Boulder

Natural Language Processing; Word Representations

YouTube Playlist

Maziar Raissi

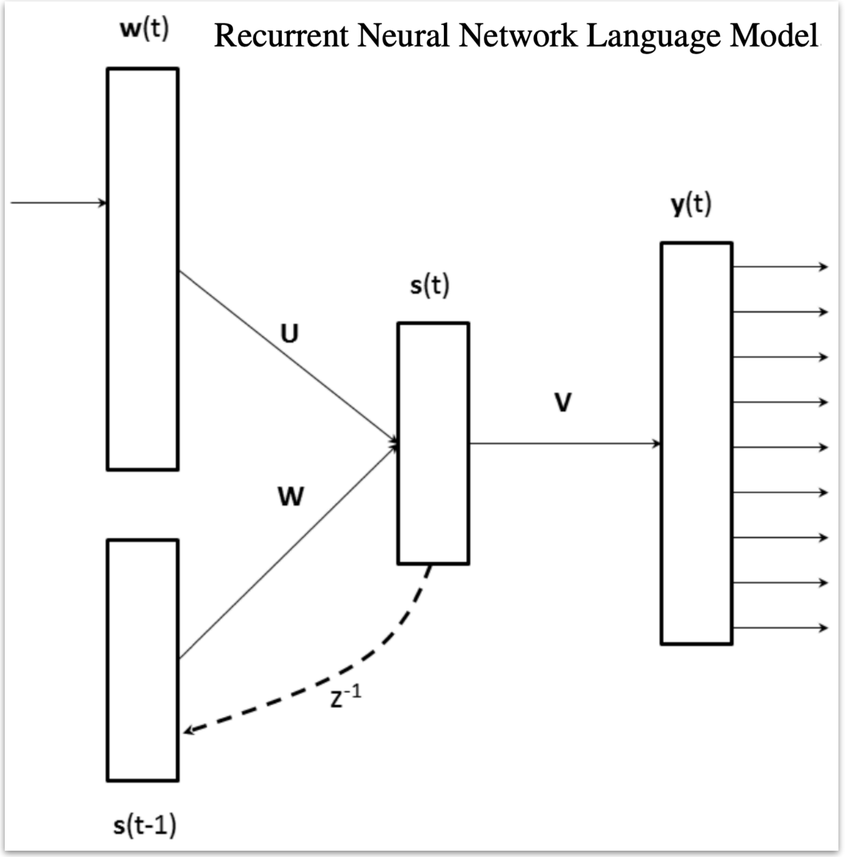
Assistant Professor

Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

Linguistic Regularities in Continuous Space Word Representations



$w(t) \rightarrow$ input vector (input word at time t encoded using 1-of-N coding)
$y(t) \rightarrow$ probability distribution over words
$s(t) \rightarrow$ hidden state maintaining a representation of the sentence history
$w(t)$ and $y(t)$ have dimensionality of the vocabulary

$$s(t) = f(Uw(t) + Ws(t-1))$$
$$y(t) = g(Vs(t))$$
$$f(z) = \frac{1}{1+e^{-z}}, g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

Each column in U represents a word

Measuring Linguistic Regularities (Syntactic & Semantic)

"a is to b as c is to ?"
"family is to families as car is to cars"
"closing is to shirt as dish is to bowl"

The Vector Offset Method

$$y = x_b - x_a + x_c$$
$$d = \arg\max_w \underbrace{\frac{x_w y}{\|x_w\|\|y\|}}_{=\cos(x_w, y)}$$

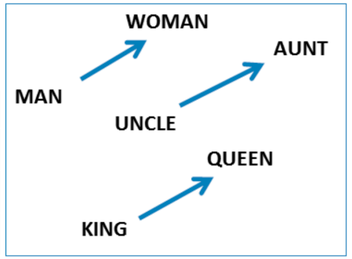
semantic regularities

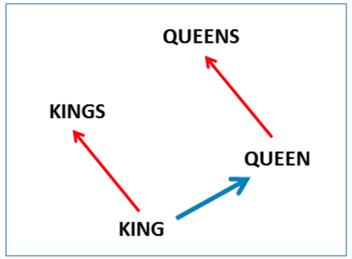
Method	Spearman's ρ	MaxDiff Acc.
LSA-640	0.149	0.364
RNN-80	0.211	0.389
RNN-320	0.259	0.408
RNN-640	0.270	0.416
RNN-1600	**0.275**	**0.418**

