Vector Space Models

Yu Meng
University of Virginia
yumeng5@virginia.edu

Sep 13, 2024

Announcement

- Project proposal due next Friday (09/20 11:59pm) – **no late days**!
- Proposal guideline: https://docs.google.com/document/d/1G-FFrENvM3QQZzlS8ranyFLqFy3PW3AEyTsNjXZPyqI/edit?usp=sharing

Overview of Course Contents

- Week 1: Logistics & Overview
- Week 2: N-gram Language Models
- Week 3: Word Senses, Semantics & Classic Word Representations
- Week 4: Word Embeddings
- Week 5: Sequence Modeling and Transformers
- Week 6-7: Language Modeling with Transformers (Pretraining + Fine-tuning)
- Week 8: Large Language Models (LLMs) & In-context Learning
- Week 9-10: Knowledge in LLMs and Retrieval-Augmented Generation (RAG)
- Week 11: LLM Alignment
- Week 12: Language Agents
- Week 13: Recap + Future of NLP
- Week 15 (after Thanksgiving): Project Presentations

(Recap) WordNet

- Word semantics is complex (multiple senses, various relations)!
- How did people represent word senses and relations in early NLP developments?
- **WordNet**: A manually curated large lexical database
- Three separate databases: one each for nouns, verbs and adjectives/adverbs
- Each database contains a set of lemmas, each one annotated with a set of senses
- Synset (synonym set): The set of near-synonyms for a sense
- Word relations (hypernym, hyponym, antonym) defined between synsets

WordNet: https://wordnet.princeton.edu/

Join at
slido.com
#4252 801

(Recap) WordNet as a Graph

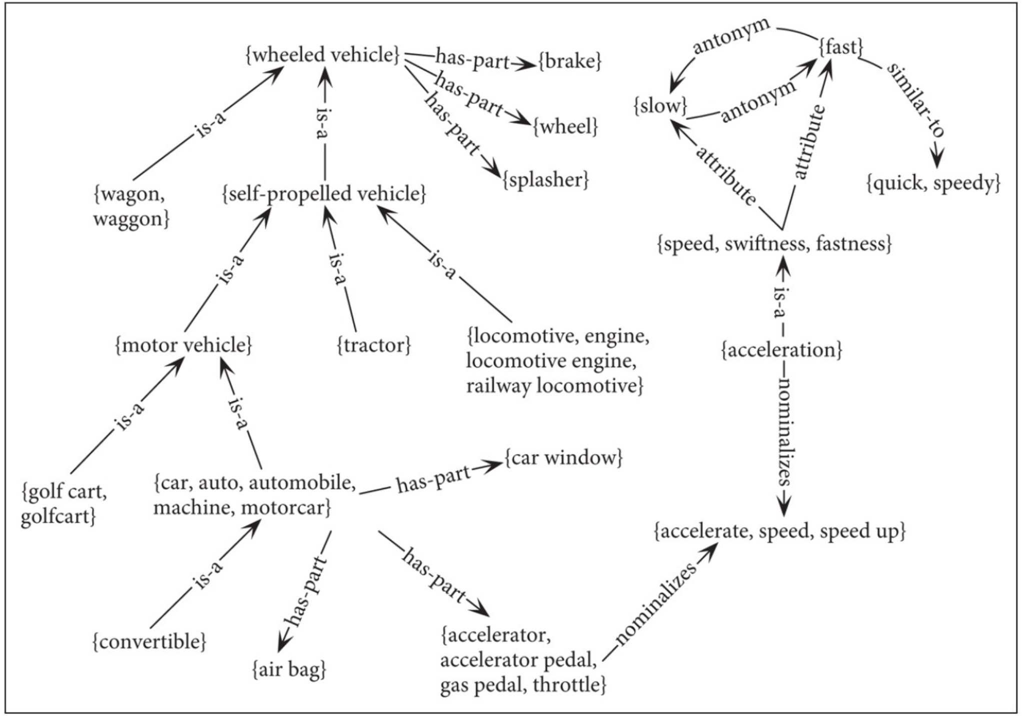


Figure source: https://academic.oup.com/edited-volume/42643/chapter/358151233

(Recap) WordNet for WSD

- All words WSD task: map all input words (nouns/verbs/adjectives/adverbs) to WordNet senses
- Strong baseline: map to the first sense in WordNet (most frequent)
- Modern approaches: sequence modeling architectures (later lectures!)

