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

by Clara Vania and Adam Lopez

Peter Danenberg (danenberg@)

May 1, 2017

An update on mobile-first

TL;DR

1. Character-level models are better than word-level models, but not as good as morphological ones.
2. Good morphology is expensive!

Lets build a case for morphology.

Word-level embeddings might discover analogies like *cat* $\rightarrow$ *cats* $\cong$ *dog* $\rightarrow$ *dogs*, but not for out-of-vocabulary things like *sloth* $\rightarrow$ *sloths*.

Morphology is only as good as its segmentizer, though.

Modeling *cats* as e.g. *cat* and *-s* is potentially useful but expensive.

Character-based models are pretty good, too.

- They can capture related orthographic mutations (e.g. -*s* and -*es* in *finches*).
- They're cheap!

Let's compare!

Let's compare language models on the same datasets while varying the following parameters:

1. Subword unit
2. Composition function
3. Morphological typology

Results are in.

- Character-level embeddings outperform word-level ones.
- Bi-LSTMs and CNNs are more effective than addition.
- Character-level embeddings aren't as good as morphological ones.
- Character-level embeddings are limited by orthography.

Segmentation is different than analysis.

word	*tries*
morphemes	*try* + *s*
morphs	*tri* + *es*
analysis	*try* + VB + 3rd + SG + Pres

Fusional languages combine features in one morpheme (English).

$$\begin{aligned} wanted &\rightarrow want + ed \\ &\rightarrow want + VB + 1st + SG + Past \end{aligned}$$

Agglutinative languages have one feature per morpheme (Turkish).

$$\begin{aligned} \textit{okursam} &\rightarrow \textit{oku} + \textit{r} + \textit{sa} + \textit{m} \\ &\rightarrow \text{``}\textit{read}\text{''} + \textit{AOR} + \textit{COND} + 1\textit{SG} \end{aligned}$$

Root and pattern languages modify roots (Arabic).

ktb (“write”) $\rightarrow$ *katab* (“wrote”)

Reduplicative languages duplicate (Indonesian).

anak (“child”) $\rightarrow$ *anak-anak* (“children”)

Language models are differentiated by subword-generation and composition.

$$\mathbf{w} = f(\mathbf{W}_s, \sigma(w))$$

where σ returns subword units; $\mathbf{W}_s$ is a parameter matrix; and f is a composition function.

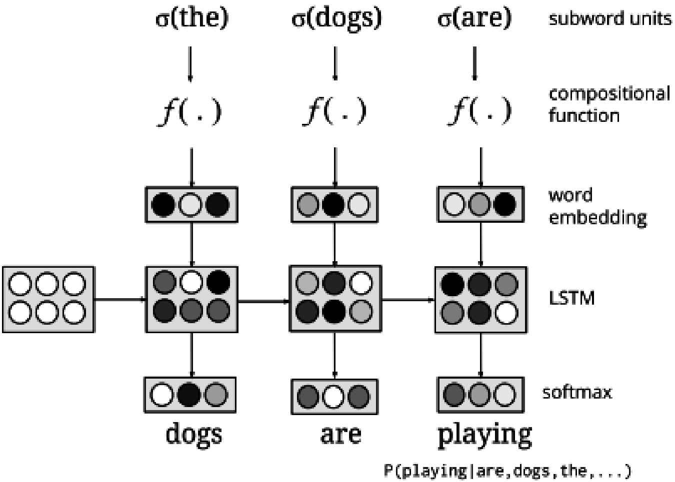
Subword units are four types.

- Character
- Character trigram
- Morfessor
- Byte-pair encoding

Composition functions are three types.

- Addition
- Bi-LSTM
- CNN

Language models are comparable using perplexity.



Results tend to favor trigram bi-LSTMs.