SemEval 2012 task of measuring relation similarity

syntactic regularities

Method	Adjectives	Nouns	Verbs	All
RNN-80	9.3	5.2	30.4	16.2
RNN-320	18.2	19.0	45.0	28.5
RNN-640	21.0	25.2	54.8	34.7
RNN-1600	**23.9**	**29.2**	**62.2**	**39.6**





Category	Relation	Patterns Tested	# Questions	Example
Adjectives	Base/Comparative	JJ/JJR, JJR/JJ	1000	good:better rough:___
Adjectives	Base/Superlative	JJ/JJS, JJS/JJ	1000	good:best rough:___
Adjectives	Comparative/ Superlative	JJS/JJR, JJR/JJS	1000	better:best rougher:___
Nouns	Singular/Plural	NN/NNS, NNS/NN	1000	year:years law:___
Nouns	Non-possessive/ Possessive	NN/NN_POS, NN_POS/NN	1000	city:city's bank:___
Verbs	Base/Past	VB/VBD, VBD/VB	1000	see:saw return:___
Verbs	Base/3rd Person Singular Present	VB/VBZ, VBZ/VB	1000	see:sees return:___
Verbs	Past/3rd Person Singular Present	VBD/VBZ, VBZ/VBD	1000	saw:sees returned:___

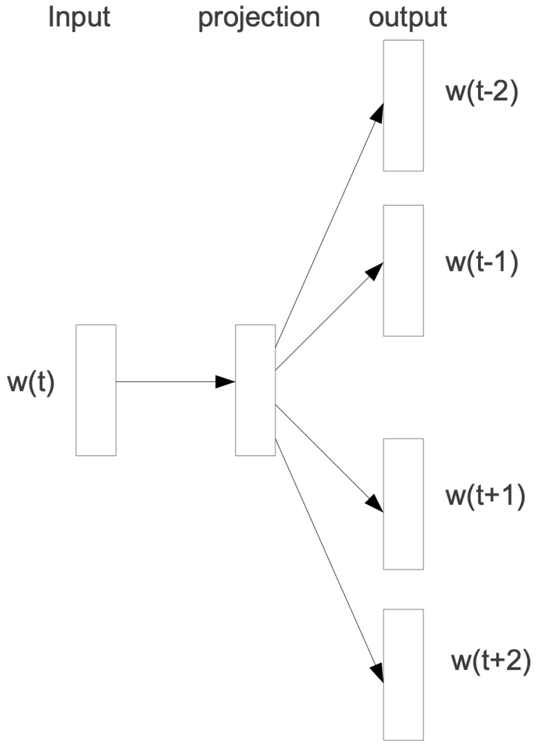
Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013.

Distributed Representations of Words and Phrases and their Compositionality

YouTube Playlist

Skip-gram Model

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t)$$



$$p(w_O|w_I)=\frac{\exp\left({v'_{w_O}}^{\top}v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\top}v_{w_I}\right)}$$

$w_1, w_2, \ldots, w_T \rightarrow$ training words
$c \rightarrow$ size of the training context
$w_I = w_t, w_O = w_{t+j}$
$W \rightarrow$ vocabulary size
$v_w \rightarrow$ input vector representation of w
$v'_w \rightarrow$ output vector representation of w
$W \approx 10^5 - 10^7$ terms

Learning Phrases

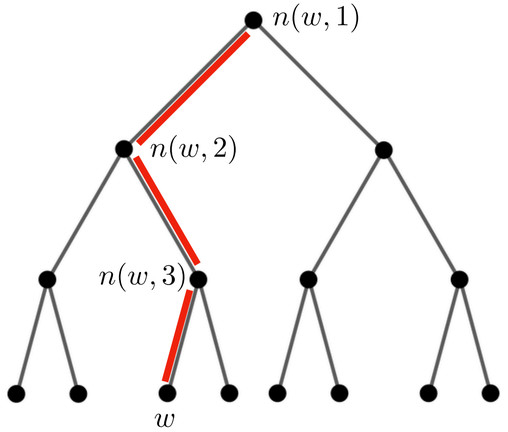
"New York Times"

$$\text{score}(w_i, w_j)=\frac{\text{count}(w_i w_j)-\delta}{\text{count}(w_i)\times\text{count}(w_j)}$$

