Software-Praktikum: Software-Technologien für Natürlichsprachliche Systeme

Word Embeddings

Sven Büchel

Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-University Jena,
Jena, Germany

https://julielab.de

Counting Cooccurrences

- *Cooccurrence*: tokens appearing together in a corpus within a window of pre-determined size

He reads a poem .

Susanne reads a novel .

The novel has 100 pages .

Her poem has 3 pages .

Susanne listens to an opera .

Peter listens to a song .

The song is in D-minor .

The opera is in D-minor .

Counting Cooccurrences

- *Cooccurrence*: tokens appearing together in a corpus within a window of pre-determined size

He reads a poem .

Susanne reads a novel .

The novel has 100 pages .

Her poem has 3 pages .

Susanne listens to an opera .

Peter listens to a song .

The song is in D-minor .

The opera is in D-minor .

Counting Cooccurrences

- *Cooccurrence*: tokens appearing together in a corpus within a window of pre-determined size

He reads a poem .
Susanne reads a novel .
The novel has 100 pages .
Her poem has 3 pages .
Susanne listens to an opera .
Peter listens to a song .
The song is in D-minor .
The opera is in D-minor .

Counting Cooccurrences

- ***Cooccurrence***: tokens appearing together in a corpus within a window of pre-determined size

He reads a poem .
Susanne reads a novel .
The novel has 100 pages .
Her poem has 3 pages .
Susanne listens to an opera .
Peter listens to a song .
The song is in D-minor .
The opera is in D-minor .

Counting Cooccurrences

- *Cooccurrence*: tokens appearing together in a corpus within a window of pre-determined size

He reads a poem .
Susanne reads a novel .
The novel has 100 pages .
Her poem has 3 pages .
Susanne listens to an opera .
Peter listens to a song .
The song is in D-minor .
The opera is in D-minor .

Cooccurrence Matrix — Raw Frequency

	read	pages	...	listen
novel	98	60	...	2
poem	67	10	...	8
...	...	...	...	...
opera	4	8	...	38

Cooccurrence Matrix — Raw Frequency

Adding marginal frequencies

	read	pages	…	listen	Σ
novel	98	60	…	2	172
poem	67	10	…	8	90
…	…	…	…	…	…
opera	4	8	…	38	166
Σ	199	229		199	**2461**

total number of cooccurrences

Cooccurrence Matrix — Relative Frequency

Divide every cell by total number of cooccurrences

	read	pages	…	listen	***P(w₁)***
novel	.049	.024	…	.001	.070
poem	.027	.004	…	.003	.037
…	…	…	…	…	…
opera	.002	.003	…	.015	.067
$P(w_2)$	.081	.093		.077	**1**

Relative frequency can be usede to estimate occurrence probability *P(w)*

Pointwise Mutual Information (PMI) Matrix

Compute PMI for each cell:

Statistical Association: How much more often than chance do 2 words cooccur?

PMI	read	pages	...	listen	$P(w_1)$
novel	0.94	0.57	...	-0.73	.070
poem	0.95	0.07	...	0.02	.037
...	...	...	...	...	...
opera	-0.43	-0.32	...	0.46	.067
$P(w_2)$	.081	.093	...	.077	1

$$PMI(w_1, w_2) := log \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)}$$

PMI for Word Meaning Analysis: *Gay* — 1900s

SEC 1 (1900): 22,097,593 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	FLOWERS	13	0	0.6	0.0	58.8
2	LAUGH	12	0	0.5	0.0	54.3
3	BRIGHT	12	0	0.5	0.0	54.3
4	GLAD	10	0	0.5	0.0	45.3
5	PARIS	9	0	0.4	0.0	40.7
6	SMILE	9	0	0.4	0.0	40.7
7	HAPPY	8	0	0.4	0.0	36.2
8	THRONG	8	0	0.4	0.0	36.2
9	GIRL	7	0	0.3	0.0	31.7
10	LAUGHTER	6	0	0.3	0.0	27.2

https://corpus.byu.edu/coha/

- compute PMI of target word with every other word
- rank cooccurring words in descending order of PMI
- manual inspection of most strongly associated words

PMI for Word Meaning Analysis: *Gay* — 2000s

SEC 2 (2000): 29,567,390 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	MARRIAGE	81	0	2.7	0.0	274.0
2	RIGHTS	57	0	1.9	0.0	192.8
3	COMMUNITY	32	0	1.1	0.0	108.2
4	BECAUSE	19	0	0.6	0.0	64.3
5	ALSO	18	0	0.6	0.0	60.9
6	LESBIAN	78	1	2.6	0.0	58.3
7	LESBIANS	16	0	0.5	0.0	54.1
8	ABORTION	14	0	0.5	0.0	47.3
9	BISEXUAL	14	0	0.5	0.0	47.3
10	ISSUES	12	0	0.4	0.0	40.6

https://corpus.byu.edu/coha/

- compute PMI of target word with every other word
- rank cooccurring words in descending order of PMI
- manual inspection of most strongly associated words

Awful — 1810s...1850s vs. 2000s

Corpus of Historical American English

SEARCH | FREQUENCY | CONTEXT | ACCOUNT

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION) [HELP...]

SEC 1 (1820, 1830, 1840, 1850, 1810): 54,403,008 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	UPON	64	0	1.2	0.0	117.6
2	STILLNESS	35	0	0.6	0.0	64.3
3	PRESENCE	29	0	0.5	0.0	53.3
4	SHALL	27	0	0.5	0.0	49.6
5	DOOM	25	0	0.5	0.0	46.0
6	SOLEMN	22	0	0.4	0.0	40.4
7	MAJESTY	21	0	0.4	0.0	38.6
8	MYSTERIOUS	21	0	0.4	0.0	38.6
9	WHOSE	21	0	0.4	0.0	38.6
10	SOUL	20	0	0.4	0.0	36.8

SEC 2 (2000): 29,567,390 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	LOT	71	0	2.4	0.0	240.1
2	HAPPENED	10	0	0.3	0.0	33.8
3	GRANDMOTHER	9	0	0.3	0.0	30.4
4	SMELL	7	0	0.2	0.0	23.7
5	MEAN	6	0	0.2	0.0	20.3
6	PRETTY	6	0	0.2	0.0	20.3
7	'RE	11	1	0.4	0.0	20.2
8	HAPPEN	5	0	0.2	0.0	16.9
9	PROBABLY	5	0	0.2	0.0	16.9
10	TASTED	5	0	0.2	0.0	16.9

Cell — 1850s vs. 2000s

Corpus of Historical American English

SEARCH | **FREQUENCY** | CONTEXT | OVERVIEW

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION) [HELP...]

