Introduction to Computational Linguistics: Lexical semantics and classification

Sharon Goldwater

27 July 2015

Computational semantics

We talked about ways to "hack" PCFGs to return better parses.

Some of these are effectively encoding semantic information/world knowledge into a syntactic grammar.

Maybe it's time to think more about semantics generally ...

School of ormat

Sharon Goldwater

Lexical semantics

Natural language understanding

Full NLU is hard! So most computational work focuses on sub-problems:

- Recognizing lexical relationships: similarity, synonymy, hyponymy (kind-of), meronymy (part-of).
- Disambiguating word senses (e.g., bank: river or finance?)
- Identifying which phrases fill the thematic roles of a verb. (J&M 19.4, 20.9)
- Recognizing entailment relations between sentences.
- Interpreting sentences to logical forms (semantic parsing), e.g., in a database query language. (J&M Ch 17-18).

Sharon Goldwater

l exical semantics

WordNet

• One way to get lexical relationships: use a database or ontology.

- WordNet (English) is a hand-built resource containing 117,000 synsets: sets of synonymous words (See http://wordnet.princeton.edu/)
- Synsets are connected by relations such as
 - hyponym/hypernym (IS-A: chair-furniture)
 - meronym (PART-WHOLE: leg-chair)
 - antonym (OPPOSITES: good-bad)
- globalwordnet.org now lists wordnets in over 50 languages (but variable size/quality/licensing)

Sharon Goldwater

Lexical semantics

Pattern	Participating Senses	Example Sentences
Animal for fur	Mink, chinchilla, rabbit, beaver, raccoon*, alpaca*, crocodile*	The <i>mink</i> drank some water / She likes to wear <i>mink</i>
Animal/Object for personality	Chicken, sheep, pig, snake, star*, rat*, doll*	The chicken drank some water / He is a chicken
Animal for meat	Chicken, lamb, fish, shrimp, salmon*, rabbit*, lobster*	The chicken drank some water / The chicken is tasty
Artifact for activity	Shower, bath, sauna, baseball,	The shower was leaking / The shower was relaxing
Body part for object part	Arm, leg, hand, face, back*, head*, foot*, shoulder*, lip*,	John's <i>arm</i> was tired / The <i>arm</i> was reupholstered
Building for people	Church, factory, school, airplane,	The <i>church</i> was built 20 years ago / The <i>church</i> sang a song
Complement Coercion	Begin, start, finish, try	John <i>began</i> reading the book / John <i>began</i> the book
Container for contents	Bottle, can, pot, pan, bowl*, plate*, box*, bucket*	The <i>bottl</i> e is made of steel / He drank half of the <i>bottl</i> e

Lexical semantics

Lexical relationships and disambiguation

Recognizing these can help with, e.g., question answering and machine translation.

- QA: Which animals love to swim? requires answers that are hyponyms of animal.
- MT: interest might translate as Zins (financial charge), Anteil (legal stake), or Interesse (concern, curiousity).

"Interest" example due to Philipp Koehn.

Sharon Goldwater

Lexical semantics Word Sense Ambiguity

- Not all problems can be solved by WordNet alone.
- Two completely different words can be spelled the same (homonyms):
 - I put my money in the *bank*. VS. He rested at the *bank* of the river. You *can* do it! She bought a *can* of soda. VS.
- More generally, words can have multiple (related or unrelated) senses (polysemes)
- Polysemous words often fall into (semi-)predictable patterns: see next slides (from Hugh Rabagliati in PPLS).

Sharon Goldwater Lexical semanti		cs
Pattern	Participating Senses	Example Sentences
Figure for Ground	Window, door, gate, goal	The window is broken / The cat walked through the window
Grinding	Apple, chair, fly	The apple was tasty / There is apple all over the table
Instrument for action	Hammer, brush, shovel, tape, lock*, bicycle*, comb*, saw*	The hammer is heavy / She hammered the nail into the wall
Instance of an entity for kind	Tennis, soccer, cat, dog, class*, dinner*, chair*, table*	Tennis was invented in England / Tennis was fun today
Location / Place at location	Bench, land, floor, ground, box*, bottle*, jail*	The bench was made of pine / The coach benched the player
Object for placing at goal	Water, paint, salt, butter, frame*, dress*, oil*	The water is cold / He watered the plant.
Object for taking from source	Milk, dust, weed, peel, pit*, skin*, juice*	The milk tastes good / He milked the cow
Material for artifact	Tin, iron, china, glass, linen*, rubber*, nickel*, fur*	Watch out for the broken glass / He filled the glass with water