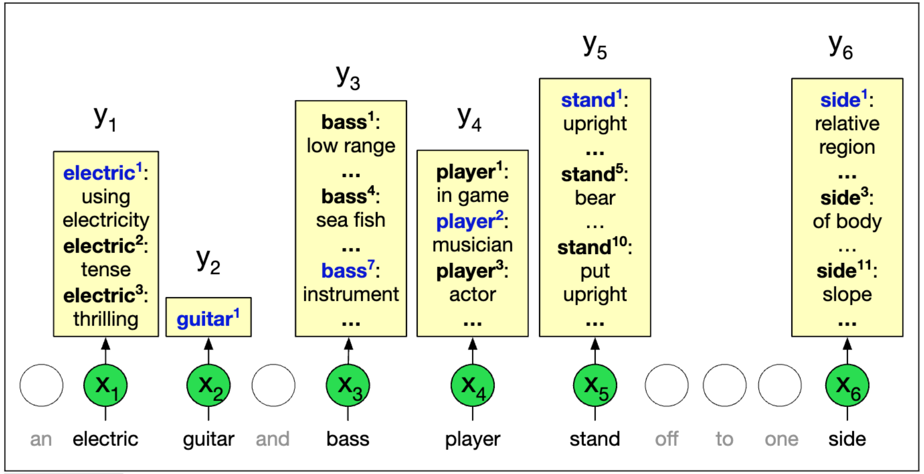


Figure source: https://web.stanford.edu/~jurafsky/slp3/G.pdf

(Recap) WordNet Limitations

- Require significant efforts to construct and maintain/update
  - Hard to keep up with rapidly evolving language usage
- Limited coverage of domain-specific terms & low-resource language
  - No coverage of specialized, domain-specific terms (e.g., medical, legal, or technical)
- Only support individual words and their meanings
  - Do not account for idiomatic expressions, phrasal verbs, or collocations

A more automatic, scalable, and contextualized word semantic learning approach is needed!

(Recap) Numerical Text Representations

- Word similarity computation is important for understanding semantics

Word similarity (on a scale from 0 to 10) manually annotated by humans

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

Word semantics can be multi-faceted

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

- How to represent words numerically? Using multi-dimensional vectors!

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

(Recap) Vector Semantics

- Represent a word as a point in a multi-dimensional semantic space
- A desirable vector semantic space: words with similar meanings are nearby in space

2D visualization of a desirable high-dimensional vector semantic space

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

(Recap) Vector Space Basics

- Vector notation: an N-dimensional vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_N] \in \mathbb{R}^N$
- Vector dot product/inner product:

$$\text{dot product}(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{v} \cdot \boldsymbol{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n = \sum_{i=1}^{N} v_i w_i$$

- Vector length/norm:

$$|\boldsymbol{v}| = \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}} = \sqrt{\sum_{i=1}^{N} v_i^2}$$

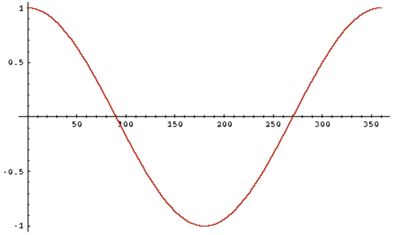
Other (less commonly-used) vector norms:
Manhattan norm, *p*-norm, infinity norm…

- Cosine similarity between vectors:

$$\cos(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}||\boldsymbol{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$

(Recap) Vector Similarity

- Cosine similarity is the most commonly used metric for similarity measurement
  - Symmetric: $\cos(\boldsymbol{v}, \boldsymbol{w}) = \cos(\boldsymbol{w}, \boldsymbol{v})$
  - Not influenced by vector length
  - Has a normalized range: [-1, 1]
  - Intuitive geometric interpretation



