Computational Linguistics

10. Semantic Similarity and Word Sense Disambiguation

Xiaojing Bai

Tsinghua University

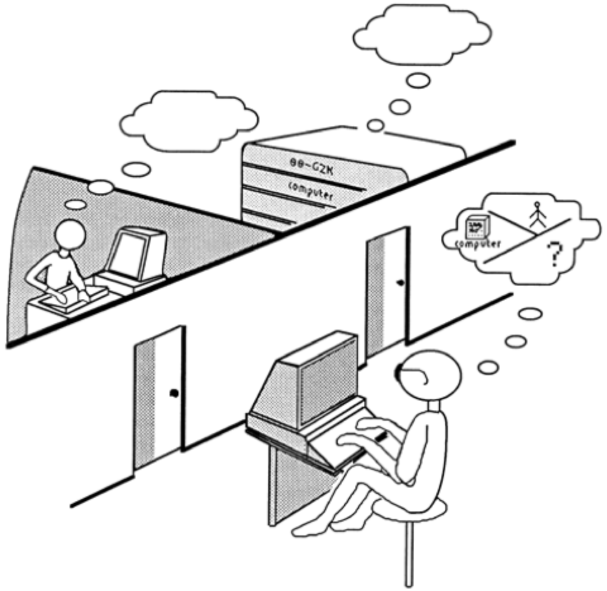
https://bxjthu.github.io/CompLing

Recap: Can a computer understand the meaning of a sentence?

- How could we tell if it did?
- Alan Turing: Can a computer think?

  Internal states vs.observable behaviors
- Natural language understanding

  Using observable behaviors to judge the capacity
- Computational approaches to natural language



Recap: meaning representation

First-order logic as a meaning representation language

- Basic elements
- Variables and quantifiers
- Lambda notation
- The semantics of FOL
- Event and state representations

Some notes: Broadly speaking, logic-based approaches to natural language semantics focus on those aspects of natural language which guide our judgments of consistency and inconsistency. The syntax of a logical language is designed to make these features formally explicit. As a result, determining properties like consistency can often be reduced to symbolic manipulation, that is, to a task that can be carried out by a computer.

Recap: representing the meaning of a word

- Dictionary entries
- Feature structures
- Relational databases
- Trees
- Synsets
- Vectors

Recap: representing the meaning of a word

- Lexical semantics
  Different aspects of word meaning: word senses, word similarity and relatedness, lexical fields and frames, connotation, etc.

- Vector semantics
  Learning computational representations of the meaning of words directly from their distributions in text

Recap: vectors as meaning representations

What's the difference?

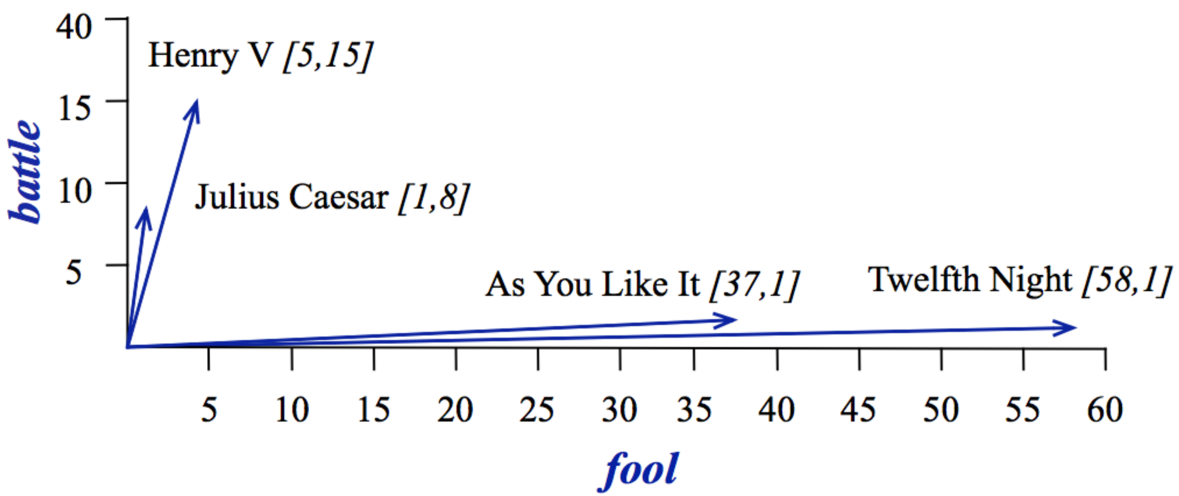
	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