$\delta \rightarrow$ discounting coefficient

Hierarchical Softmax

$\log_2(W)$ evaluations, rather than W



$n(w, j) \rightarrow j$th node on the path from the root to w
$L(w) \rightarrow$ length of this path
$n(w, 1) \rightarrow$ root
$n(w, L(w)) = w$
$\text{ch}(n) \rightarrow$ an arbitrary fixed child (e.g., left) of a node n
$[\![TRUE]\!] = 1, [\![FALSE]\!] = -1$

$$p(w|w_I)=\prod_{j=1}^{L(w)-1}\sigma\left([\![n(w,j+1)=\text{ch}(n(w,j))]\!]\cdot {v'_{n(w,j)}}^{\top}v_{w_I}\right)$$

$$p(w|w_I)=\sigma\left({v'_{n(w,1)}}^{T}v_{w_I}\right)\sigma\left(-{v'_{n(w,2)}}^{T}v_{w_I}\right)\sigma\left({v'_{n(w,3)}}^{T}v_{w_I}\right)$$

binary Huffman tree: assign short codes to the frequent words

Negative Sampling

Noise Contrastive Estimation (NCE): a good model should be able to differentiate data from noise by means of logistic regression

$$\log p(w_O|w_I)\approx\log\sigma({v'_{w_O}}^{\top}v_{w_I})+\sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}\left[\log\sigma(-{v'_{w_i}}^{\top}v_{w_I})\right]$$

noise distribution

$P_n(w) \propto U(w)^{3/4}$ unigram distribution

Countering the imbalance between rare and frequent words

each word w_i in the training set is discarded with probability

$$P(w_i)=1-\sqrt{\frac{t}{f(w_i)}}$$

$t = 10^{-5} \rightarrow$ threshold, $f \rightarrow$ frequency

Evaluation

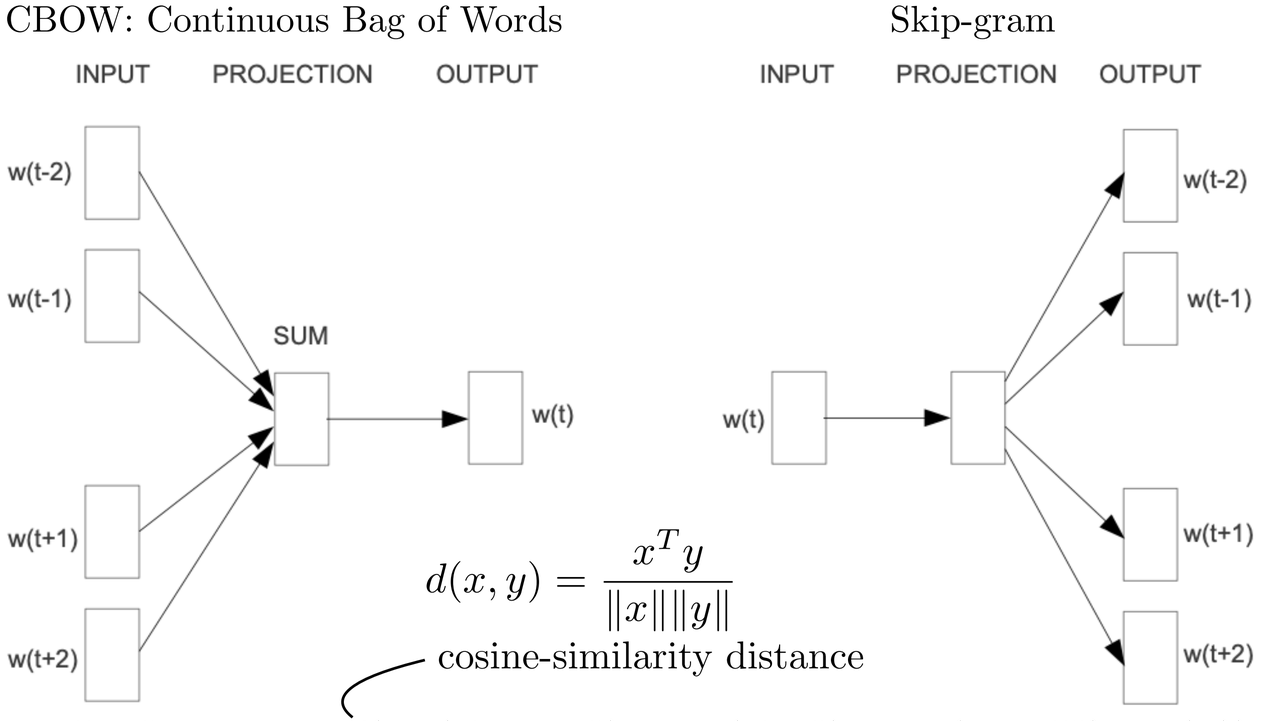
syntactic & semantic analogies

"Paris" $= \arg\min_w d(\text{vec}(\text{"Berlin"}) - \text{vec}(\text{"Germany"}) + \text{vec}(\text{"France"}), \text{vec}(w))$