SEC 1 (1850): 16,471,649 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	QUEEN	9	0	0.5	0.0	54.6
2	CONDEMNED	7	0	0.4	0.0	42.5
3	THY	6	0	0.4	0.0	36.4
4	HONEY	6	0	0.4	0.0	36.4
5	BEES	6	0	0.4	0.0	36.4
6	YOUNG	6	0	0.4	0.0	36.4
7	ROYAL	5	0	0.3	0.0	30.4
8	HERMIT	5	0	0.3	0.0	30.4
9	GLOOMY	5	0	0.3	0.0	30.4
10	FELON	7	1	0.4	0.0	12.6

SEC 2 (2000): 29,567,390 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PHONE	917	0	31.0	0.0	3,101.4
2	PHONES	258	0	8.7	0.0	872.6
3	STEM	97	0	3.3	0.0	328.1
4	-	71	0	2.4	0.0	240.1
5	YOU	65	0	2.2	0.0	219.8
6	)	43	0	1.5	0.0	145.4
7	RESEARCH	43	0	1.5	0.0	145.4
8	N'T	42	0	1.4	0.0	142.0
9	TALKING	40	0	1.4	0.0	135.3
10	RANG	36	0	1.2	0.0	121.8

Cell — 1850s vs. 2000s

Corpus of Historical American English

SEARCH | FREQUENCY | CONTEXT | OVERVIEW

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION) [HELP...]

SEC 1 (1850): 16,471,649 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	QUEEN	9	0	0.5	0.0	54.6
2	CONDEMNED	7	0	0.4	0.0	42.5
3	THY	6	0	0.4	0.0	36.4
4	HONEY	6	0	0.4	0.0	36.4
5	BEES	6	0	0.4	0.0	36.4
6	YOUNG	6	0	0.4	0.0	36.4
7	ROYAL	5	0	0.3	0.0	30.4
8	HERMIT	5	0	0.3	0.0	30.4
9	GLOOMY	5	0	0.3	0.0	30.4
10	FELON	7	1	0.4	0.0	12.6

SEC 2 (2000): 29,567,390 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PHONE	917	0	31.0	0.0	3,101.4
2	PHONES	258	0	8.7	0.0	872.6
3	STEM	97	0	3.3	0.0	328.1
4	-	71	0	2.4	0.0	240.1
5	YOU	65	0	2.2	0.0	219.8
6	)	43	0	1.5	0.0	145.4
7	RESEARCH	43	0	1.5	0.0	145.4
8	N'T	42	0	1.4	0.0	142.0
9	TALKING	40	0	1.4	0.0	135.3
10	RANG	36	0	1.2	0.0	121.8

Downside: statistical association does not necessarily entail a semantic relationship

Cell — 1850s

Corpus of Historical American English

SEARCH | FREQUENCY | CONTEXT

SECTION: 1850 **(9)** (SHUFFLE)

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?]

1	1853	NF	LangstrothOnHive	A	B	C	proceed as follows: With a very sharp knife, carefully cut out a **queen** **cell**, on a piece of comb an inch or
2	1853	NF	MysteriesBee-keeping	A	B	C	ADVANTAGES OF THIS METHOD. It is very plain that a **queen** from such finished **cell** must be ready to d
3	1853	NF	MysteriesBee-keeping	A	B	C	then, until further evidence contradicts it, that the first perfect **queen** leaving her **cell**, makes it her busi
4	1853	NF	MysteriesBee-keeping	A	B	C	pieces by the time the bee gets out. The covering to the **queen**'s **cell** is like the drone's, but larger in dia
5	1853	NF	MysteriesBee-keeping	A	B	C	of growth, as well as the eggs. Fig. 1 represents a **queen**'s **cell** just commenced. They are usually started
6	1853	NF	MysteriesBee-keeping	A	B	C	and removed by the workers. It will be perceived that each finished **queen**'s **cell** contains as much wax
7	1853	NF	MysteriesBee-keeping	A	B	C	foregoing conditions of the stock may require their use). STATE OF **QUEEN**'S **CELL** WHEN USED. They are
8	1853	NF	MysteriesBee-keeping	A	B	C	one of these methods could be relied upon. Instead of constructing a **queen**'s **cell**, and then removing t
9	1853	NF	MysteriesBee-keeping	A	B	C	the fact, that a few times I have found a quantity remaining in the **cell** after the **queen** had left. The con

PMI Matrix

PMI	read	pages	...	listen	$P(w_1)$
novel	0.94	0.57	...	-0.73	.070
poem	0.95	0.07	...	0.02	.037
...	...	...	...	...	...
opera	-0.43	-0.32	...	0.46	.067
$P(w_2)$	.081	.093	...	.077	1

$$PMI(w_1, w_2) := log \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)}$$