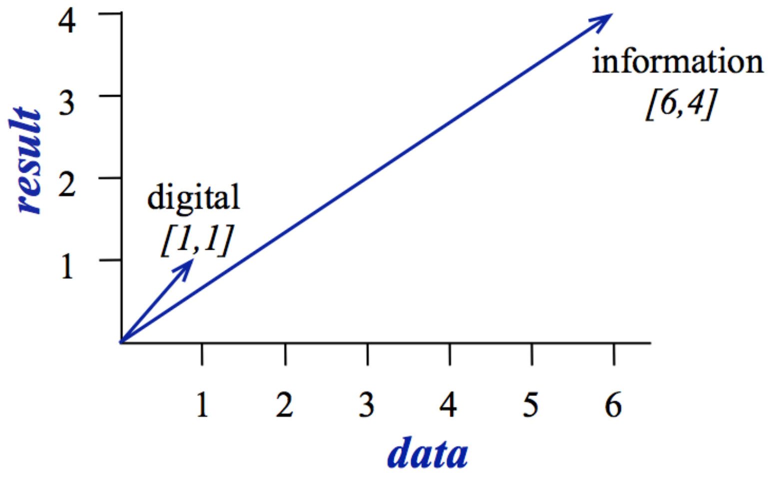
	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

Recap: vectors as meaning representations

What's the difference?



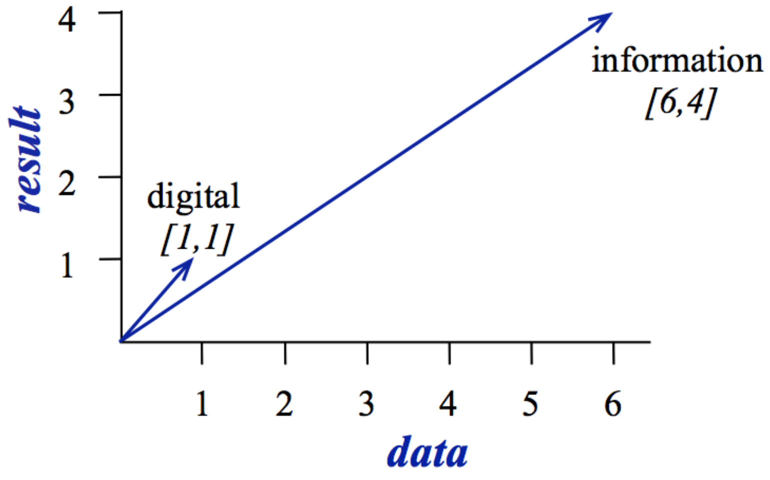


At the end of this session you will

- know about how word similarity can be measured
- understand what are word senses and the possible relations between them
- understand how word senses are defined in WordNet
- understand the goal and applications of WSD
- know about the types of algorithms for WSD

Cosine for measuring word similarity

- Word similarity as vector similarity
- Vector similarity as the cosine of the angle between the vectors: $\cos(\vec{v}, \vec{w})$



Cosine for measuring word similarity

$\text{dot-product}(\vec{v}, \vec{w}) = \vec{v} \cdot \vec{w}$

- Algebraic definition: $\vec{v} \cdot \vec{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + ... + v_N w_N$
- Geometric definition: $\vec{v} \cdot \vec{w} = |\vec{v}||\vec{w}| \cos\theta$

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

Cosine for measuring similarity: an example

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

	large	**data**	**computer**
apricot	2	0	0
digital	0	1	2
information	1	6	1

$$\cos(\text{apricot,information}) = \frac{2+0+0}{\sqrt{4+0+0}\sqrt{1+36+1}} = \frac{2}{2\sqrt{38}} = .16$$

$$\cos(\text{digital,information}) = \frac{0+6+2}{\sqrt{0+1+4}\sqrt{1+36+1}} = \frac{8}{\sqrt{38}\sqrt{5}} = .58$$

Food for your thought

Is the simple frequency of co-occurrence the best measure of association between words?

Are word vectors based on raw frequencies informative and discriminative enough to represent word meaning?

The rationale behind the tf-idf vector model

	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

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

The tf-idf weighting of the value for word *t* in document *d*:

$w_{t,d}$ = term frequency × inverse document frequency

PPMI: measure of the association between words

- Simple frequency isn't the best measure!