Sharon Goldwater

27 July 2015

Pattern	Participating Senses	Example Sentences
Place for an event	Vietnam, Korea, Waterloo, Iraq	It is raining in Vietnam / John was shot during Vietnam
Place for an institution	White House, Washington, Hollywood, Pentagon, Wall Street*, Supreme Court	The White House is being repainted / The White House made an announcement
Plant for food or material	Corn, broccoli, coffee, cotton, lettuce*, eggs*, oak*, pine*	The large field of <i>corn /</i> The <i>corn</i> is delicious
Portioning	Water, beer, jam	She drank some <i>water /</i> She bought three <i>waters</i>
Publisher for product	Newspaper, magazine, encyclopedia, Wall Street Journal*, New York Times*,	The newspaper is badly printed / The newspaper fired three employees
Artist for product	Writer, artist, composer, Shakespeare, Dickens*, Mozart*, Picasso*	The writer drank a lot of wine / The writer is hard to understand
Object for contents	Book, CD, DVD, TV*, magazine*, newspaper*	The heavy, leather- bound <i>book /</i> The <i>book</i> is funny.
Visual Metaphor	Word Bug t column sticks	Most of the weight rests on the beam / There was a beam of light

- S1: a sense of concern with and curiosity about someone or something, Synonym: involvement
- S2: the power of attracting or holding one's interest (because it is unusual or exciting etc.), Synonym: interestingness
- S3: a reason for wanting something done, Synonym: sake
- S4: a fixed charge for borrowing money; usually a percentage of the amount borrowed
- S5: a diversion that occupies one's time and thoughts (usually pleasantly), Synonyms: pastime, pursuit
- \bullet S6: a right or legal share of something; a financial involvement with something, Synonym: stake
- S7: (usually plural) a social group whose members control some field of activity and who have common aims, Synonym: interest group

Sharon Goldwater

Lexical semantics

Word sense disambiguation (WSD)

- · For many applications, we would like to disambiguate senses
 - we may be only interested in one sense
 - searching for $\ensuremath{\operatorname{chemical}}$ plant on the web, we do not want to know about chemicals in bananas
- Task: Given a polysemous word, find the sense in a given context
- Typical approach uses context words as features to train a supervised classifier.

How many senses?

- How many senses does the word interest have?
 - She pays 3% interest on the loan.
 - He showed a lot of **interest** in the painting.
 - Microsoft purchased a controlling **interest** in Google.
 - It is in the national **interest** to invade the Bahamas.
 - I only have your best **interest** in mind.
 - Playing chess is one of my **interests**.
 - Business **interests** lobbied for the legislation.
- Are these seven different senses? Four? Three?

"Interest" example due to Philipp Koehn.

Lexical semantics

Polysemy in WordNet

- Polysemous words are part of multiple synsets
- This is why relationships are defined between synsets, not words
- On average,

Sharon Goldwater

- nouns have 1.24 senses (2.79 if exluding monosemous words)
- verbs have 2.17 senses (3.57 if exluding monosemous words)
- Some argue Wordnet is too fine-grained.

Stats from: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

Sharon Goldwater

11

Lexical semantics Classification (aka Categorization)

- Important in human learning and processing of language:
 - phonetic categorization
 - spoken word recognition
 - learning syntactic categories
- And in NLP and linguistics:
 - Word sense disambiguation
 - Classifying text: into different topics, spam/not-spam, 1-5 star review
 - Author attribution: male/female, specific author, healthy/mental illness

Sharon Goldwater

Lexical semantics

Formalizing the classification task

- Assume we've made some observations \vec{x} about the thing we want to classify. (\vec{x} are observed variables).
- y (a hidden variable) is the class label, Y the set of class labels. We want:

 $\hat{y} = \operatorname*{argmax}_{y \in Y} P(y|\vec{x})$

- Text classification: \vec{x} are words in a document, y is spam/not spam.
- WSD: \vec{x} are features of the ambiguous word, y is the sense.