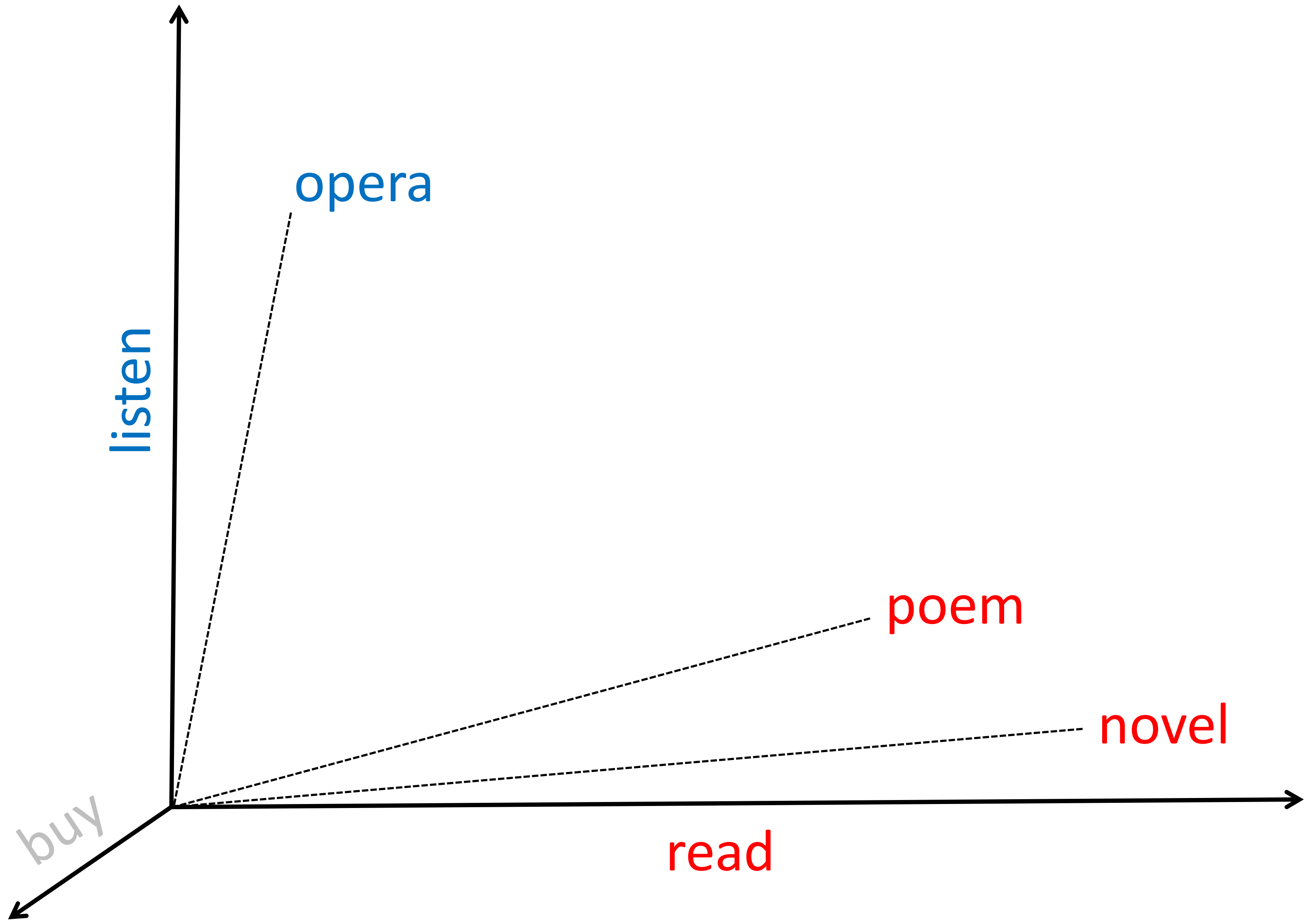
PMI Matrix

PMI	read	pages	...	listen
novel	0.94	0.57	...	-0.73
poem	0.95	0.07	...	0.02
...	...	...	...	...
opera	-0.43	-0.32	...	0.46

PMI Matrix

PMI	read	pages	...	listen
novel	0.94	0.57	...	-0.73
poem	0.95	0.07	...	0.02
...	...	...	...	...
opera	-0.43	-0.32	...	0.46

Vector Space Interpretation of PMI Matrix



Vector Space Interpretation of PMI Matrix

Cosine between vectors encodes semantic similarity

$$\cos(a,b) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{a_1 \times b_1 + a_2 \times b_2 + \ \ldots}{\sqrt{a_1^2 + a_2^2 + \cdots} \times \sqrt{b_1^2 + b_2^2 + \ \ldots}}$$

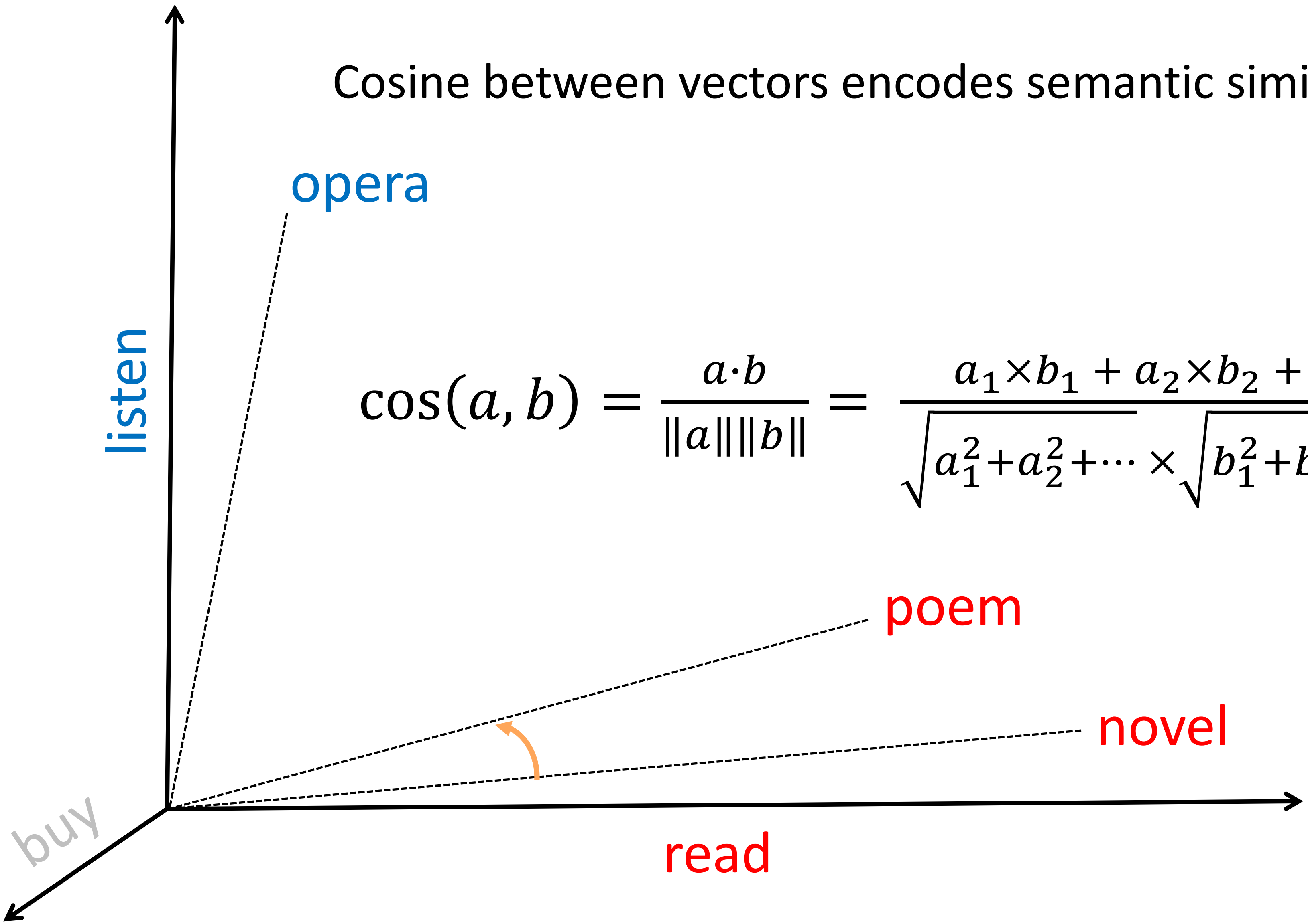
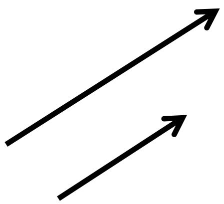
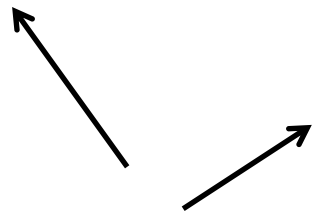