Cosine function values under different angles

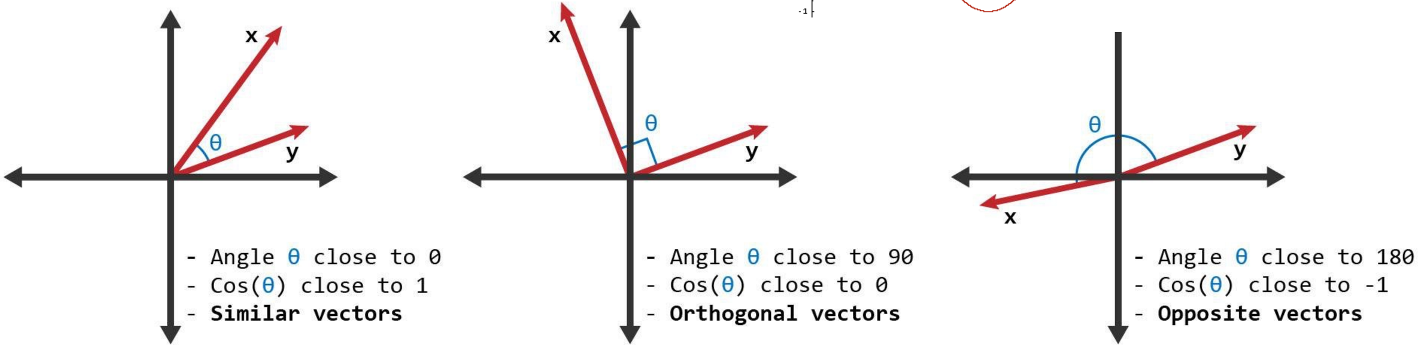


Figure source: https://www.learndatasci.com/glossary/cosine-similarity/

(Recap) How to Represent Words as Vectors?

- Given a vocabulary $\mathcal{V} = \{\text{good}, \text{feel}, \text{I}, \text{sad}, \text{cats}, \text{have}\}$
- Most straightforward way to represent words as vectors: use their indices
- One-hot vector: only one high value (1) and the remaining values are low (0)
- Each word is identified by a unique dimension

$$
\begin{aligned}
\boldsymbol{v}_{\text{good}} &= [1, 0, 0, 0, 0, 0] \\
\boldsymbol{v}_{\text{feel}} &= [0, 1, 0, 0, 0, 0] \\
\boldsymbol{v}_{\text{I}} &= [0, 0, 1, 0, 0, 0] \\
\boldsymbol{v}_{\text{sad}} &= [0, 0, 0, 1, 0, 0] \\
\boldsymbol{v}_{\text{cats}} &= [0, 0, 0, 0, 1, 0] \\
\boldsymbol{v}_{\text{have}} &= [0, 0, 0, 0, 0, 1]
\end{aligned}
$$

UNIVERSITY of VIRGINIA

Join at slido.com #4252 801

Represent Sequences by Word Occurrences

- Consider the mini-corpus with three documents

$$d_1 = \text{“I feel good”}$$
$$d_2 = \text{“I feel sad”}$$
$$d_3 = \text{“I have cats”}$$

$$\boldsymbol{v}_{\text{good}} = [1, 0, 0, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{feel}} = [0, 1, 0, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{I}} = [0, 0, 1, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{sad}} = [0, 0, 0, 1, 0, 0]$$
$$\boldsymbol{v}_{\text{cats}} = [0, 0, 0, 0, 1, 0]$$
$$\boldsymbol{v}_{\text{have}} = [0, 0, 0, 0, 0, 1]$$

- Straightforward way of representing documents: look at which words are present

$$\boldsymbol{v}_{d_1} = [1, 1, 1, 0, 0, 0]$$
$$\boldsymbol{v}_{d_2} = [0, 1, 1, 1, 0, 0]$$
$$\boldsymbol{v}_{d_3} = [0, 0, 1, 0, 1, 1]$$

Document vector similarity $\Rightarrow$

$$\cos(\boldsymbol{v}_{d_1}, \boldsymbol{v}_{d_2}) = \frac{2}{3}$$
$$\cos(\boldsymbol{v}_{d_1}, \boldsymbol{v}_{d_3}) = \frac{1}{3}$$
$$\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = \frac{1}{3}$$

Join at
slido.com
#4252 801

Word-Document Matrix

- With larger text collections, word frequencies in documents entail rich information
- Consider the four plays by Shakespeare and obtain the word frequency statistics
- Look at 4 manually-picked words: “battle” “good” “fool” “wit”

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

There are many more words!