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

Efficient Estimation of Word Representations in Vector Space

YouTube Video



$$d(x, y) = \frac{x^T y}{\|x\|\|y\|}$$

cosine-similarity distance

$$\text{“smallest”} = \arg\min_w d(\text{vec}(\text{“biggest”}) - \text{vec}(\text{“big”}) + \text{vec}(\text{“small”}), \text{vec}(w))$$

$$\text{“Paris”} = \arg\min_w d(\text{vec}(\text{“Berlin”}) - \text{vec}(\text{“Germany”}) + \text{vec}(\text{“France”}), \text{vec}(w))$$

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness
	Semantic Accuracy [%]	Syntactic Accuracy [%]	Test Set [20]
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

GloVe: Global Vectors for Word Representation

YouTube Video

1) Global matrix factorization methods
2) Local context window methods

$X \rightarrow$ matrix of word-word co-occurrence counts

$X_{ij} \rightarrow$ # times word j occurs in the context of word i

$\log(X_{ij}) = w_i^T \tilde{w}_j + b_i + \tilde{b}_j \rightarrow$ assumption

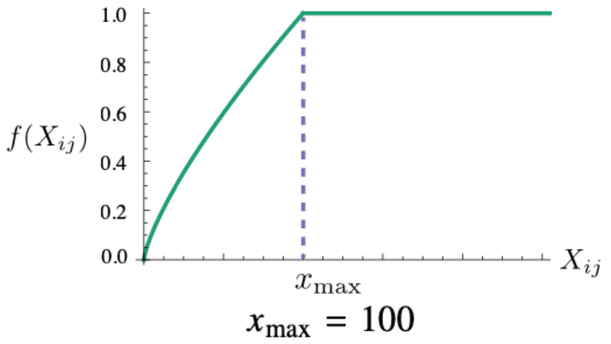
$w_i \in \mathbb{R}^d \rightarrow$ word vector

$\tilde{w}_j \in \mathbb{R}^d \rightarrow$ context word vector

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

weighted least squares

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \qquad \alpha = 3/4$$



$x_{\max} = 100$

Relationship to skip-gram

$$Q_{ij} = \frac{\exp(w_i^T \tilde{w}_j)}{\sum_{k=1}^{V} \exp(w_i^T \tilde{w}_k)} \rightarrow \text{probability that word } j \text{ appears in the context of word } i$$

$$J = -\sum_{\substack{i \in \text{corpus} \\ j \in \text{context}(i)}} \log Q_{ij}$$

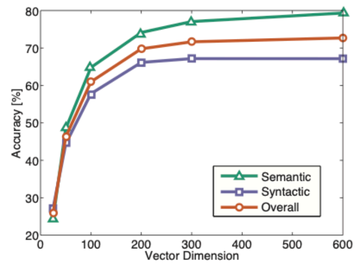
Model	Dim.	Size	Sem.	Syn.	Tot.
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7

$$J = -\sum_{i=1}^{V}\sum_{j=1}^{V} X_{ij} \log Q_{ij} \rightarrow \text{group together those terms that have the same values for } i \text{ and } j$$

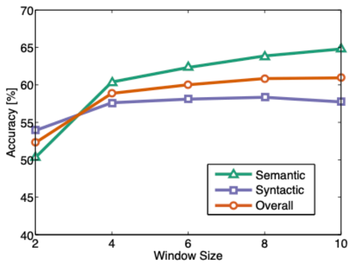
$$X_i = \sum_k X_{ik} \rightarrow \text{\# times any word occurs in the context of word } i$$

$$P_{ij} = \frac{X_{ij}}{X_i} \rightarrow \text{prob. that word } j \text{ occurs in the context of word } i$$

