WORD SIMILARITY

David Kauchak
CS159 Fall 2024

1

Admin

Assignment 4

Grading

Quiz #2 Thursday
- 45 minutes
- Open book and notes

Assignment 5
- Two part assignment
- A due Thursday after fall break
- Have a proper fall break!
- B due a week later

2

Quiz #2

Topics
- Linguistics 101
- Parsing
  - Grammars, CFGs, PCFGs
  - Top-down vs. bottom-up
  - CKY algorithm
  - Grammar learning
  - Evaluation
  - Improved models

3

Text Similarity

A common question in NLP is how similar are texts

score: sim(,) = ?

rank: ?

4

Bag of words representation

For now, let's ignore word order:

Obama said banana repeatedly last week on tv, "banana, banana, banana"

banana	obama	said	california	across	tv	wrong	capital	
4	1	1	0	0	1	0	0	…

"Bag of words representation": multi-dimensional vector, one dimension per word in our vocabulary

Frequency of word occurrence

5

Vector based word

A

a1: When	1
a2: the	2
a3: defendant	1
a4: and	1
a5: courthouse	0
…	

B

b1: When	1
b2: the	2
b3: defendant	1
b4: and	0
b5: courthouse	1
…	

Multi-dimensional vectors, one dimension per word in our vocabulary

6

TF-IDF

One of the most common weighting schemes

TF = term frequency

IDF = inverse document frequency

$$a'_i = \underbrace{a_i}_{\text{TF}} \times \underbrace{\log N / \mathrm{df}_i}_{\text{IDF (word importance weight)}}$$

We can then use this with any of our similarity measures!

7

Normalized distance measures

Cosine

$$sim_{\cos}(A,B) = A \cdot B = \sum_{i=1}^{n} a'_i b'_i = \frac{\sum_{i=1}^{n} a_i b_i}{\boxed{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}}$$

L2

$$dist_{L2}(A,B) = \sqrt{\sum_{i=1}^{n} (a'_i - b'_i)^2}$$

L1

$$dist_{L1}(A,B) = \sum_{i=1}^{n} |a'_i - b'_i|$$

a' and b' are length normalized versions of the vectors

8

Our problems

Which of these have we addressed?

- word order
- length
- synonym
- spelling mistakes
- word importance
- word frequency

A model of word similarity!

9

Word overlap problems

A: When the defendant and his *lawyer* walked into the *court*, some of the victim supporters *turned* their backs to him.

B: When the defendant walked into the *courthouse* with his *attorney*, the crowd *truned* their backs on him.

10

Word similarity

How similar are two words?

score: $sim(w_1, w_2) = ?$

rank: w ? w_1 w_2 w_3

applications?

list: w_1 and w_2 are synonyms

11

Word similarity applications

General text similarity

Thesaurus generation

Automatic evaluation

Text-to-text

- paraphrasing
- summarization
- machine translation

information retrieval (search)

12

Word similarity

How similar are two words?

score: $sim(w_1, w_2) = ?$

rank: w ? w_1 w_2 w_3

ideas? useful resources?

list: w_1 and w_2 are synonyms

13

Word similarity

Four categories of approaches (maybe more)

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

14

Character-based similarity

sim(*turned*, *truned*) = ?

How might we do this using only the words (i.e. no outside resources?

15

Edit distance (Levenshtein distance)

The edit distance between w_1 and w_2 is the minimum number of operations to transform w_1 into w_2

Operations:

- insertion
- deletion
- substitution

EDIT(turned, truned) = ?
EDIT(computer, commuter) = ?
EDIT(banana, apple) = ?
EDIT(wombat, worcester) = ?

16

Edit distance

EDIT(turned, truned) = 2
- delete u
- insert u

EDIT(computer, commuter) = 1
- replace p with m

EDIT(banana, apple) = 5
- delete b
- replace n with p
- replace a with p
- replace n with l
- replace a with e

EDIT(wombat, worcester) = 6

17

Better edit distance

Are all operations equally likely?
- No

Improvement: give different weights to different operations
- replacing a for e is more likely than z for y

Ideas for weightings?
- Learn from actual data (known typos, known similar words)
- Intuitions: phonetics
- Intuitions: keyboard configuration

18

Vector character-based word similarity

sim(*turned*, *truned*) = ?

Any way to leverage our vector-based similarity approaches from last time?

19

Vector character-based word similarity

sim(*turned*, *truned*) = ?

a:	0		a:	0
b:	0		b:	0
c:	0		c:	0
d:	1		d:	1
e:	1		e:	1
f:	0		f:	0
g:	0		g:	0
…			…	

Generate a feature vector based on the characters
(or could also use the set based measures at the character level)

problems?

