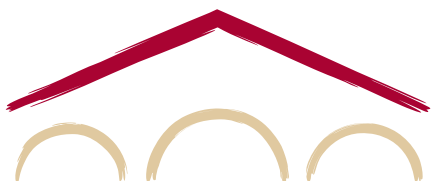
Distributed word representations: High-level goals and guiding hypotheses

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



Meaning latent in co-occurrence patterns

	:)	:/	:D	:|	;p	abandon	abc	ability	able	⋯
:)	74	1	0	0	0	1	0	2	2	
:/	1	306	0	0	0	0	0	0	17	
:D	0	0	16	0	0	0	6	1	1	
:|	0	0	0	120	0	0	0	1	9	
;p	0	0	0	0	516286	0	0	0	0	⋯
abandon	1	0	0	0	0	370	24	65	235	
abc	0	0	6	0	0	24	7948	77	291	
ability	2	0	1	1	0	65	77	4820	1807	
able	2	17	1	9	0	235	291	1807	14328	
⋮					⋮					

Meaning latent in co-occurrence patterns

Class	Word
neg	awful
neg	terrible
neg	lame
neg	worst
neg	disappointing
pos	nice
pos	amazing
pos	wonderful
pos	good
pos	awesome

A hopeless learning scenario

Meaning latent in co-occurrence patterns

Class	Word
neg	awful
neg	terrible
neg	lame
neg	worst
neg	disappointing
pos	nice
pos	amazing
pos	wonderful
pos	good
pos	awesome

$\Pr(\text{Class} = \text{pos})$	Word
?	w_1
?	w_2
?	w_3
?	w_4

A hopeless learning scenario

Meaning latent in co-occurrence patterns

Class	Word	excellent	terrible
neg	awful	6	113
neg	terrible	8	309
neg	lame	1	69
neg	worst	9	202
neg	disappointing	19	29
pos	nice	118	2
pos	amazing	91	6
pos	wonderful	66	7
pos	good	21	9
pos	awesome	67	2

A promising learning scenario

Meaning latent in co-occurrence patterns

Class	Word	excellent	terrible
neg	awful	6	113
neg	terrible	8	309
neg	lame	1	69
neg	worst	9	202
neg	disappointing	19	29
pos	nice	118	2
pos	amazing	91	6
pos	wonderful	66	7
pos	good	21	9
pos	awesome	67	2

Pr(Class=pos)	Word	excellent	terrible
≈0	w_1	4	82
≈0	w_2	5	84
≈1	w_3	49	3
≈1	w_4	41	5

A promising learning scenario

High-level goals

1. Begin thinking about how vectors can encode the meanings of linguistic units.

2. Foundational concepts for vector-space model (VSMs) a.k.a. embeddings.

3. A foundation for deep learning NLU models.

Guiding hypotheses

Firth (1957)

“You shall know a word by the company it keeps.”

Harris (1954)

“distributional statements can cover all of the material of a language without requiring support from other types of information.”

Wittgenstein (1953)

“the meaning of a word is its use in the language”

Turney and Pantel (2010)

“If units of text have similar vectors in a text frequency matrix, then they tend to have similar meanings.”

Great power, a great many design choices

tokenization
annotation
tagging
parsing
feature selection
⋮ cluster texts by date/author/discourse context/. . .
⇓ ⇙

Matrix design	Reweighting	Dimensionality reduction	Vector comparison
word × document	probabilities	LSA	Euclidean
word × word	length norm.	PLSA	Cosine
word × search proximity	TF-IDF	LDA	Dice
adj. × modified noun	PMI	PCA	Jaccard
word × dependency rel.	Positive PMI	NNMF	KL
⋮	⋮	⋮	⋮

Nearly the full cross-product to explore; only a handful of the combinations are ruled out mathematically. Models like GloVe and word2vec offer packaged solutions to design/weighting/reduction and reduce the importance of the choice of comparison method. Contextual embeddings dictate many preprocessing choices.

References I

John R. Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. The MacMillan Company, New York. Translated by G. E. M. Anscombe.