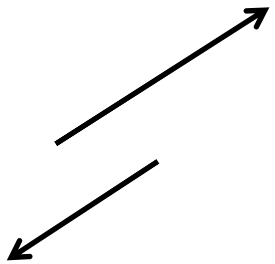
Illustration of Cosine Similarity



cos(a,b) = 1



cos(a,b) = 0



cos(a,b) = -1

PMI	read	pages	listen
novel	0.94	0.57	-0.73
poem	0.95	0.07	0.02

$$\cos(novel, poem) = \frac{.94 \times .95 + \ .57 \times .07 \ - \ .73 \times .02}{\sqrt{.94^2 + .57^2 + .73^2} \times \sqrt{.95^2 + .07^2 + .02^2}} = \mathbf{.73}$$

Statistical Association vs. Semantic Similarity

- „Shallow“ text feature vs. „deep“ text feature
- Syntagmatic vs. paradigmatic
- Examples:

associated:	*mashed*	*potatoes*
similar:	*potatoes*	*fries*
associated and similar:	*potato*	*salad*
neither:	*potato*	*transcendent*

Empirically Validation: Word Similarity Lists

Word 1	Word 2	Similarity
Love	Sex	6.77
Tiger	Cat	7.35
Tiger	Tiger	10.00
Book	Paper	7.46

Example entries WordSim353 (Finkelstein et al., 2002)

- Ask 20 people how similar two words are on 1-to-10 scale
- Average responses for each word (“ground truth”)
- Compute similarity of word pairs with cosine
- PMI vectors agree more with ground truth than two human raters with each other (Levy et al., TACL 2015)

Curse of Dimensionality

	read	pages	buy	eat	listen	…
novel	98	60	3	0	2	
poem	67	10	1	0	8	
opera	4	8	0	0	38	
⋮						

Typically 100,000 x 100,000 words

= 10 billion combinations!

Compressing the PMI Matrix

Singular Value Decomposition: Mathematical procedure allowing to reduce the number of columns of PMI matrix

	Opaque Dimension 1	Opaque Dimension 2	Opaque Dimension 3	…
novel	0.5	0.1	0.2	
poem	0.3	0.0	0.3	
opera	0.1	-0.1	0.5	
⋮				

Typically 100,000 words x 300 dimensions
= 30 million combinations!

- ***Word Embeddings***: Dense (no or few zeros), low dimensional (50-1000) vector represenations of words

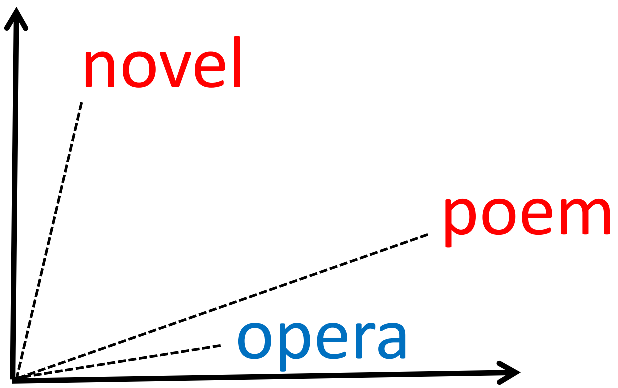
-0.13102 -0.054447 -0.051866 -0.10289 -0.072061 0.16523 -0.17298 0.21865 0.041183 -0.010858 0.074741 0.35226
0.42662 -0.071747 0.25112 0.12082 -0.33192 -0.4728 -0.0090568 0.0030266 0.032861 0.074323 -0.38017 0.091399
-0.16034 -0.050232 -0.094194 0.16656 0.40901 0.069625 0.059306 0.01991 -0.35846 -0.14549 0.24894 0.50184 -
0.0073098 -0.4589 -0.10073 -0.099315 0.30583 -0.40577 0.16586 0.055741 0.26776 -0.13515 0.28127 0.069221 -
0.20907 0.092053 0.39419 -0.2412 0.01173 -0.16856 -0.0053851 0.14282 0.17513 0.34775 0.178 0.35883 -0.17684
0.53104 0.04751 -0.30134 -0.53297 -0.22041 0.097703 0.052288 0.10849 0.12409 -0.11369 0.19042 0.19554 -
0.14949 -0.29675 -0.14285 0.22217 0.21503 -0.2309 0.4381 0.22739 -0.052386 -0.20003 0.19725 -0.032432 -
0.14307 0.021958 0.36876 -0.10084 -0.18536 0.27691 -0.43856 0.087418 -0.33836 0.083161 -0.40672 0.14497 -
0.41334 0.0012195 -0.32266 0.067225 0.18359 0.010442 -0.15499 -0.82943 -0.069867 -0.26416 0.42656 0.26765 -
0.12262 -0.116 -0.076926 -0.16992 0.055428 -0.20699 -0.090381 0.082171 -0.31509 -0.12135 0.055464 0.9075
0.18585 -0.20836 0.019945 0.17853 -0.31707 0.054172 0.40715 0.32685 -0.20493 0.099457 0.15329 -0.28035
0.36088 0.31671 -0.2216 -0.094332 0.33993 -0.23604 0.44507 -0.025739 0.2082 -0.28423 0.18867 -0.30867 -
0.015983 0.13985 0.035387 0.25648 -0.18241 0.50119 -0.31602 -0.19771 -0.3002 0.048059 0.14868 -0.45165
0.11831 0.045376 0.31328 -0.052771 0.08615 -0.18376 0.071614 0.30406 0.26742 -0.22895 0.17671 0.33062
0.17738 0.042157 -0.29211 -0.10786 -0.064557 -0.10006 0.39087 -0.21173 -0.085387 -0.040239 -0.1044 -0.019623 -
0.32887 0.15656 0.039189 -0.30531 0.235 -0.025831 0.041146 0.30737 -0.16955 -0.18446 -0.11642 0.038028
0.094888 -0.25135 -0.011466 0.18069 0.44957 -0.28939 -0.46813 0.035372 0.045633 0.1507 -0.098108 -0.31644 -
0.19265 -0.3108 0.32345 0.57775 0.042428 0.2334 -0.093899 -0.50785 -0.68498 0.088108 -0.25361 -0.018187 -
0.50159 -0.19892 -0.12127 -0.21447 0.22551 0.021314 0.078556 -0.0828 -0.27046 -0.19486 0.13457 0.44123
0.13542 -0.37831 0.36109 -0.04392 0.21795 -0.092712 -0.12707 -0.1428 -0.021229 -0.13407 -0.12783 -0.099737 -
0.055585 0.042925 -0.41051 -0.044614 -0.2326 -0.033486 -0.1761 -0.042099 -0.20191 -0.042496 -0.08971 0.062699
-0.39227 0.2632 0.13261 -0.45002 -0.2213 0.31223 0.43488 -0.05547 0.22954 0.70868 -0.37327 0.2844 -0.24495 -
0.28255 0.21883 -0.053093 -0.3006 -0.34203 -0.11602 0.36381 0.11346 0.1853 -0.014843 0.21921 0.047219 -
0.0054492 0.2878 0.51144 0.17271 -0.026182 0.00051472 0.033597 -0.061401 0.25367 -0.13141 -0.056602 -
0.0025169 0.44398 -0.26233 0.21532 0.34318 -0.081855 -0.030759 -0.022955 -0.1757 0.44088 -0.062219