- Document vector representation with word frequencies:

$$\boldsymbol{v}_{d_1} = [1, 114, 36, 20] \quad \boldsymbol{v}_{d_2} = [0, 80, 58, 15] \quad \boldsymbol{v}_{d_3} = [7, 62, 1, 2] \quad \boldsymbol{v}_{d_4} = [13, 89, 4, 3]$$

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Document Similarity

- Document vector representation with word frequencies:

$$\boldsymbol{v}_{d_1} = [1, 114, \boxed{36, 20}] \quad \boldsymbol{v}_{d_2} = [0, 80, \boxed{58, 15}] \quad \boldsymbol{v}_{d_3} = [7, 62, 1, 2] \quad \boldsymbol{v}_{d_4} = [13, 89, 4, 3]$$

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- "fool" and "wit" occur much more frequently in d_1 and d_2 than d_3 and d_4
- d_1 and d_2 are comedies $\cos(\boldsymbol{v}_{d_1}, \boldsymbol{v}_{d_2}) = 0.95$ $\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = 0.81$
- Word frequencies in documents do reflect the semantic similarity between documents!

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

UNIVERSITY of VIRGINIA

Join at
slido.com
#4252 801

Words Represented with Documents

- “Battle”: “the kind of word that occurs in Julius Caesar and Henry V (history plays)”
- “Fool”: “the kind of word that occurs in comedies”

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- Represent words using their co-occurrence counts with documents:

$$\boldsymbol{v}_{\text{battle}} = [1, 0, 7, 13]$$
$$\boldsymbol{v}_{\text{good}} = [114, 80, 62, 89]$$
$$\boldsymbol{v}_{\text{fool}} = [36, 58, 1, 4]$$
$$\boldsymbol{v}_{\text{wit}} = [20, 15, 2, 3]$$

Words Represented with Documents

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

$$\boldsymbol{v}_{\text{battle}} = [1, 0, 7, 13]$$
$$\boldsymbol{v}_{\text{good}} = [114, 80, 62, 89]$$
$$\boldsymbol{v}_{\text{fool}} = [36, 58, 1, 4]$$
$$\boldsymbol{v}_{\text{wit}} = [20, 15, 2, 3]$$

$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{wit}}) = 0.93$$
$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{battle}}) = 0.09$$

Previously:

$$\boldsymbol{v}_{\text{battle}} = [1, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{good}} = [0, 1, 0, 0]$$
$$\boldsymbol{v}_{\text{fool}} = [0, 0, 1, 0]$$
$$\boldsymbol{v}_{\text{wit}} = [0, 0, 0, 1]$$

$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{wit}}) = 0$$
$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{battle}}) = 0$$

Document co-occurrence statistics provide coarse-grained contexts

Fine-Grained Contexts: Word-Word Matrix

Instead of using documents as contexts for words, we can also use words as contexts

4 words to the left	center word	4 words to the right
is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

UNIVERSITY of VIRGINIA

Join at slido.com #4252 801

Fine-Grained Contexts: Word-Word Matrix

Count how many times words occur in a ±4 word window around the center word

context word

center word

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Counts derived from the Wikipedia corpus

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

UNIVERSITY of VIRGINIA

Join at slido.com #4252 801

Word Similarity Based on Word Co-occurrence

- Word-word matrix with ±4 word window

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

- “digital” and “information” both co-occur with “computer” and “data” frequently
- “cherry” and “strawberry” both co-occur with “pie” and “sugar” frequently
- Word co-occurrence statistics reflect word semantic similarity!
- Issues? Sparsity!

Is Raw Frequency A Good Representation?