20

Vector character-based word similarity

sim(*restful*, *fluster*) = ?

Character level loses a lot of information

a:	0	a:	0
b:	0	b:	0
c:	0	c:	0
d:	1	d:	1
e:	1	e:	1
f:	0	f:	0
g:	0	g:	0
…		…	

ideas?

21

Vector character-based word similarity

sim(*restful*, *fluster*) = ?

Use character bigrams or even trigrams

aa:	0	aa:	0
ab:	0	ab:	0
ac:	0	ac:	0
…		…	
es:	1	er:	1
…		…	
fu:	1	fl:	1
…		…	
re:	1	lu:	1
…		…	

22

Word similarity

Four general categories
- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

23

WordNet

Lexical database for English
- 155,287 words
- 206,941 word senses
- 117,659 synsets (synonym sets)
- ~400K relations between senses
- Parts of speech: nouns, verbs, adjectives, adverbs

Word graph, with word senses as nodes and edges as relationships

Psycholinguistics
- WN attempts to model human lexical memory
- Design based on psychological testing

Created by researchers at Princeton
- http://wordnet.princeton.edu/

Lots of programmatic interfaces

24

WordNet relations

- synonym
- antonym
- hypernyms
- hyponyms
- holonym
- meronym
- troponym
- entailment
- (and a few others)

25

WordNet relations

synonym – X and Y have similar meaning

antonym – X and Y have opposite meanings

hypernyms – subclass
- beagle is a hypernym of dog

hyponyms – superclass
- dog is a hyponym of beagle

holonym – contains part
- car is a holonym of wheel

meronym – part of
- wheel is a meronym of car

26

WordNet relations

troponym – for verbs, a more specific way of doing an action
- run is a troponym of move
- dice is a troponym of cut

entailment – for verbs, one activity leads to the next
- sleep is entailed by snore

(and a few others)

27

WordNet

Graph, where nodes are words and edges are relationships

There is some hierarchical information, for example with hyp-er/o-nomy

28

WordNet: run

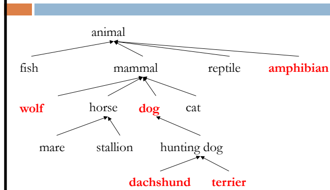
29

WordNet: run

- S: (n) **run**, tally (a score in baseball made by a runner touching all four bases safely) *"the Yankees scored 3 runs in the bottom of the 9th"; "their first tally came in the 3rd inning"*
  - ***direct hyponym*** / *full hyponym*
    - S: (n) earned run (a run that was not scored as the result of an error by the other team)
    - S: (n) unearned run (a run that was scored as a result of an error by the other team)
    - S: (n) run batted in, rbi (a run that is the result of the batter's performance) *"he had more than 100 rbi last season"*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

30

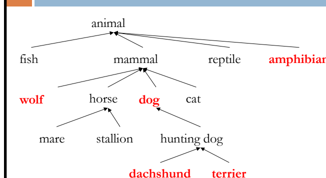
WordNet-like Hierarchy



To utilize WordNet, we often want to think about some graph-based measure.

31

WordNet-like Hierarchy

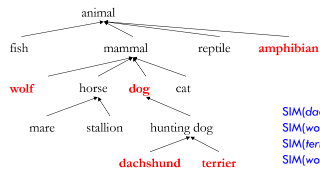


Rank the following based on similarity:

SIM(*wolf*, *dog*)
SIM(*wolf*, *amphibian*)
SIM(*terrier*, *wolf*)
SIM(*dachshund*, *terrier*)

32

WordNet-like Hierarchy

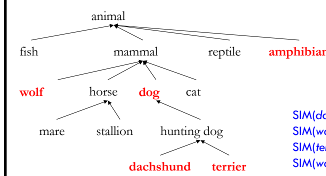


SIM(*dachshund*, *terrier*)
SIM(*wolf*, *dog*)
SIM(*terrier*, *wolf*)
SIM(*wolf*, *amphibian*)

What information/heuristics did you use to rank these?