-0.13102 -0.054447 -0.051866 -0.10289 -0.072061 0.16523 -0.17298 0.21865 0.041183 -0.010858 0.074741 0.35226 0.42662 -0.071747 0.25112 0.12082 -0.33192 -0.4728 -0.0090568 0.0030266 0.032861 0.074323 -0.38017 0.091399 -0.16034 -0.050232 -0.094194 0.16656 0.40901 0.069625 0.059306 0.01991 -0.35846 -0.14549 0.24894 0.50184 -0.0073098 -0.4589 -0.10073 -0.099315 0.30583 -0.40577 0.16586 0.055741 0.26776 -0.13515 0.28127 0.069221 -0.20907 0.092053 0.39419 -0.2412 0.01173 -0.16856 -0.0053851 0.14282 0.17513 0.34775 0.178 0.35883 -0.17684 0.53104 0.04751 -0.30134 -0.53297 -0.22041 0.097703 0.052288 0.10849 0.12409 -0.11369 0.19042 0.19554 -0.14949 -0.29675 -0.14285 0.22217 0.21503 -0.2309 0.4381 0.22739 -0.052386 -0.20003 0.19725 -0.032432 -0.14307 0.021958 0.36876 -0.10084 -0.18536 0.27691 -0.43856 0.087418 -0.33836 0.083161 -0.40672 0.14497 -0.41334 0.0012195 -0.32266 0.067225 0.18359 0.010442 -0.15499 -0.82943 -0.069867 -0.26416 0.42656 0.26765 -0.12262 -0.116 -0.076926 -0.16992 0.055428 -0.20699 -0.090381 0.082171 -0.31509 -0.12135 0.055464 0.9075 0.18585 -0.20836 0.019945 0.17853 -0.31707 0.054172 0.40715 0.32685 -0.20493 0.099457 0.15329 -0.28035 0.36088 0.31671 -0.2216 -0.094332 0.33993 -0.23604 0.44507 -0.025739 0.2082 -0.28423 0.18867 -0.30867 -0.015983 0.13985 0.035387 0.25648 -0.18241 0.50119 -0.31602 -0.19771 -0.3002 0.048059 0.14868 -0.45165 0.11831 0.045376 0.31328 -0.052771 0.08615 -0.18376 0.071614 0.30406 0.26742 -0.22895 0.17671 0.33062 0.17738 0.042157 -0.29211 -0.10786 -0.064557 -0.10006 0.39087 -0.21173 -0.085387 -0.040239 -0.1044 -0.019623 -0.32887 0.15656 0.039189 -0.30531 0.235 -0.025831 0.041146 0.30737 -0.16955 -0.18446 -0.11642 0.038028 0.094888 -0.25135 -0.011466 0.18069 0.44957 -0.28939 -0.46813 0.035372 0.045633 0.1507 -0.098108 -0.31644 -0.19265 -0.3108 0.32345 0.57775 0.042428 0.2334 -0.093899 -0.50785 -0.68498 0.088108 -0.25361 -0.018187 -0.50159 -0.19892 -0.12127 -0.21447 0.22551 0.021314 0.078556 -0.0828 -0.27046 -0.19486 0.13457 0.44123 0.13542 -0.37831 0.36109 -0.04392 0.21795 -0.092712 -0.12707 -0.1428 -0.021229 -0.13407 -0.12783 -0.099737 -0.055585 0.042925 -0.41051 -0.044614 -0.2326 -0.033486 -0.1761 -0.042099 -0.20191 -0.042496 -0.08971 0.062699 -0.39227 0.2632 0.13261 -0.45002 -0.2213 0.31223 0.43488 -0.05547 0.22954 0.70868 -0.37327 0.2844 -0.24495 -0.28255 0.21883 -0.053093 -0.3006 -0.34203 -0.11602 0.36381 0.11346 0.1853 -0.014843 0.21921 0.047219 -0.0054492 0.2878 0.51144 0.17271 -0.026182 0.00051472 0.033597 -0.061401 0.25367 -0.13141 -0.056602 -0.0025169 0.44398 -0.26233 0.21532 0.34318 -0.081855 -0.030759 -0.022955 -0.1757 0.44088 -0.062219

(*sunshine*)

Word2Vec

(Mikolov et al., NIPS 2013)

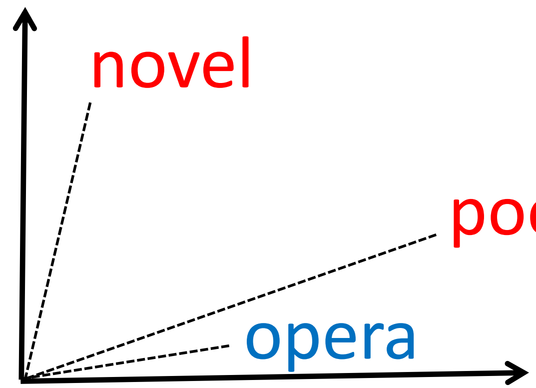


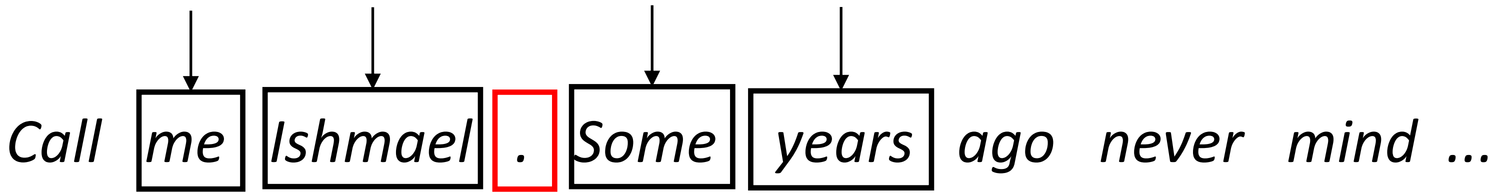
Call | *me* | *Ishmael* | *.* | *Some* *years ago never mind ...*