- On the one hand, high frequency can imply semantic similarity
- On the other hand, there are words with universally high frequencies

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- Can we reweight the raw frequencies so that distinctively high frequency terms are highlighted?

Term Frequency (TF)

- A word appearing 100 times in a document doesn't make it 100 times more likely to be relevant to the meaning of the document
- Instead of using the raw counts, we squash the counts with log scale

$$\mathrm{TF}(w,d) = \begin{cases} 1 + \log_{10} \mathrm{count}(w,d) & \mathrm{count}(w,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Document Frequency (DF)

- Motivation: Give a higher weight to words that occur only in a few documents
  - Terms that are limited to a few documents are more discriminative
  - Terms that occur frequently across the entire collection aren't as helpful
- Document frequency (DF): count how many documents a word occurs in

$$\mathrm{DF}(w) = \sum_{i=1}^{N} \mathbb{1}(w \in d_i) \longrightarrow$$

Evaluates to 1 if w occurs in d_i
otherwise evaluates to 0

- DF is NOT defined to be the total count of a word across all documents (collection frequency)!

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Inverse Document Frequency (IDF)

- We want to emphasize discriminative words (with low DF)
- Inverse document frequency (IDF): total number of documents (N) divided by DF, in log scale

$$\mathrm{IDF}(w) = \log_{10}\left(\frac{N}{\mathrm{DF}(w)}\right)$$

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

DF & IDF statistics in the Shakespeare corpus (37 documents)

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

TF-IDF Weighting

The TF-IDF weighted value characterizes the "salience" of a term in a document

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \text{IDF}(w)$$

TF-IDF weighted

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.246	0	0.454	0.520
good	0	0	0	0
fool	0.030	0.033	0.0012	0.0019
wit	0.085	0.081	0.048	0.054

$$\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = 0.10 \quad \cos(\boldsymbol{v}_{d_3}, \boldsymbol{v}_{d_4}) = 0.99$$

Raw counts

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

$$\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = 0.81 \quad \cos(\boldsymbol{v}_{d_3}, \boldsymbol{v}_{d_4}) = 0.99$$

UNIVERSITY of VIRGINIA

Join at
slido.com
#4252 801

How to Define Documents?

- The concrete definition of documents is usually open to different design choices
  - Wikipedia article/page
  - Shakespeare play
  - Book chapter/section
  - Paragraph/sentence
  - …
- Larger documents provide broader context; smaller ones provide focused insights
- Depends on the analysis need: interested in global trends across documents (e.g., news articles) vs. more local patterns (e.g., specific sections of a legal document)?

Probability-Based Weighting

- TF-IDF weighting scheme is based on heuristics
- Can we weigh the raw counts with probabilistic approaches?
- Intuition: the association between two words can be reflected by **how much they co-occur more than by chance**

context word | summed counts

center word | summed counts

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Word Association Based on Probability

- In probability theory, when two random variables A & B are independent, we have

  Joint probability $p(A, B) = p(A)p(B)$

- When two words co-occur by chance, we expect their probabilities to satisfy the independence assumption: $p(w_1, w_2) = p(w_1)p(w_2)$
- When $p(w_1, w_2) > p(w_1)p(w_2)$, two words co-occur more often than would be expected by chance
- How to develop a probabilistic metric to characterize this association?

Pointwise Mutual Information (PMI)

- PMI compares the probability of two words co-occurring with the probabilities of the words occurring independently

$$\mathrm{PMI} = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} = \log_2 \frac{\#(w_1, w_2)}{\#(w_1)\#(w_2)}$$

- PMI = 0: Two words co-occur as expected by chance => no particular association
- PMI > 0: Two words co-occur more often than by chance => the higher the PMI, the stronger the association between the words
- PMI < 0: Two words co-occur less often than expected by chance => negative associations; not much actionable insight
- Positive PMI (PPMI): replaces all negative PMI values with zero

$$\mathrm{PPMI} = \max\left(\log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}, 0\right)$$

