Enriching Word Vectors with Subword Information

Piotr Bojanowski and Edouard Grave and Armand Joulin and Tomas Mikolov - Facebook AI Research 2017

Louis (Yiqing) Luo

July 13th, 2018

Continuous Representation of Words

- Previously largely based on occurrence of words
- Recently employed FF-NN to learn embeddings.
- However, all methods ignore the internal structure of words.
  - Because many word formations follow rules, it is possible to improve vector representations for morphologically rich languages by using character level information. level information

Word2vec Skipgram (Mikolov et al 2013)

- Framed as binary classification problem with negative sampling
- maximizes score function s(w,c) between words and their context

$$\mathrm{s(w,c)} = w^T \mathrm{c}$$

- note that it ignores morphology of words

Subword Model

- Represent a word as bag of character n-grams

$$\text{skiiing} = \{\wedge\text{skiing\$}, \wedge\ \text{ski, skii, kiin, iing, ing\$}\}$$

- With G_w being the set of n-grams appearing in word w union with the original word w:

$$\text{s(w,c)} = \Sigma_{g \epsilon G_w} g^T c$$

Technical details

- n-grams between 3 and 6 characters used
- Hashing to map n-grams to integers
- SGD to min log-likelihood
- Subsampling of frequent words
- Less than 2 times slower than word2vec skipgram model

Experiments - Word Analogy (A is to B as C is to ?)

- All models are trained on Wikipedia data

		sg	cbow	sisg
Cs	Semantic	25.7	27.6	27.5
	Syntactic	52.8	55.0	77.8
De	Semantic	66.5	66.8	62.3
	Syntactic	44.5	45.0	56.4
En	Semantic	78.5	78.2	77.8
	Syntactic	70.1	69.9	74.9
It	Semantic	52.3	54.7	52.3
	Syntactic	51.5	51.8	62.7

Table 2: Accuracy of our model and baselines on word analogy tasks for Czech, German, English and Italian. We report results for semantic and syntactic analogies separately.

Comparison to State-of-the-Art

	DE		EN		ES	FR
	GUR350	ZG222	WS353	RW	WS353	RG65
Luong et al. (2013)	-	-	64	34	-	-
Qiu et al. (2014)	-	-	65	33	-	-
Soricut and Och (2015)	64	22	71	42	47	67
sisg	73	43	73	48	54	69
Botha and Blunsom (2014)	56	25	39	30	28	45
sisg	66	34	54	41	49	52

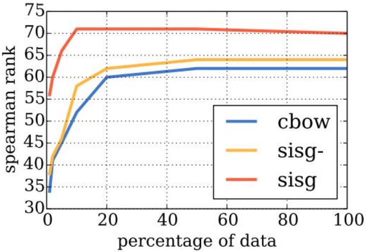
Table 3: Spearman's rank correlation coefficient between human judgement and model scores for different methods using morphology to learn word representations. We keep all the word pairs of the evaluation set and obtain representations for out-of-vocabulary words with our model by summing the vectors of character n-grams. Our model was trained on the same datasets as the methods we are comparing to (hence the two lines of results for our approach).

Experiments - Language Modeling

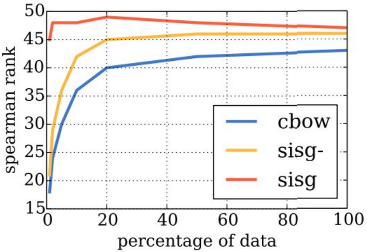
	Cs	De	Es	Fr	Ru
Vocab. size	46k	37k	27k	25k	63k
CLBL	465	296	200	225	304
CANLM	371	239	165	184	261
LSTM	366	222	157	173	262
sg	339	216	150	162	237
sisg	**312**	**206**	**145**	**159**	**206**

Table 5: Test perplexity on the language modeling task, for 5 different languages. We compare to two state of the art approaches: CLBL refers to the work of Botha and Blunsom (2014) and CANLM refers to the work of Kim et al. (2016).

Experiments - Size of Training Data (German and English)



(a) DE-GUR350



(b) EN-RW

Experiments - Size of n-gram

	2	3	4	5	6
2	57	64	67	69	69
3		65	68	70	70
4			70	70	**71**
5				69	**71**
6					70

(a) DE-GUR350

	2	3	4	5	6
2	59	55	56	59	60
3		60	58	60	62
4			62	62	63
5				64	64
6					**65**

(b) DE Semantic

	2	3	4	5	6
2	45	50	53	54	55
3		51	55	55	**56**
4			54	**56**	**56**
5				**56**	**56**
6					54

(c) DE Syntactic

	2	3	4	5	6
2	41	42	46	47	**48**
3		44	46	**48**	**48**
4			47	**48**	**48**
5				**48**	**48**
6					**48**

(d) EN-RW

	2	3	4	5	6
2	78	76	75	76	76
3		78	77	78	77
4			79	79	79
5				**80**	79
6					**80**

(e) EN Semantic

	2	3	4	5	6
2	70	71	73	74	73
3		72	74	**75**	74
4			74	**75**	**75**
5				74	74
6					72

(f) EN Syntactic

Table 4: Study of the effect of sizes of n-grams considered on performance. We compute word vectors by using character n-grams with n in $\{i, \ldots, j\}$ and report performance for various values of i and j. We evaluate this effect on German and English, and represent out-of-vocabulary words using subword information.

Discussion

- Possible to compute word vector for out-of-vocabulary words
- Learn better representation from small amount of data
- Short n-gram ($n = 4$) good to capture syntatic information
- Longer n-gram ($n=6$) good to capture semantic information

Experiments - Difference in Attention Paid to n-grams

	word	n-grams		
DE	autofahrer	fahr	fahrer	auto
	freundeskreis	kreis	kreis>	<freun
	grundwort	wort	wort>	grund
	sprachschule	schul	hschul	sprach
	tageslicht	licht	gesl	tages
EN	anarchy	chy	<anar	narchy
	monarchy	monarc	chy	<monar
	kindness	ness>	ness	kind
	politeness	polite	ness>	eness>
	unlucky	<un	cky>	nlucky
	lifetime	life	<life	time
	starfish	fish	fish>	star
	submarine	marine	sub	marin
	transform	trans	<trans	form
FR	finirais	ais>	nir	fini
	finissent	ent>	finiss	<finis
	finissions	ions>	finiss	sions>

Table 6: Illustration of most important character n-grams for selected words in three languages. For each word, we show the n-grams that, when removed, result in the most different representation.

Experiments - Examples

query	tiling	tech-rich	english-born	micromanaging	eateries	dendritic
`sisg`	tile	tech-dominated	british-born	micromanage	restaurants	dendrite
	flooring	tech-heavy	polish-born	micromanaged	eaterie	dendrites
`sg`	bookcases	technology-heavy	most-capped	defang	restaurants	epithelial
	built-ins	.ixic	ex-scotland	internalise	delis	p53

Table 7: Nearest neighbors of rare words using our representations and `skipgram`. These hand picked examples are for illustration.

Pretrained models

- available online at www.fasttext.cc