Sharon Goldwate

Sharon Goldwater

12

Lexical semantics

13

15

WSD as example classification task

- $\bullet\,$ disambiguate three senses of the target word $\underline{\operatorname{plant}}$
- \vec{x} are, e.g., the words and POS tags in the document the target word occurs in
- \boldsymbol{y} is the latent sense. Assume three possibilities:
 - y =sense
 - 1 Noun: a member of the plant kingdom

Lexical semantics

- 2 Verb: to place in the ground
- 3 Noun: a factory

Naive Bayes classifier

• Start with our usual step of applying Bayes' rule:

$$\hat{y} = \underset{y \in Y}{\operatorname{argmax}} P(y|\vec{x})$$
$$= \underset{y \in Y}{\operatorname{argmax}} P(\vec{x}|y) P(y)$$

Lexical semantics
Application to WSD

- $x_1 = POS$ of target word (obtained automatically, so not perfect)

Naive Bayes classifier

• Start with our usual step of applying Bayes' rule:

 $\hat{y} = \operatorname*{argmax}_{y \in Y} P(y|\vec{x})$ $= \operatorname*{argmax}_{y \in Y} P(\vec{x}|y) P(y)$

• Then, make a **Naive Bayes** assumption: features are conditionally independent given class. Therefore,

 $P(\vec{x}|y) = P(x_1, x_2, \dots, x_n|y)$ $\approx P(x_1|y)P(x_2|y)\dots P(x_n|y)$

17

19

21

Lexical semantics Application to WSD

- Let's suppose the following features:
 - $-x_1 = POS$ of target word (obtained automatically, so not perfect)
 - $x_2 =$ word to left of target word
 - x_3 = word to right of target word
 - x_4 = document contains the word animal
- In this case we might expect:
 - $P(x_1 = \mathbb{NN} \mid y = 1)$ very high, and $P(x_1 = \mathbb{NN} \mid y = 2)$ very low
 - $P(x_2 = \text{chemical} \mid y = 1)$ much lower than $P(x_2 = \text{chemical} \mid y = 3)$
- etc.

Sharon Goldwater

Sharon Goldwater

Sharon Goldwater

• Let's suppose the following features:

- x_2 = word to left of target word

- x_3 = word to right of target word

- x_4 = document contains the word animal

Lexical semantics

Training the model

As usual, we can estimate these probabilities from an annotated corpus.

- The prior distribution over classes P(y) (proportion of things in each class).
- The feature probabilities $P(x_i|y)$ for each possible class y.

Given the probabilites, just apply them to features observed in test cases to find the highest probability class.

Sharon Goldwater

16

Lexical semantics Advantages of Naive Bayes

- Very easy to implement
- Very fast to train and test
- Doesn't require as much training data as some other methods
- Usually works reasonably well
- This should be your baseline method for any classification task

Sharon Goldwate

Lexical semantics

Problems with Naive Bayes

- Naive Bayes assumption is naive!
- Consider our WSD categories for plant.
- Are the features we used really independent given the category?
 - POS tag and word to the left?
 - word to the left and word to the right?
 - animal in doc and word to left?

Sharon Goldwater

Sharon Goldwater

20

Lexical semantics

Problems with Naive Bayes

- Naive Bayes assumption is naive!
- Consider our WSD categories for plant.
- Are the features we used independent given the category?
 - POS tag and word to the left?
 - word to the left and word to the right?
 - animal in doc and word to left?
- Clearly not, in some cases more than others.

Non-independent features

- Features are not usually independent given the class
- Adding multiple feature types (e.g., words and morphemes) often leads to even stronger correlations between features
- Accuracy of classifier can sometimes still be ok, but it will be highly overconfident in its decisions.
 - Ex: NB sees 5 features that all point to class 1, treats them as five independent sources of evidence.
 - Like asking 5 friends for an opinion when some got theirs from each other.

A different approach to modeling

- so far, all our models have been generative
- discriminative models can address some of the above issues (although they will introduce others)

Sharon Goldwater

Lexical semantics

Generative probabilistic models

- Model the joint probability $P(\vec{x}, \vec{y})$
 - \vec{x} : the observed variables
 - \vec{y} : the latent variables (for Naive Bayes, just one y).