  Words that are frequent but not informative or discriminative: *the*, *it*, *they*

- Positive Pointwise Mutual Information (PPMI)

  How much more are the two words co-occurring in our corpus than we would have a priori expected them to appear by chance?

$$PMI(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

$$PPMI(w,c) = max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

Word senses

carpet vs. carpets

sing vs. sing, sang, sung

Word senses

carpet vs. carpets

sing vs. sing, sang, sung

- Lemma or citation form: the grammatical form of a word that is used to represent a word in dictionaries and thesaurus
- Wordform: the full inflected or derived form of a lemma

Word senses

carpet vs. carpets

sing vs. sing, sang, sung

- Lemma or citation form: the grammatical form of a word that is used to represent a word in dictionaries and thesaurus
- Wordform: the full inflected or derived form of a lemma

Word sense : a discrete representation of one aspect of the meaning of a lemma

E.g.

Instead, a bank can hold the investments in a custodial account in the client's name.

But as agriculture burgeons on the east bank, the river will shrink even more.

Relations between the senses of a word

- Senses coincidentally sharing an orthographic form but not related
- Related terms
  - Homonym: e.g. bank ("financial institution") vs. bank ("sloping mound")
  - Homonymy
  - Homograph: e.g. bank; bat("club for hitting a ball") vs. bat ("nocturnal flying animal")
  - Homophone: e.g. write vs. right; piece vs. peace
- Related problems for NLP

Relations between the senses of a word

- Senses semantically related
- Related terms
  - Polysemy: e.g. bank, school, university, hospital
  - Metonymy: e.g. the White House, Jane Austen, Plums

Relations between senses (rather than words)

- Synonymy
- Antonymy
- Hypernymy
- Hyponymy
- Meronymy

How to define the meaning of a word sense?

Examples from *American Heritage Dictionary* (Morris, 1985)

right	*adj.*	located nearer the right hand esp. being on the right when facing the same direction as the observer
left	*adj.*	located nearer to this side of the body than the right
red	*n.*	the color of blood or a ruby
blood	*n.*	the red liquid that circulates in the heart, arteries and veins of animals

WordNet: a database of lexical relations

- Defining a sense through its relationship with other senses
- The most commonly used resource for English sense relations
- Three databases: 1) nouns, 2) verbs, 3) adjectives and adverbs
- Representing a concept in logical terms vs. represents a concept as a list of the word senses that can be used to express the concept
- WordNet 3.1

Noun relations in WordNet

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$\textit{breakfast}^1 \rightarrow \textit{meal}^1$
Hyponym	Subordinate	From concepts to subtypes	$\textit{meal}^1 \rightarrow \textit{lunch}^1$
Instance Hypernym	Instance	From instances to their concepts	$\textit{Austen}^1 \rightarrow \textit{author}^1$
Instance Hyponym	Has-Instance	From concepts to concept instances	$\textit{composer}^1 \rightarrow \textit{Bach}^1$
Member Meronym	Has-Member	From groups to their members	$\textit{faculty}^2 \rightarrow \textit{professor}^1$
Member Holonym	Member-Of	From members to their groups	$\textit{copilot}^1 \rightarrow \textit{crew}^1$
Part Meronym	Has-Part	From wholes to parts	$\textit{table}^2 \rightarrow \textit{leg}^3$
Part Holonym	Part-Of	From parts to wholes	$\textit{course}^7 \rightarrow \textit{meal}^1$
Substance Meronym		From substances to their subparts	$\textit{water}^1 \rightarrow \textit{oxygen}^1$
Substance Holonym		From parts of substances to wholes	$\textit{gin}^1 \rightarrow \textit{martini}^1$
Antonym		Semantic opposition between lemmas	$\textit{leader}^1 \iff \textit{follower}^1$
Derivationally Related Form		Lemmas w/same morphological root	$\textit{destruction}^1 \iff \textit{destroy}^1$

Verb relations in WordNet

Relation	**Definition**	**Example**
Hypernym	From events to superordinate events	*fly*^{9} $\rightarrow$ *travel*^{5}
Troponym	From events to subordinate event (often via specific manner)	*walk*^{1} $\rightarrow$ *stroll*^{1}
Entails	From verbs (events) to the verbs (events) they entail	*snore*^{1} $\rightarrow$ *sleep*^{1}
Antonym	Semantic opposition between lemmas	*increase*^{1} $\Longleftrightarrow$ *decrease*^{1}
Derivationally Related Form	Lemmas with same morphological root	*destroy*^{1} $\Longleftrightarrow$ *destruction*^{1}

More details about WordNet

- WordNet online
- Database statistics
- A glossary of WordNet terms
- Five Papers on WordNet
- Frequently Asked Questions
- Wordnet with NLTK

Measuring word similarity with WordNet

They didn't have newspapers, books and even cell phones to transmit their viewpoints like we do.