- Start with random vectors of chosen dimensionality
- Predict surrounding words based on similarity of current vectors
- Iteratively update vectors to reduce error (machine learning)

Word2Vec

(Mikolov et al., NIPS 2013)

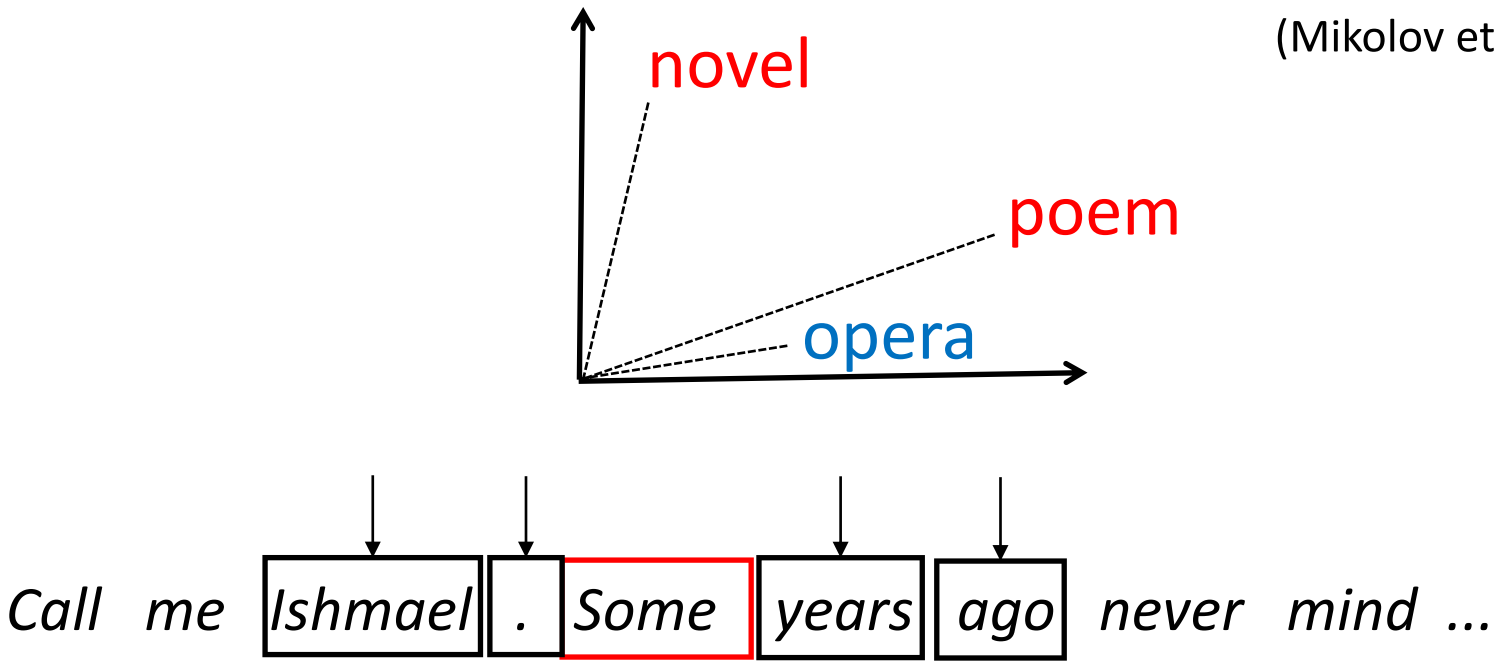




- Start with random vectors of chosen dimensionality
- Predict surrounding words based on similarity of current vectors
- Iteratively update vectors to reduce error (machine learning)

Word2Vec

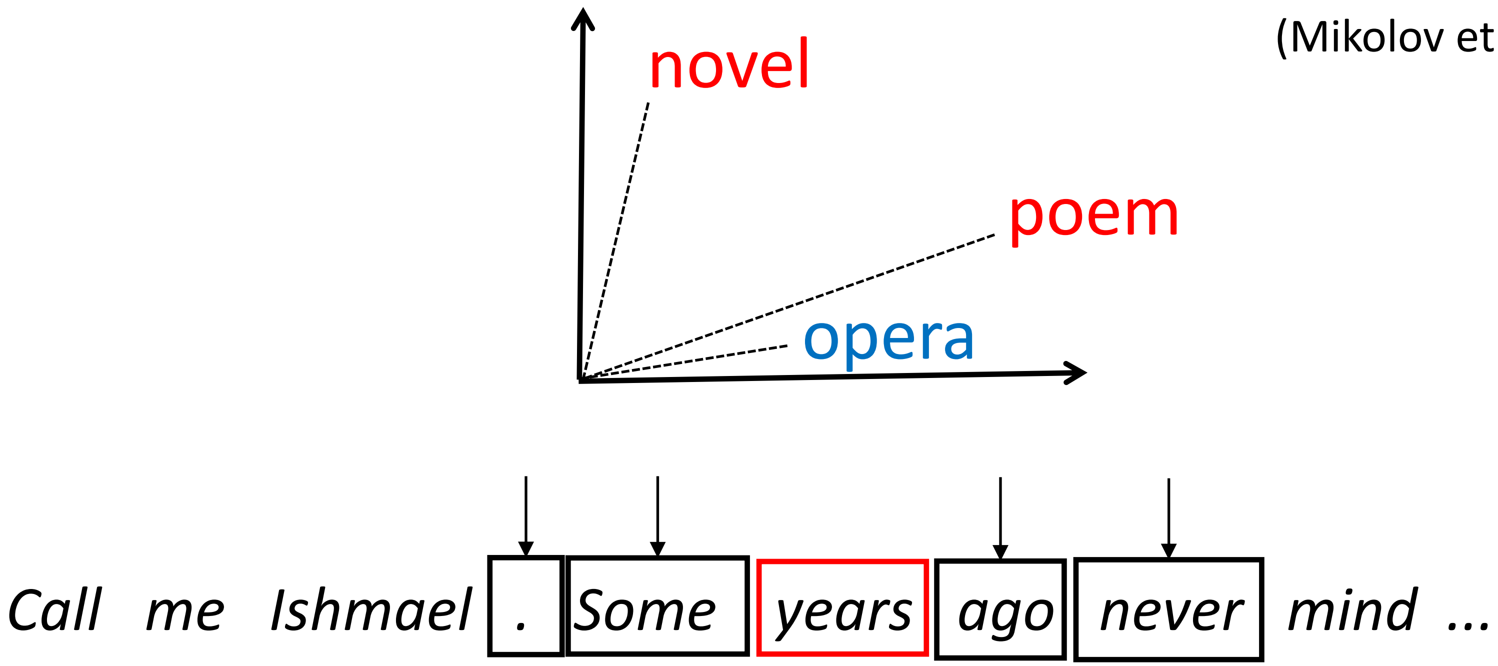
(Mikolov et al., NIPS 2013)



