Distributed word representations: overview

Chris Potts
Stanford Linguistics

CS 244U: Natural language understanding

A typical starting point

Upper left corner of a matrix derived from the training portion of this IMDB data release: `http://ai.stanford.edu/~amaas/data/sentiment/`.

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

Guiding hypotheses

Guiding hypotheses

John Rupert Firth, (1957, ‘A synopsis of linguistic theory’)

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

Guiding hypotheses

John Rupert Firth, (1957, ‘A synopsis of linguistic theory’)

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

Firth (1957, ‘A synopsis of linguistic theory’)

“the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously.”

Guiding hypotheses

John Rupert Firth, (1957, 'A synopsis of linguistic theory')

"You shall know a word by the company it keeps."

Firth (1957, 'A synopsis of linguistic theory')

"the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously."

Zellig Harris (1954, 'Distributional structure')

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

Guiding hypotheses

John Rupert Firth, (1957, ‘A synopsis of linguistic theory’)

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

Firth (1957, ‘A synopsis of linguistic theory’)

“the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously.”

Zellig Harris (1954, ‘Distributional structure’)

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

Turney and Pantel (2010, ‘From frequency to meaning’)

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

General questions for vector-space modelers

For a word $\times$ document matrix:

- How do the rows relate to each other?
- How do the columns relate to each other?
- For a given group of documents D, which words epitomize D?
- For a given a group of words W, which documents epitomize W (IR)?

Some matrix designs

Word × document

Upper left corner of a matrix derived from the training portion of this IMDB data release: `http://ai.stanford.edu/~amaas/data/sentiment/`.

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

Word × word

Upper left corner of a matrix derived from the training portion of this IMDB data release: http://ai.stanford.edu/~amaas/data/sentiment/.

	against	age	agent	ages	ago	agree	ahead	ain't	air	aka	al
against	2003	90	39	20	88	57	33	15	58	22	24
age	90	1492	14	39	71	38	12	4	18	4	39
agent	39	14	507	2	21	5	10	3	9	8	25
ages	20	39	2	290	32	5	4	3	6	1	6
ago	88	71	21	32	1164	37	25	11	34	11	38
agree	57	38	5	5	37	627	12	2	16	19	14
ahead	33	12	10	4	25	12	429	4	12	10	7
ain't	15	4	3	3	11	2	4	166	0	3	3
air	58	18	9	6	34	16	12	0	746	5	11
aka	22	4	8	1	11	19	10	3	5	261	9
al	24	39	25	6	38	14	7	3	11	9	861

Word × discourse context

Upper left corner of an interjection × dialog-act tag matrix derived from the Switchboard Dialog Act Corpus:

	%	+	^2	^g	^h	^q	aa
absolutely	0	2	0	0	0	0	95
actually	17	12	0	0	1	0	4
anyway	23	14	0	0	0	0	0
boy	5	3	1	0	5	2	1
bye	0	1	0	0	0	0	0
bye-bye	0	0	0	0	0	0	0
dear	0	0	0	0	1	0	0
definitely	0	2	0	0	0	0	56
exactly	2	6	1	0	0	0	294
gee	0	3	0	0	2	1	1
goodness	1	0	0	0	2	0	0

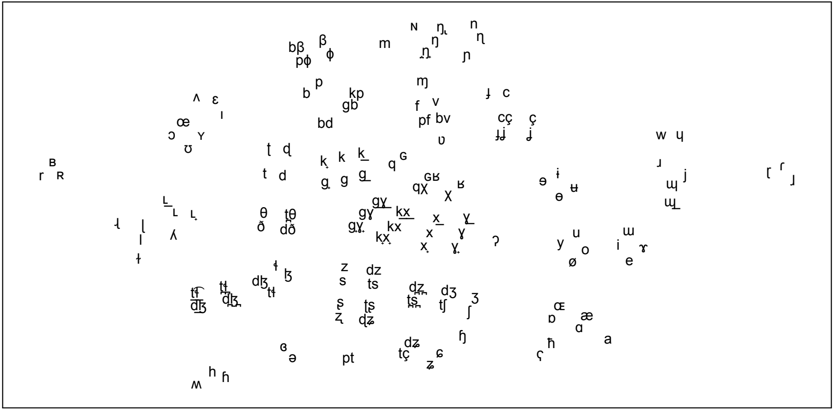
Phonological segment × feature values

Derived from http://www.linguistics.ucla.edu/people/hayes/120a/.
Dimensions: (141 × 28).

	syllabic	stress	long	consonantal	sonorant	continuant	delayed.release	approximant	tap	trill	…
ɒ	1	−1	−1	−1	1	1	0	1	−1	−1	
ɑ	1	−1	−1	−1	1	1	0	1	−1	−1	
ɶ	1	−1	−1	−1	1	1	0	1	−1	−1	
a	1	−1	−1	−1	1	1	0	1	−1	−1	
æ	1	−1	−1	−1	1	1	0	1	−1	−1	…
ʌ	1	−1	−1	−1	1	1	0	1	−1	−1	
ɔ	1	−1	−1	−1	1	1	0	1	−1	−1	
o	1	−1	−1	−1	1	1	0	1	−1	−1	
ɤ	1	−1	−1	−1	1	1	0	1	−1	−1	
ə	1	−1	−1	−1	1	1	0	1	−1	−1	
⋮					⋮						

Phonological segment × feature values

Derived from http://www.linguistics.ucla.edu/people/hayes/120a/.
Dimensions: (141 × 28).



Other designs

- word $\times$ dependency rel.
- word $\times$ syntactic context
- adj. $\times$ modified noun
- word $\times$ search query
- person $\times$ product
- word $\times$ person
- word $\times$ word $\times$ pattern
- verb $\times$ subject $\times$ object

$\vdots$

Great power, a great many design choices

Great power, a great many design choices

Matrix type
word $\times$ document
word $\times$ word
word $\times$ search proximity
adj. $\times$ modified noun
word $\times$ dependency rel.
$\vdots$

Great power, a great many design choices

tokenization
annotation
tagging
parsing
feature selection
$\vdots$ cluster texts by date/author/discourse context/...
$\Downarrow$ $\Swarrow$

Matrix type
word $\times$ document
word $\times$ word
word $\times$ search proximity
adj. $\times$ modified noun
word $\times$ dependency rel.
$\vdots$

Great power, a great many design choices

tokenization
annotation
tagging
parsing
feature selection
$\vdots$ cluster texts by date/author/discourse context/...
$\Downarrow$ $\swarrow$

Matrix type	Reweighting
word $\times$ document	probabilities
word $\times$ word	length normalization
word $\times$ search proximity	TF-IDF
adj. $\times$ modified noun	PMI
word $\times$ dependency rel.	Positive PMI
$\vdots$	$\vdots$

Great power, a great many design choices

tokenization
annotation
tagging
parsing
feature selection
⋮ cluster texts by date/author/discourse context/…
$\Downarrow$ $\swarrow$

Matrix type	Reweighting	Dimensionality reduction
word × document	probabilities	LSA
word × word	length normalization	PLSA
word × search proximity	TF-IDF	LDA
adj. × modified noun	PMI	PCA
word × dependency rel.	Positive PMI	IS
⋮	⋮	⋮

Great power, a great many design choices

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

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

Great power, a great many design choices

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

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

(Nearly the full cross-product to explore; only a handful of the combinations are ruled out mathematically, and the literature contains relatively little guidance.)