Language	word	character		char trigrams		BPE		Morfessor		%imp
		bi-lstm	CNN	add	bi-lstm	add	bi-lstm	add	bi-lstm	
Czech	41.46	34.25	36.60	42.73	**33.59**	49.96	33.74	47.74	36.87	18.98
English	46.40	43.53	44.67	45.41	**42.97**	47.51	43.30	49.72	49.72	7.39
Russian	34.93	28.44	29.47	35.15	**27.72**	40.10	28.52	39.60	31.31	20.64
Finnish	24.21	20.05	20.29	24.89	**18.62**	26.77	19.08	27.79	22.45	23.09
Japanese	98.14	98.14	**91.63**	101.99	101.09	126.53	96.80	111.97	99.23	6.63
Turkish	66.97	54.46	55.07	**50.07**	54.23	59.49	57.32	62.20	62.70	25.24
Arabic	48.20	42.02	43.17	50.85	**39.87**	50.85	42.79	52.88	45.46	17.28
Hebrew	38.23	31.63	33.19	39.67	**30.40**	44.15	32.91	44.94	34.28	20.48
Indonesian	46.07	45.47	46.60	58.51	45.96	59.17	**43.37**	59.33	44.86	5.86
Malay	54.67	53.01	**50.56**	68.51	50.74	68.99	51.21	68.20	52.50	7.52

Table 5: Language model perplexities on test. The best model for each language is highlighted in **bold** and the improvement of this model over the word-level model is shown in the final column.

How about with hand-annotated morphology?

Languages	Addition	bi-LSTM
Czech	51.8	**30.07**
Russian	41.82	**26.44**

What if we increase the amount of unannotated data?

#tokens	word	char trigram bi-LSTM	char CNN
1M	39.69	32.34	35.15
2M	37.59	36.44	35.58
3M	36.71	35.60	35.75
4M	35.89	32.68	35.93
5M	35.20	34.80	37.02
10M	35.60	35.82	39.09

What about automatic annotation?

Using MADAMIRA for Arabic, the perplexity of bi-LSTMs is still 42.85 vs. 39.87 with character trigrams.

Lastly, what if we restrict ourselves to nouns and verbs?

Inflection	Model	all	frequent	rare
Czech nouns	word	61.21	56.84	72.96
	characters	51.01	47.94	59.01
	char-trigrams	50.34	48.05	56.13
	BPE	53.38	49.96	62.81
	morph. analysis	**40.86**	**40.08**	**42.64**
Czech verbs	word	81.37	74.29	99.40
	characters	70.75	68.07	77.11
	char-trigrams	65.77	63.71	70.58
	BPE	74.18	72.45	78.25
	morph. analysis	**59.48**	**58.56**	**61.78**
Russian nouns	word	45.11	41.88	48.26
	characters	37.90	37.52	38.25
	char-trigrams	36.32	34.19	38.40
	BPE	43.57	43.67	43.47
	morph. analysis	**31.38**	**31.30**	**31.50**
Russian verbs	word	56.45	47.65	69.46
	characters	45.00	40.86	50.60
	char-trigrams	42.55	39.05	47.17
	BPE	54.58	47.81	64.12

Character-level models lose the meaning of root morphemes.

Model	Frequent Words			Rare Words		OOV words	
	man	*including*	*relatively*	*unconditional*	*hydroplane*	*uploading*	*foodism*
word	person anyone children men	like featuring include includes	extremely making very quite	nazi fairly joints supreme	molybdenum your imperial intervene	- - - -	- - - -
BPE LSTM	ii hill text netherlands	called involve like creating	newly never essentially least	unintentional ungenerous unanimous unpalatable	emphasize heartbeat hybridized unplatable	upbeat uprising handling hand-colored	vigilantism pyrethrum pausanias footway
char-trigrams LSTM	mak vill cow maga	include includes undermining under	resolutely regeneratively reproductively commonly	unconstitutional constitutional unimolecular medicinal	selenocysteine guerrillas scrofula seleucia	drifted affected conflicted convicted	tuaregs quft subjectivism tune-up
char-LSTM	mayr many mary may	inclusion insularity includes include	relates replicate relativity gravestones	undamaged unmyelinated unconditionally uncoordinated	hydrolyzed hydraulics hysterotomy hydraulic	musagte mutualism mutualists meursault	formulas formally fecal foreland
char-CNN	mtn mann jan nun	include includes excluding included	legislatively lovely creatively negatively	unconventional unintentional unconstitutional untraditional	hydroxyproline hydrate hydrangea hyena	unloading loading upgrading upholding	fordism dadaism popism endemism

Conclusion

- Might be some utility in semi-supervised learning from partially annotated data.