Model	\vec{x}	$ec{y}$
Naive Bayes	features	class
HMM	words	tags
PCFG	words	rules in tree

Generative models have a "generative story"

Lexical semantics

- a probabilistic process that describes how the data were created
 - Multiplying probabilities of each step gives us $P(\vec{x}, \vec{y})$.
- Naive Bayes: For each item *i* to be classified,
 - Generate its class $y^{(i)}$

24

26

Sharon Goldwater

- Generate its features $x_1^{(i)} \dots x_n^{(i)}$ conditioned on $y^{(i)}$
- See previous lectures for HMM and PCFG generative stories.

Sharon Goldwater

Lexical semantics

Inference in generative models

• At test time, given only \vec{x} , we infer \vec{y} using Bayes' rule:

$$P(\vec{y}|\vec{x}) = \frac{P(\vec{x}|\vec{y})P(\vec{y})}{P(\vec{x})}$$

• So, we actually model $P(\vec{x}, \vec{y})$ as $P(\vec{x}|\vec{y})P(\vec{y})$.

- You can confirm this for each of the previous models.

Lexical semantics

27

20

25

Discriminative probabilistic models

- Model $P(\vec{y}|\vec{x})$ directly
- No model of $P(\vec{x}, \vec{y})$
- No generative story
- No Bayes' rule

Sharon Goldwater

Sharon Goldwate

• Need not be probabilistic.

nearest neighbor methods.

l exical semantics Discriminative models more broadly

• Examples: support vector machines, artificial neural networks, decision trees,

• Here, we consider only one method: Maximum Entropy (MaxEnt) models.

• Trained to *discriminate* correct vs. wrong values of \vec{y} , given input \vec{x} .

28

Lexical semantics

MaxEnt classifiers

- Used widely in many different fields, under many different names
- Most commonly, multinomial logistic regression
 - multinomial if more than two possible classes
 - otherwise (or if lazy) just logistic regression
- Also: log-linear model, single neuron classifier, harmonic grammar, etc ...

Sharon Goldwate

Sharon Goldwater

Defining a MaxEnt model for WSD

- Define features $f_i(\vec{x}, y)$ that depend on both observed and latent variables.
- Each feature f_i has a real-valued weight w_i (learned in training).

 $\begin{array}{ll} f_1: & {\rm POS(tgt)} = {\rm NN} \ \& \ y = 1 \\ f_2: & {\rm POS(tgt)} = {\rm NN} \ \& \ y = 2 \\ f_3: & {\rm preceding_word(tgt)} = {\rm `chemical'} \ \& \ y = 3 \end{array}$

 $f_4:$ doc_contains('animal') & y=1

where tgt is the target word

• For senses {1: member of plant kingdom; 2: put in ground; 3: factory}, which weights are likely to be positive? Negative?

Sharon Goldwater

Lexical semantics

Feature templates

• In practice, features are usually defined using templates

POS(tgt)=t & y
preceding_word(tgt)=w & y
doc_contains(w) & y

- instantiate with all possible POSs t or words w and classes y
- usually filter out features occurring very few times
- templates can also define real-valued or integer-valued features
- NLP tasks often have a few templates, but 1000s or 10000s of features
- Whereas in statistical analysis, we try to have very few features (independent variables), to understand which affect the dependent variable.

Sharon Goldwater

Lexical semantics Which features are active?

• Example doc: [... animal/NN ... chemical/JJ plant/NN ...]

- Notice that zero-valued features have no effect on the final probability
- Other features will be multiplied by their weights, summed, then exp.

Feature templates

• In practice, features are usually defined using templates

POS(tgt)=t & y preceding_word(tgt)=w & y doc_contains(w) & y

- instantiate with all possible POSs t or words w and classes y
- usually filter out features occurring very few times
- templates can also define real-valued or integer-valued features

Sharon Goldwater

32

34

Lexical semantics

33

35

Classification with MaxEnt

• Choose the class that has highest probability according to

$$P(y|\vec{x}) = \frac{1}{Z} \exp\left(\sum_{i} w_i f_i(\vec{x}, y)\right)$$

where

$$-\exp(x) = e^x$$

- $\sum_i w_i f_i$ is the *dot product* of vectors \vec{w} and \vec{f} , also written $\vec{w} \cdot \vec{f}$.
- The normalization constant $Z = \sum_{y'} \exp(\sum_i w_i f_i(\vec{x}, y'))$

```
Sharon Goldwater
```