UNIVERSITY of VIRGINIA

Join at
slido.com
#4252 801

PPMI Example

Raw counts

	computer	data	result	pie	sugar
cherry	2	8	9	442	25
strawberry	0	0	1	60	19
digital	1670	1683	85	5	4
information	3325	3982	378	5	13

PPMI-weighted matrix

	computer	data	result	pie	sugar
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Issue: biased toward infrequent events (rare words tend to have very high PMI values)

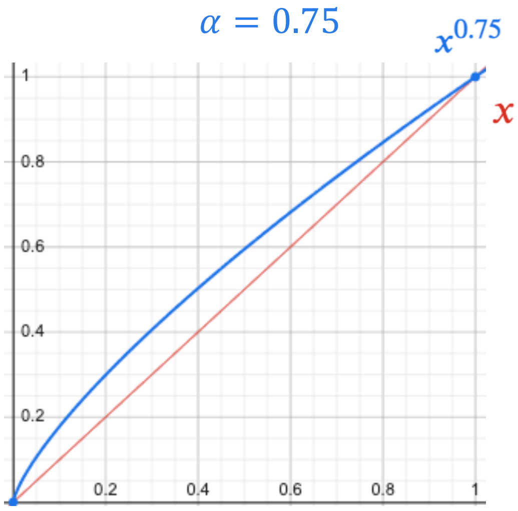
PPMI with Power Smoothing

Power smoothing: Manually boost low probabilities by raising to a power α

$$\text{PPMI} = \max\left(\log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}, 0\right)$$

Original: $$p(w) = \frac{\#(w)}{\sum_{w' \in \mathcal{V}} \#(w')}$$

Power smoothed: ($\alpha < 1$) $$p_\alpha(w) = \frac{\#(w)^\alpha}{\sum_{w' \in \mathcal{V}} \#(w')^\alpha}$$



PPMI with Add-*k* Smoothing

- Another way of increasing the counts of rare occurrences is to apply add-*k* smoothing

	computer	data	result	pie	sugar
cherry	2	8	9	442	25
strawberry	0	0	1	60	19
digital	1670	1683	85	5	4
information	3325	3982	378	5	13

Add a constant *k* to all counts

- The larger the *k* (*k* can be larger than 1), the more we boost the probability of rare occurrences

TF-IDF vs. PMI Weighting

- TF-IDF
  - Measures the importance of a word in a document relative to other documents (corpus)
  - Context granularity: document level
  - Based on heuristics
  - High TF-IDF = frequent in a document but infrequent across the corpus
- PMI:
  - Measures the strength of association between two words
  - Context granularity: word pair level (usually based on local context windows)
  - Based on probability assumptions
  - High PMI = words co-occur more often than expected by chance, a strong association

UNIVERSITY of VIRGINIA

Join at slido.com #4252 801

Summary: Word Semantics & Senses

- Understanding word semantics & senses help us build better language models!
- Word semantics is complex
  - Polysemy: a single word having multiple meanings
  - Multi-faceted: word meanings entail various aspects (e.g., valence, arousal, dominance)
- Many types of word relations: synonyms, antonyms, hyponyms & hypernyms…
- Word relations are usually not binarized (e.g., perfect synonyms are rare); word similarity is usually a more flexible measure

University of Virginia

Join at slido.com #4252 801

Summary: Classic Word Representations

- Large-scale lexical databases (WordNet) were constructed in early NLP developments
- WordNet consists of manually curated synsets linked by relation edges
- WordNet can be used as a database for word sense disambiguation
- WordNet has significant limitations:
  - Require significant efforts to construct and maintain/update
  - Limited coverage of domain-specific terms & low-resource language
  - Only support individual words and their meanings

Summary: Vector Space Models

- Vector semantic space: use vector representations to reflect word semantics
- Cosine similarity is the most-commonly used metric for vector similarity
- Word-document & word-word co-occurrence statistics provide valuable semantic information – count-based vector representations work decently well
- Raw counts are not good representations (e.g., biased to universally frequent terms)
- TF-IDF highlights the important words in a document relative to other documents
- PMI measures the strength of association between two words based on probabilistic (independence) assumptions

Thank You!

Yu Meng
University of Virginia
yumeng5@virginia.edu