33

WordNet-like Hierarchy



SIM(*dachshund*, *terrier*)
SIM(*wolf*, *dog*)
SIM(*terrier*, *wolf*)
SIM(*wolf*, *amphibian*)

- path length is important (but not the only thing)
- words that share the same ancestor are related
- words lower down in the hierarchy are finer grained and therefore closer

34

WordNet similarity measures

path length doesn't work very well

Some ideas:
- path length scaled by the depth (Leacock and Chodorow, 1998)

With a little cheating:
- Measure the "information content" of a word using a corpus: how specific is a word?
  - words higher up tend to have less information content
  - more frequent words (and ancestors of more frequent words) tend to have less information content

35

WordNet similarity measures

Utilizing information content:
- information content of the lowest common parent (Resnik, 1995)
- information content of the words minus information content of the lowest common parent (Jiang and Conrath, 1997)
- information content of the lowest common parent divided by the information content of the words (Lin, 1998)

36

Word similarity

Four general categories

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

37

Dictionary-based similarity

Word	Dictionary blurb
aardvark	a large, nocturnal, burrowing mammal, *Orycteropus afer, ofcentral and southern Africa, feeding on ants and termites andhaving a long, extensile tongue, strong claws, and long ears.*
beagle	One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat.
dog	Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.

38

Dictionary-based similarity

Utilize our text similarity measures

sim(*dog*, *beagle*) =

sim(One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat. ,

Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.)

39

Dictionary-based similarity

–noun

1. a domesticated canid, *Canis familiaris,* bred in many varieties.

2. any carnivore of the dogfamily Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.

3. the male of such an animal.

4. any of various animals resembling a dog.

5. a despicable man or youth.

6. *Informal* . a fellow in general: *a lucky dog.*

7. dogs, *Slang* . feet.

8. *Slang* .
- **a.** something worthless or of extremely poor quality: *That used car you bought is a dog.*
- **b.** an utter failure; flop: *Critics say his new play is a dog.*

9. *Slang* . an ugly, boring, or crude person.

10. *Slang* . hot dog.

11. (*initial capital letter*) *Astronomy* . either of two constellations, Canis Major or Canis Minor.

12. *Machinery* .
- **a.** any of various mechanical devices, as for gripping or holding something.
- **b.** a projection on a moving part for moving steadily or for tripping another part with which it engages.

13. Also called **gripper**, **nipper.** *Metalworking* . a device on a drawbench for drawing the work through the die.

14. a cramp binding together two timbers.

15. an iron bar driven into a stone or timber to provide a means of lifting it.

16. an andiron; firedog.

17. *Meteorology* . a sundog or fogdog.

18. a word formerly used in communications to represent the letter D.

What about words that have multiple senses/parts of speech?

40

Dictionary-based similarity

–noun

1. a domesticated canid, *Canis familiaris,* bred in many varieties.
2. any carnivore of the dogfamily Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.
3. the male of such an animal.
4. any of various animals resembling a dog.
5. a despicable man or youth.
6. *Informal* . a fellow in general: *a lucky dog.*
7. dogs, *Slang* . feet.
8. *Slang* .
   a. something worthless or of extremely poor quality: *That used car you bought is a dog.*
   b. an utter failure; flop: *Critics say his new play is a dog.*
9. *Slang* . an ugly, boring, or crude person.
10. *Slang* . hot dog.
11. (*initial capital letter*) *Astronomy* . either of two constellations, Canis Major or Canis Minor.
12. *Machinery* .
    a. any of various mechanical devices, as for gripping or holding something.
    b. a projection on a moving part for moving steadily or for tripping another part with which it engages.
13. Also called **gripper**, **nipper.** *Metalworking* . a device on a drawbench for drawing the work through the die.
14. a cramp binding together two timbers.
15. an iron bar driven into a stone or timber to provide a means of lifting it.
16. an andiron; firedog.
17. *Meteorology* . a sundog or fogdog.
18. a word formerly used in communications to represent the letter D.

1. part of speech tagging
2. word sense disambiguation
3. most frequent sense
4. average similarity between all senses
5. max similarity between all senses
6. sum of similarity between all senses

41

Dictionary + WordNet

WordNet also includes a "gloss" similar to a dictionary definition

Other variants include the overlap of the word senses as well as those word senses that are related (e.g. hypernym, hyponym, etc.)

- incorporates some of the path information as well
- Banerjee and Pedersen, 2003

42

Word similarity

Four general categories

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

43

Corpus-based approaches

Word — **ANY** blurb with the word

aardvark

beagle

dog

Ideas?

44

Corpus-based

The **Beagle** is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg

Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern **Beagle** can be traced in Ancient Greece[2] back to around the 5th century BC.

From medieval times, **beagle** was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard **Beagle** type was beginning to develop: the distinction between the North Country Beagle and Southern

45

Corpus-based: feature extraction

The **Beagle** is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg

We'd like to utilize our vector-based approach

How could we we create a vector from these occurrences?