Lexical semantics Training the model

• Given given items $x^{(1)}\dots x^{(N)}$ with labels $y^{(1)}\dots y^{(N)},$ choose weights that make the labels most probable under the model:

$$\hat{w} = \operatorname*{argmax}_{\vec{w}} \sum_{j} \log P(y^{(j)} | x^{(j)})$$

• called conditional maximum likelihood estimation (CMLE)

Sharon Goldwate

Sharon Goldwater

Lexical semantics

Training the model

• Given given items $x^{(1)}\ldots x^{(N)}$ with labels $y^{(1)}\ldots y^{(N)}$, choose weights that make the labels most probable under the model:

$$\hat{v} = \operatorname*{argmax}_{\vec{w}} \sum_{j} \log P(y^{(j)} | x^{(j)})$$

- called conditional maximum likelihood estimation (CMLE)
- Like MLE, CMLE will overfit, so we use tricks (regularization) to avoid that.
- Training isn't just counting things; instead requires iterative methods that gradually update the weights: can be slow.
- Implemented in many existing packages (e.g., MALLET, scikit-learn)

38

Sharon Goldwater

39

MaxEnt and Optimality Theory

Lexical semantics

Suppose our classification problem is: Which surface form y is best, given underlying form x and constraints $\vec{f}?$

/k o t . z/	IDENT-VOICE	*Insert	*Delete	Faith-Voice
[kot.z]	1	0	0	0
[kot.iz]	0	1	0	0
[kot.zii]	1	2	0	0
[kot.]	0	0	1	0
[kot.s]	0	0	0	1

Sharor

36

Sharon Goldwater

MaxEnt and Optimality Theory

- MaxEnt is similar to OT, except
 - features can have positive or negative weights (vs. constraints: violations always bad)
 - mulitple low-ranked active features can gang up and outweigh a single high-ranked feature.
- In fact, Harmonic Grammar (precursor to OT) is MaxEnt.
- Linguistically motivated reasons for moving to OT, but current work looks at advantages/disadvantes of each (in typology, learning, etc).

For more, see e.g. Goldwater and Johnson (2003); Hayes and Wilson (2008); Johnson et al. (2015)

Sharon Goldwater

Lexical semantics

Introduction to Computational Linguistics: Distributional semantics

Sharon Goldwater

27 July 2015

informatics

Sharon Goldwater

27 July 2015

Meaning from context(s)

Distributional semantics

• Consider the example from J&M (quoted from earlier sources):

a bottle of *tezgüino* is on the table everybody likes *tezgüino tezgüino* makes you drunk we make *tezgüino* out of corn Sharon Goldwater

Sharon Goldwater

Distributional semantics

Distributional hypothesis

- perhaps we can infer meaning just by looking at the contexts a word occurs in
- perhaps meaning IS the contexts a word occurs in (!)
- either way, similar contexts imply similar meanings:
 - this idea is known as the distributional hypothesis

Sharon Goldwate

Distributional semantics

"Distribution": a polysemous word

- Probability distribution: a function from outcomes to real numbers
- Linguistic distribution: the set of contexts that a particular item (here, word) occurs in

Sharon Goldwater

Distributional semantics

Distributional semantics: basic idea

- Represent each word w_i as a vector of its contexts
- Ex: each dimension is a context word; = 1 if it co-occurs with w_i , otherwise 0.

	pet	bone	fur	run	brown	screen	mouse	fetch
$w_1 =$	1	1	1	1	1	0	0	1
$w_2 =$	1	0	1	0	1	0	1	0
$w_3 =$	0	0	0	1	0	1	1	0

• Note: real vectors would be far more sparse

Sharon Goldwater

Distributional semantics

References

- Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*, pages 113–122, Stockholm University.
- Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Johnson, M., Pater, J., Staubs, R., and Dupoux, E. (2015). Sign constraints on feature weights improve a joint model of word segmentation and phonology. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 303–313, Denver, Colorado. Association for Computational Linguistics.

Lexical semantics

How to represent other aspects of word meaning?

- QA ex: What is a good way to remove wine stains?
 - To know that Salt is a great way to eliminate wine stains is a good answer,
 - need to know that good and great have similar (in fact graded) meanings.
- There is some work on inferring similarity using WordNet
- But distributional representations are much more common.