- Start with random vectors of chosen dimensionality
- Predict surrounding words based on similarity of current vectors
- Iteratively update vectors to reduce error (machine learning)

Word2Vec

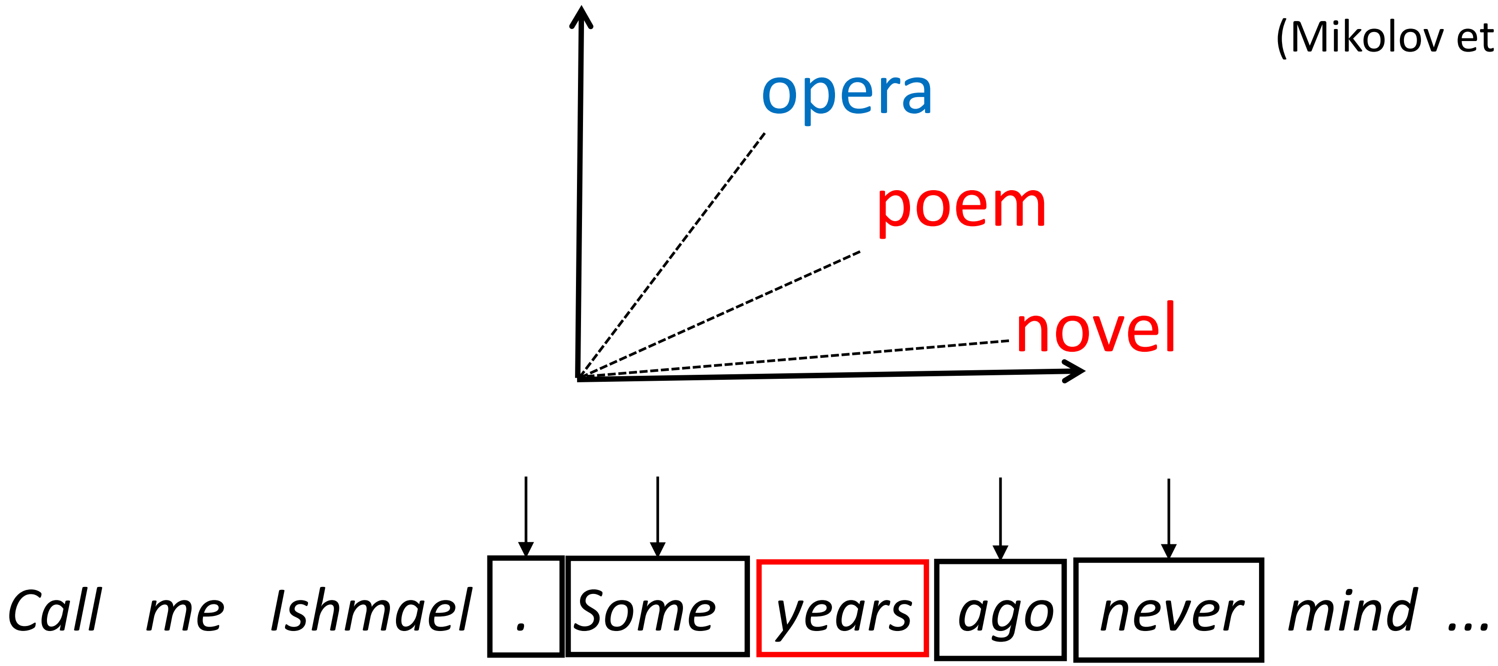
(Mikolov et al., NIPS 2013)



- Start with random vectors of chosen dimensionality
- Predict surrounding words based on similarity of current vectors
- Iteratively update vectors to reduce error (machine learning)

Word2Vec

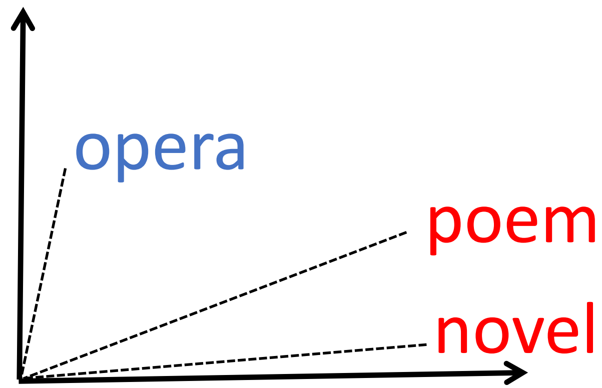
(Mikolov et al., NIPS 2013)

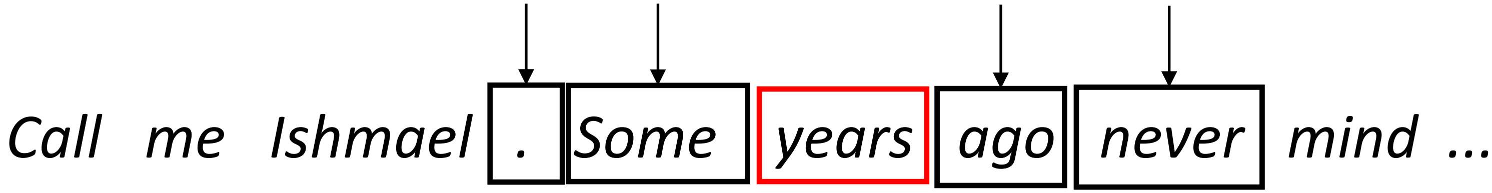


- Start with random vectors of chosen dimensionality
- Predict surrounding words based on similarity of current vectors
- Iteratively update vectors to reduce error (machine learning)