- collect word counts from all documents with the word in it
- collect word counts from all sentences with the word in it
- collect all word counts from all words within ***X*** words of the word
- collect all words counts from words in specific relationship: subject-object, etc.

46

Word-context co-occurrence vectors

The **Beagle** is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg

Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern **Beagle** can be traced in Ancient Greece[2] back to around the 5th century BC.

From medieval times, **beagle** was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard **Beagle** type was beginning to develop: the distinction between the North Country Beagle and Southern

47

Word-context co-occurrence vectors

The **Beagle** is a breed

Beagles are intelligent, and

to the modern **Beagle** can be traced

From medieval times, **beagle** was used as

1840s, a standard **Beagle** type was beginning

the:	2
is:	1
a:	2
breed:	1
are:	1
intelligent:	1
and:	1
to:	1
modern:	1
…	

Often do some preprocessing like lowercasing and removing stop words

48

Corpus-based similarity

sim(*dog*, *beagle*) =

sim(*context_vector(dog)*, *context_vector(beagle)*)

the:	5
is:	1
a:	4
breeds:	2
are:	1
intelligent:	5
…	

the:	2
is:	1
a:	2
breed:	1
are:	1
intelligent:	1
and:	1
to:	1
modern:	1
…	

49

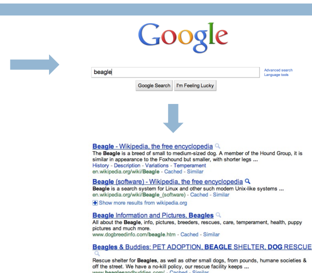
Web-based similarity

Ideas?

50

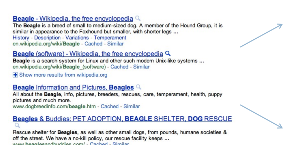
Web-based similarity

beagle



51

Web-based similarity



Concatenate the snippets for the top *N* results

Concatenate the web page text for the top *N* results

52

Another feature weighting

TF- IDF weighting takes into account the general importance of a feature

For distributional similarity, we have the feature (f_i), but we also have the word itself (w) that we can use for information

sim(*context_vector*(**dog**), *context_vector*(**beagle**))

the:	5	the:	2
is:	1	is:	1
a:	4	a:	2
breeds:	2	breed:	1
are:	1	are:	1
intelligent:	5	intelligent:	1
…		and:	1
		to:	1
		modern:	1
		…	

53

Another feature weighting

Feature weighting ideas given this additional information?

sim(*context_vector*(**dog**), *context_vector*(**beagle**))

the:	5	the:	2
is:	1	is:	1
a:	4	a:	2
breeds:	2	breed:	1
are:	1	are:	1
intelligent:	5	intelligent:	1
…		and:	1
		to:	1
		modern:	1
		…	

54

Another feature weighting

count *how likely* feature f_i and word w are to occur together

- incorporates co-occurrence
- but also incorporates how often w and f_i occur in other instances

sim(*context_vector*(**dog**), *context_vector*(**beagle**))

Does IDF capture this?

Not really. IDF only accounts for f_i regardless of w

55

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

When will this be high and when will this be low?
What happens if x and y are independent/dependent?

56

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if x and y are independent (i.e. one occurring doesn't impact the other occurring) then:

$$p(x,y) =$$

57

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if x and y are independent (i.e. one occurring doesn't impact the other occurring) then:

$$p(x,y) = p(x)p(y)$$

What does this do to the sum?

58

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if they are dependent then:

$$p(x,y) = p(x)p(y \mid x) = p(y)p(x \mid y)$$

$$\Downarrow$$

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(y \mid x)}{p(y)}$$

59

Mutual information

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(y \mid x)}{p(y)}$$

What is this asking?
When is this high?

How much more likely are we to see y given x has a particular value!

60

Point-wise mutual information

Mutual information

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

How related are two variables (i.e. over all possible values/events)

Point-wise mutual information

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

How related are two particular events/values

61

PMI weighting

Mutual information is often used for feature selection in many problem areas

PMI weighting weights co-occurrences based on their correlation (i.e. high PMI)

context_vector(beagle)

the:	2
is:	1
a:	2
breed:	1
are:	1
intelligent:	1
and:	1
to:	1
modern:	1
…	

$\log \frac{p(beagle,the)}{p(beagle)p(the)}$

$\log \frac{p(beagle,breed)}{p(beagle)p(breed)}$

How do we calculate these?

62