Questions to consider

- What defines "context"? (What are the dimensions, what counts as cooccurrence?)
- How to weight the context words (Boolean? counts? other?)
- How to measure similarity between vectors?

Defining the context

- Usually ignore **stopwords** (function words and other very frequent/uninformative words)
- Usually use a large window around the target word (e.g., 100 words, maybe even whole document)
- Can use just cooccurrence within window, or may require more (e.g., dependency relation from parser)
- Note: all of these for *semantic* similarity; for *syntactic* similarity, use a small window (1-3 words) and track *only* frequent words.

Sharon Goldwater

Distributional semantics

How to weight the context words

- binary indicators not very informative
- presumably more frequent co-occurrences matter more
- but, is frequency good enough?
 - frequent words are expected to have high counts in the context vector
 - regardless of whether they occur more often with this word than with others

Sharon Goldwater

Sharon Goldwater

Distributional semantics Collocations

- We want to know which words occur *surprisingly* often in the context of w
- Put another way, what collocations include w?

Sharon Goldwater

Distributional semantics Collocations

- We want to know which words occur surprisingly often in the context of w
- Put another way, what collocations include w?
- Collocations used not just for word similarity (as in next slides).
- In general, they tell us about word associations.
- For example, which concepts associate with positive vs. negative words? (sentiment analysis).

Distributional semantics Mutual information

• Recall the definition of pointwise mutual information:

 $\mathsf{PMI}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \Leftarrow \mathsf{Actual prob of seeing words } x \mathsf{ and } y \mathsf{ together}$

- How much more/less likely is the cooccurrence than if the words were independent?
- Defn of *coocurrence* depends on task, but here: "within context window".

Sharon Goldwater

Distributional semantics

A problem with PMI

- In practice, PMI is computed with counts (using MLE)
- Result: it is over-sensitive to the chance co-occurrence of infrequent words
- See next slide: ex. PMIs from bigrams with 1 count in 1st 1000 documents of NY Times corpus



Distributional semantics

11

Example PMIs (Manning & Schütze, 1999, p181)

I_{1000}	w^1	w^2	$w^1 w^2$	Bigram
16.95	5	1	1	Schwartz eschews
15.02	1	19	1	fewest visits
13.78	5	9	1	FIND GARDEN
12.00	5	31	1	Indonesian pieces
9.82	26	27	1	Reds survived
9.21	13	82	1	marijuana growing
7.37	24	159	1	doubt whether
6.68	687	9	1	new converts
6.00	661	15	1	like offensive
3.81	159	283	1	must think

Distributional semantics

10

Alternatives to PMI for finding collocations

- There are a lot, all ways of measuring statistical (in)dependence.
 - Student t-test
 - Pearson's χ^2 statistic
 - Dice coefficient
 - likelihood ratio test (Dunning, 1993)
 - Lin association measure (Lin, 1998)
 - and many more...
- Of those listed here, Dunning LR test probably most reliable for low counts.
- However, which works best may depend on particular application/evaluation.

How to measure similarity

- So, let's assume we have context vectors for two words \vec{v} and \vec{w}
- Each contains PMI values for all context words
- One way to think of these vectors: as points in high-dimensional space