- A fundamental task for semantic models is to predict how similar two words' meanings are
- Applications: query expansion, learning sentiment lexicons, paraphrasing…
- Thesaurus methods
- Goal: measure how close the two target words are within the hierarchy
- WordNet::Similarity
- Computing semantic similarity with NLTK

Word sense disambiguation (WSD)

- Lexical ambiguity and an avalanche of competing interpretations
- WSD: the task of selecting the correct sense for a word
  - Input: a word in context along with a fixed inventory of potential senses
  - Output: the correct word sense for that use

  Reports said the plant was likely to close in December, leaving many jobless.

  plant 1: leafy green organism
  plant 2: equipment and fixtures for manufacturing

- Applications

Types of algorithms for WSD

- Supervised

  We know the answers for many examples and can use them to learn from their (automatically determinable) characteristics. We then apply the learned model to a comparable set of examples (not the same ones).

- **Weakly supervised (knowledge-based)**

  We start with no known answers, but we use secondary texts (dictionary glosses) to infer underlying relationships through the Lesk algorithm.

- Semi-supervised

  We know the answers for a small number of examples, and can gain more examples from the data by finding similar cases and inferring the answers they should have through bootstrapping.

- Unsupervised

  We start with no known answers, and no predefined senses. The set of “senses” is created automatically from the instances of each word in the training set.

Weakly supervised or knowledge-based WSD

- Indirect supervision
- Knowledge from dictionaries, thesauruses or similar knowledge bases
- The original Lesk algorithm (Lesk, 1986)
- The simplified Lesk algorithm (Kilgarriff and Rosenzweig, 2000)

The simplified Lesk algorithm

```
function SIMPLIFIED LESK(word, sentence) returns best sense of word
  best-sense ← most frequent sense for word
  max-overlap ← 0
  context ← set of words in sentence
  for each sense in senses of word do
    signature ← set of words in the gloss and examples of sense
    overlap ← COMPUTEOVERLAP(signature, context)
    if overlap > max-overlap then
        max-overlap ← overlap
        best-sense ← sense
  end
  return(best-sense)
```

The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list.

The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

bank1

Gloss: a financial institution that accepts deposits and channels the money into lending activities

Examples: *"he cashed a check at the bank"*, *"that bank holds the mortgage on my home"*

bank2

Gloss: sloping land (especially the slope beside a body of water)

Examples: *"they pulled the canoe up on the bank"*, *"he sat on the bank of the river and watched the currents"*

The simplified Lesk algorithm

- Choosing the sense whose dictionary gloss or definition shares the most words with the target word's neighborhood

- The original Lesk algorithm (Lesk, 1986)

  Choosing the sense whose dictionary gloss or definition shares the most words with the dictionary glosses or definitions of the target word's neighborhood

pine	1	kinds of evergreen tree with needle-shaped leaves
	2	waste away through sorrow or illness
cone	1	solid body which narrows to a point
	2	something of this shape whether solid or hollow
	3	fruit of certain evergreen trees

Two Python implementations of the Lesk algorithms

http://www.nltk.org/howto/wsd.html

https://github.com/alvations/pywsd

```
>>> wn.synset('car.n.01').definition()
'a motor vehicle with four wheels; usually propelled by an internal combustion engine'

>>> wn.synset('car.n.01').examples()
['he needs a car to get to work']
```

At the end of this session you will

- know about how word similarity can be measured
- understand what are word senses and the possible relations between them
- understand how word senses are defined in WordNet
- understand the goal and applications of WSD
- know about the types of algorithms for WSD

Homework

- Read/Review
  - J+M_6 (6.1-6.7)
  - J+M_C
- Practice
  - http://www.nltk.org/book/ch02.html#wordnet

Next session

Semantic Role Labeling and Computational Discourse