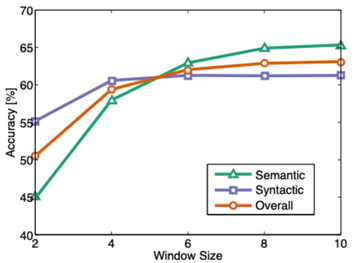
$$J = -\sum_{i=1}^{V} X_i \sum_{j=1}^{V} P_{ij} \log Q_{ij} = \sum_{i=1}^{V} X_i \underbrace{H(P_i, Q_i)}_{\text{cross-entropy}}$$



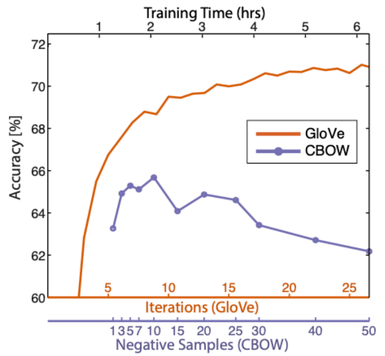
(a) Symmetric context



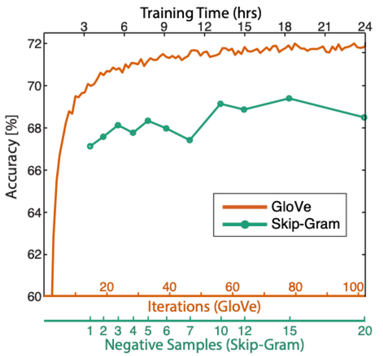
(b) Symmetric context



(c) Asymmetric context



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Enriching Word Vectors with Subword Information

YouTube Video

$w \in \{1, 2, \ldots, W\}$
size of the vocabulary

$w_1, w_2, \ldots, w_T \rightarrow$ a large training corpus

$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c \mid w_t)$$

context

$s : (w_t, w_c) \mapsto s(w_t, w_c) \in \mathbb{R}$
scoring function

$$p(w_c \mid w_t) = \frac{e^{s(w_t,\, w_c)}}{\sum_{j=1}^{W} e^{s(w_t,\, j)}}$$

presence (or absence) of context words

$$\log\left(1 + e^{-s(w_t,\, w_c)}\right) + \sum_{n \in \mathcal{N}_{t,c}} \log\left(1 + e^{s(w_t,\, n)}\right)$$

set of negative examples

$$\sum_{t=1}^{T} \left[\sum_{c \in \mathcal{C}_t} \ell(s(w_t,\, w_c)) + \sum_{n \in \mathcal{N}_{t,c}} \ell(-s(w_t,\, n)) \right]$$

$\ell : x \mapsto \log(1 + e^{-x}) \rightarrow$ logistic loss function

$s(w_t, w_c) = \mathbf{u}_{w_t}^{\top} \mathbf{v}_{w_c}$

$\mathbf{u}_w$ & $\mathbf{v}_w \in \mathbb{R}^d \rightarrow$ input & output vectors

Subword model

where and $n = 3$ character n-grams

<wh, whe, her, ere, re>

<where>

dictionary of n-grams of size G
$3 \leq n \leq 6$

$\mathcal{G}_w \subset \{1, \ldots, G\}$
set of n-grams appearing in w

$$s(w, c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^{\top} \mathbf{v}_c$$

$$\mathbf{u}_w = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g$$

Subword Information Skip Gram

		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

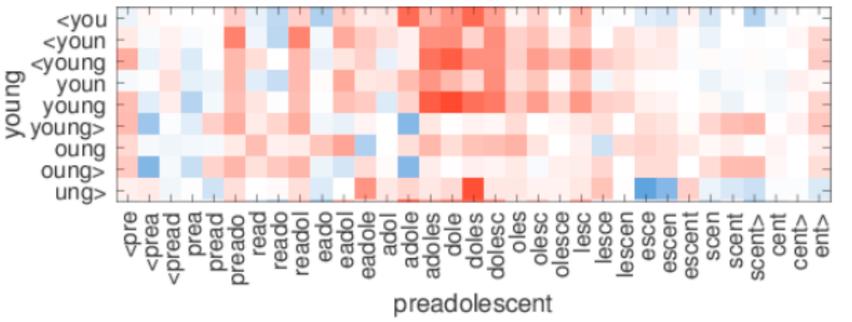
most important character n-grams

	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>

$$u_{w \setminus g} = \sum_{g' \in \mathcal{G} - \{g\}} z_{g'}$$

Shows the n-grams that, when removed, result in the most different representation

Red indicates positive cosine, while blue negative



Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5 (2017): 135-146.

Boulder

Questions?

YouTube Playlist