Sharon Goldwater	Distributional semantics	14 Sh	aron Goldwater	Distributional semantics 15	
	How to measure similarity			Vector space representation	
• So, let's assum	e we have context vectors for two words $ec{v}$ and $ec{w}$	•	• Ex. in 2-dim space: $cat = (v_1, v_2), computer = (w_1, w_2)$		
• Each contains l	PMI values for all context words			dog • cat	
• One way to thi	nk of these vectors: as points in high-dimensional space			•	
SVD) to cre	often use dimensionality reduction methods (PCA, LS ate a more compact (but still high-dim!) representation wh stances as much as possible.				
Sharon Goldwater	Distributional semantics	<u>16 Sh</u>	aron Goldwater	Distributional semantics 17	
	Euclidean distance			Dot product	
• We could meas	ure (dis)similarity using Euclidean distance: $\left(\sum_i (v_i-w_i)^2\right)^2$	1/2 •	Another possib	ility: take the dot product of $ec v$ and $ec w$:	
	dog • cat Euclidean			$sim_{DP}(\vec{v}, \vec{w}) = \vec{v} \cdot \vec{w}$ $= \sum_{i} v_{i} w_{i}$	
• But doesn't wo	ork well if even one dimension has an extreme value		large value to $-$ When v_i is	d w_i are both large (share a context word), this contributes a o the sum. large but w_i is small (inconsistent contexts), this does not such to the sum.	
Sharon Goldwater	Distributional semantics	18 Sh	aron Goldwater	Distributional semantics 19	
	Normalized dot product			Normalized dot product	
• Some vectors a	re longer than others (have higher values):	•	Some vectors a	re longer than others (have higher values):	
[[5, 2.3, 0, 0.2, 2.1] vs. [0.1, 0.3, 1, 0.4, 0.1]			[5, 2.3, 0, 0.2, 2.1] vs. $[0.1, 0.3, 1, 0.4, 0.1]$	
	ontext word counts, these will be <i>frequent</i> words 2MI values, these are likely to be <i>infrequent</i> words			ontext word counts, these will be <i>frequent</i> words MI values, these are likely to be <i>infrequent</i> words	
• Dot product is	generally larger for longer vectors, regardless of similarity	•	Dot product is	generally larger for longer vectors, regardless of similarity	
		•	To correct for t	his, we normalize : divide by the length of each vector:	
				$sim_NDP(ec v, ec w) = (ec v \cdot ec w)/(ec v ec w)$	

Normalized dot product = cosine

• The normalized dot product is just the cosine of the angle between vectors.



Other similarity measures

- Again, many alternatives
 - Jaccard measure
 - Dice measure
 - Jenson-Shannon divergence
 - etc.
- Again, may depend on particular application/evaluation
- Ranges from -1 (vectors pointing opposite directions) to 1 (identical vectors)

Sharon Goldwater Distri	butional semantics 22	Sharon Goldwater	Distributiona	l semantics	23
How do we evaluati	on these representations?		Relatedness	judgments	
 We can use task-based evaluation information retrieval 	on: use the representations in a system, e.g.,	 Participants an concepts? 	re asked, e.g.: on a scal	e of 1-10, how related are the	following
 question answering automatic essay grading 			LEMON	FLOWER	
• Or we can evaluate against human judgements, e.g.,		Usually given	some examples initially	to set the scale , e.g.	
 relatedness judgments word association 		– LEMON-T – LEMON-O	RUTH = 1 $RANGE = 10$		

Sharon Goldwater

24

26

• But still a funny task, and answers depend a lot on how the question is asked ('related' vs. 'similar' vs. other terms)

Sharon Goldwater

Distributional semantics Word association

- Participants see/hear a word, say the first word that comes to mind
- Data collected from lots of people provides probabilities of each answer:

		ORANGE	0.16
		SOUR	0.11
		TREE	0.09
LEMON	\Rightarrow	YELLOW	0.08
		TEA	0.07
		JUICE	0.05

Example data from the Edinburgh Associative Thesaurus: http://www.eat.rl.ac.uk/

Sharon Goldwater

Current work: neural networks

Distributional semantics

- Another method for learning vector space representations
- Recent work has argued these representations capture important linguistic regularities, not just similarity (Mikolov et al., 2013)



Comparing to human data

• Human judgments provide a ranked list of related words/associations for each word w

Distributional semantics

- Computer system provides a ranked list of most similar words to w
- Compute the Spearman rank correlation between the lists (how well do the rankings match?)
- Often report on several data sets, as their details differ

Distributional semantics

27

25

Current work: compositionality

- One definition of collocations: non-compositional phrases
 - White House: not just a house that is white
 - barn raising: involves more than the parts imply
- But a lot of language is compositional
 - red barn: just a barn that is red
 - wooden plank: nothing special here
- Can we capture compositionality in a vector space model?

Sharon Goldwater

Compositionality in a vector space

 $\bullet\,$ More formally, compositionality implies some operator $\oplus\,$ such that

 $meaning(w_1w_2) = meaning(w_1) \oplus meaning(w_2)$

- Current work investigates possible operators
- vector addition (doesn't work very well)
- tensor product
- nonlinear operations learned by neural networks
- \bullet One problem: words like not—more like operators than points in space.

References

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Sharon Goldwater

Distributional semantics

30

Sharon Goldwater

Distributional semantics

31