Word2Vec

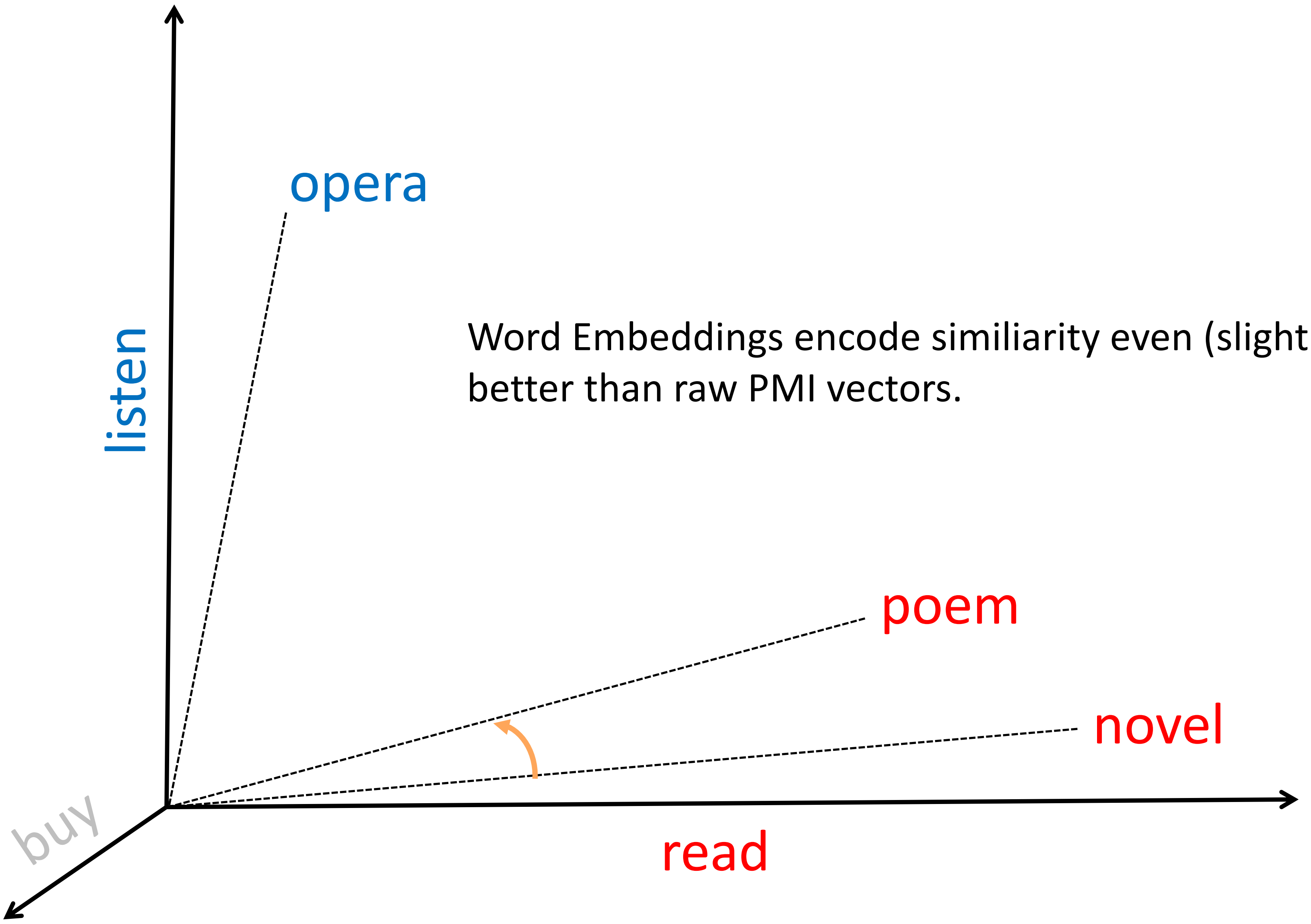
(Mikolov et al., NIPS 2013)





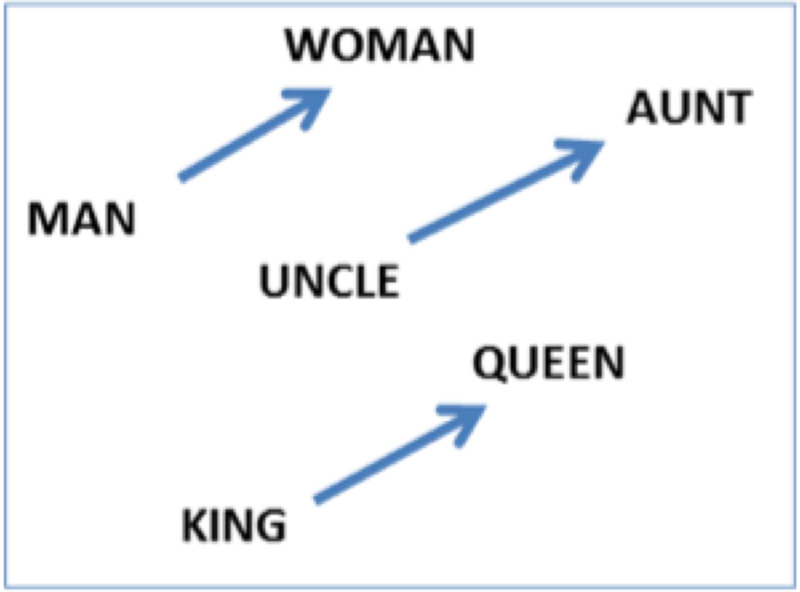
- Start with random vectors of chosen dimensionality
- Predict surrounding words based on similarity of current vectors
- Iteratively update vectors to reduce error (machine learning)

Computing Similarity



Word Embeddings encode similiarity even (slightly) better than raw PMI vectors.

Computing Word Analogies



(Mikolov et al., NAACL 2013)

- Semantic relationships are encoded by vectors, too
- Questions like „What is to *king* as *woman* is to *man*?" can be answered with vector arithmetic

Surprising „Content“ of Word Embeddings

- Morphological relationships: sg.-pl., comparatives
  (Mikolov et al., NAACL 2013)
- Emotion: *terrorism* vs. *sunshine*
  (Buechel & Hahn, NAACL 2018)
- Abstractness: *freedom* vs. *laptop*
  (Köper & Schulte im Walde, LREC 2016)
- Geolocation, GDP, fertility rate and many other referential attributes of country names (*France, Italy, Spain,...)*
  (Gupta et al., EMNLP 2015)