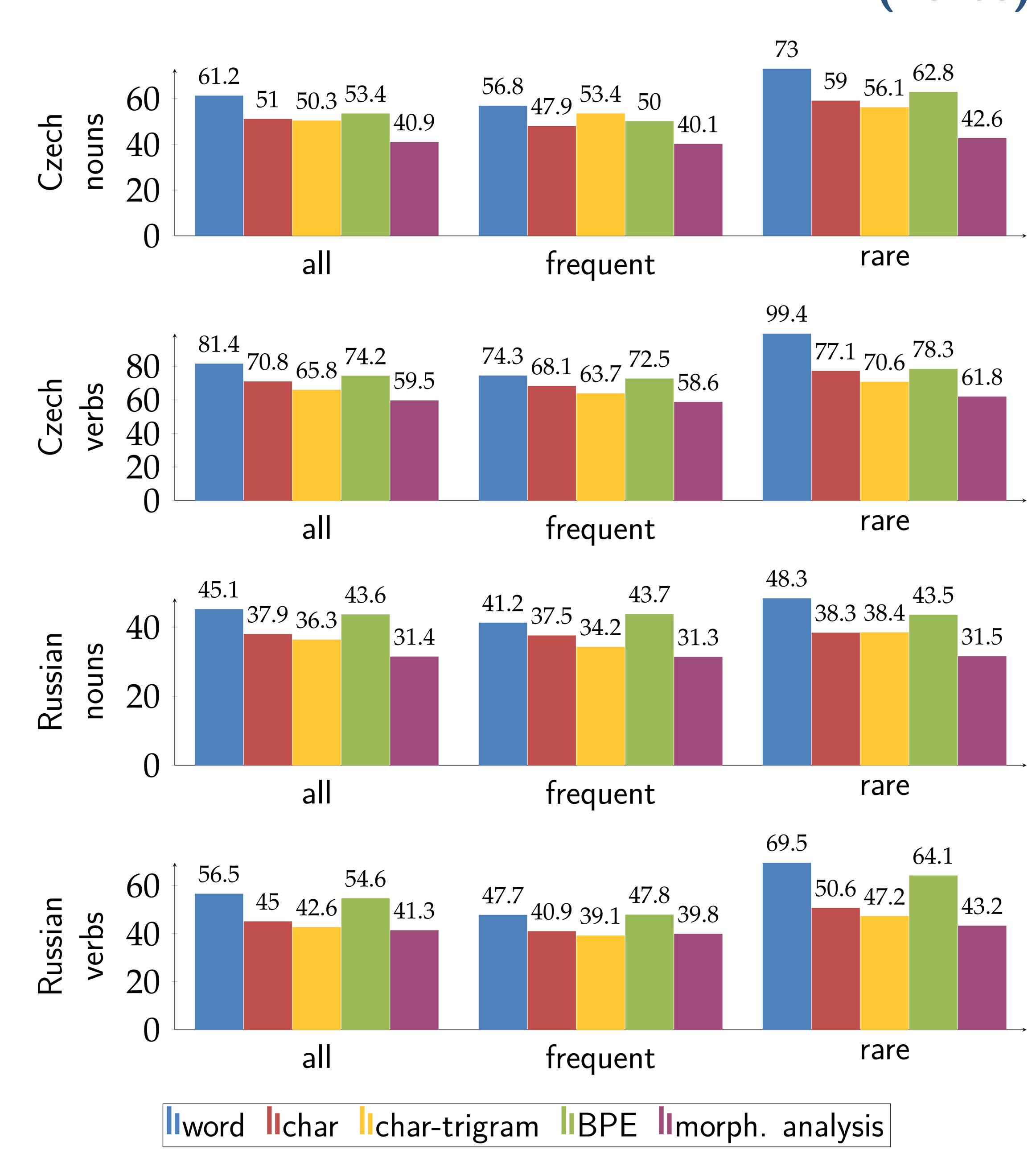
How much training data is needed to reach perplexity obtained using annotated data?



How do we know if these representations actually affect the predictions?

Analyze perplexities when the inflected words of interest are in the most recent history.

"The girl **loves dogs** but the boy **does not** ."
(verbs)

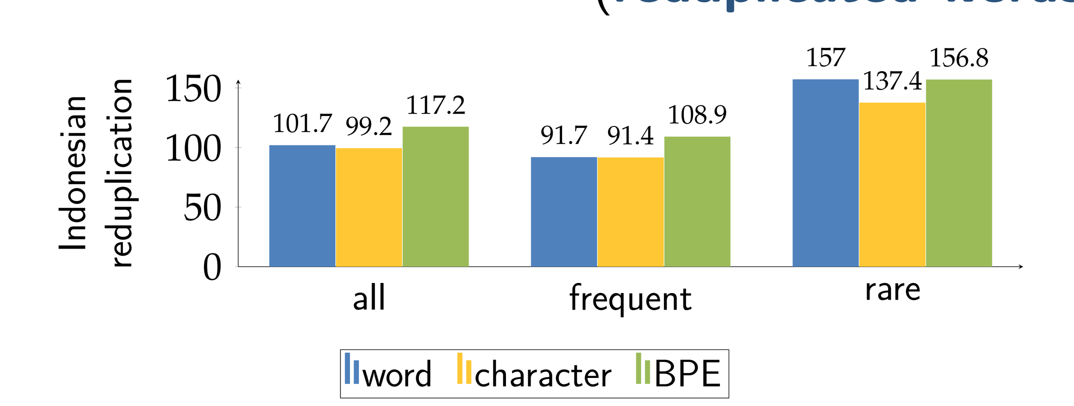


Which model most effectively captures reduplication?

Percentage of full reduplication on the training data:

Language	type-level (%)	token-level (%)
Indonesian	1.1	2.6
Malay	1.3	2.9

"Saya membeli **buku-buku itu** kemarin ."
*I bought **those books** yesterday .*
(**reduplicated words**)



Conclusion

- Character-level models are effective for many languages, but these models do not match the predictive accuracy of model with explicit knowledge of morphology.
- In this study, a previously unstudied combination of character trigram composed with bi-LSTM outperform most others.
- Our qualitative analysis suggests that they learn orthographic similarity of affixes.
- Other factors such as morphology and orthography affect the utility of these representations.