Intro to Text Processing
Lecture 8

Behrang Mohit

Some slides are borrowed or adapted from Noah Smith, Dan Jurafsky and Chris Manning lectures.

So far ...

• Talked about:
  – Text processing
    • IE, Doc classification
  – Basics of linguistics
  – Syntax
    • POS tagging
Words as symbols

• that empty sky was full of light

• that/DT empty/ADJ sky/Noun was/VBD full/ADJ of/IN light/NN

• 22/DT 59/ADJ 103/Noun 4/VBD 58/ADJ 9/IN 97/NN

Word Relationships

• We want computers to know how lexical items (words) like *giraffe* are related to each other.
  — e.g., so we can infer:
    • saw (Sara, giraffe) ➔ saw(Sara, animal)
  — Disambiguate:
    • In my office there are two tables and one *chair*.
    • This year the *chair* of conference is Prof. Ali.
Learning words

• Computers have no view of the world

Learning words

• No understanding of dictionaries, Wikipedia, scientific text

• Understand symbols, statistics and graph structures
  – Lexical database
Lexical semantics

• Syntax
  – Word-level: POS tagging
  – Sentence-level: Parse Trees

• Semantics (meaning)
  – **Lexical Semantics**: Study of the meaning of words and the connections between words

Words and their Relations

• **Lexical item** or a *word form*: One or more words which form the basic elements of a language’s lexicon (vocabulary)
  – Horse
  – traffic light

• Sense: a meaning of a word form
  – Chair: an object used for sitting
  – Chair: The head of a group of humans
Lexical Matrix

- Many:Many mappings between word forms and meanings.
- F1 and F2 are *synonyms*.

Lexical Matrix

- Chair and sofa are synonyms
- Word Relations
  - Synonymy
  - Antonymy
  - [a-z]+nymy
Synonyms

• Synonym: A word that can be substituted for another without changing the meaning of the sentence.
  – e.g. Drive (Mariem, car) vs. Drive(Mariem, automobile).
  – Cases were words are not quite synonyms
    • Ali likes to drink strong tea.
    • Ali likes to drink powerful tea.
  – Genre, style, etc. influence word usage.

Synonymy and word sense

• Synonymy is really between senses of words
  – Zeinab likes fish and chips.
  – *Tariq likes fish and circuits.
  – The computer has many chips.
  – The computer has many circuits.
Antonyms

- Formal definition?
  - Not always not-x
    - *Not rich* ≠ *poor*
  - Some where at the opposite ends of the a scale

  ![Image](image.png)

- Or changes:
  - rise/fall

[a-z]+nyms

- *is-a* relationship:
  - **Hyponyms**: specifications
    - *Horse* is a hyponym of *animal*.
  - **Hypernyms**: generalization
    - *Animal* is a hypernym of *horse*.

- Part-whole relationship
  - Meronym
    - Wheel is a meronym of car
  - Holonym
    - Car is a holonym of wheel
Hyponymy and Hypernymy

• One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  – *car* is a hyponym of *vehicle*
  – *mango* is a hyponym of *fruit*

• Conversely **hypernym/superordinate** (“hyper is super”)
  – *vehicle* is a hypernym of *car*
  – *fruit* is a hypernym of *mango*

Hyponymy more formally

• Extensional:
  – The class denoted by the superordinate extensionally includes the class denoted by the hyponym

• Entailment:
  – A sense A is a hyponym of sense B if *being an A* entails *being a B*

• Hyponymy is usually transitive
  – (A hypo B and B hypo C entails A hypo C)

• Another name: the **IS-A hierarchy**
  – A IS-A B (or A ISA B)
  – B subsumes A
WordNet

• Lexical database of English sense relations:
  – Nouns, Verb, Adjectives, Adverbs

• Developed at Princeton University
  – Over 20 years effort
  – Global WordNet: in many languages (e.g. Arabic)

• Organized by synsets:
  – Set of near synonyms
    • Chair, sofa
    • Chair, chairman, head

Senses of “bass” in Wordnet

Noun

• S: (n) bass (the lowest part of the musical range)
• S: (n) bass, bass part (the lowest part in polyphonic music)
• S: (n) bass, basso (an adult male singer with the lowest voice)
• S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
• S: (n) freshwater bass, bass (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
• S: (n) bass, bass voice, basso (the lowest adult male singing voice)
• S: (n) bass (the member with the lowest range of a family of musical instruments)
• S: (n) bass (nontechnical name for any of numerous edible marine and freshwater spiny–finned fishes)

Adjective

• S: (adj) bass, deep (having or denoting a low vocal or instrumental range) “a deep voice”; “a bass voice is lower than a baritone voice”; “a bass clarinet”
Synsets for dog

- S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
- S: (n) dog (informal term for a man) "you lucky dog"
- S: (n) cad, bounder, blackguard, dog, hound, heel (someone who is morally reprehensible) "you dirty dog"
- S: (n) frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
- S: (n) pawl, detent, click, dog (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
- S: (n) andiron, firedog, dog, dog-iron (metal supports for logs in a fireplace) "the andirons were too hot to touch"
- ...

WordNet Noun Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Also called</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyponym</td>
<td>Superordinate</td>
<td>From concepts to superordinates</td>
<td>breakfast₁ → meal₁</td>
</tr>
<tr>
<td>Hyponym</td>
<td>Subordinate</td>
<td>From concepts to subtypes</td>
<td>meal₁ → lunch₁</td>
</tr>
<tr>
<td>Member Meronym</td>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty₂ → professor₁</td>
</tr>
<tr>
<td>Has-Instance</td>
<td></td>
<td>From concepts to instances of the concept</td>
<td>composer₁ → Bach¹</td>
</tr>
<tr>
<td>Instance</td>
<td></td>
<td>From instances to their concepts</td>
<td>Austin⁴ → author⁴</td>
</tr>
<tr>
<td>Member Holonym</td>
<td>Member-Of</td>
<td>From members to their groups</td>
<td>copilot⁴ → crew⁴</td>
</tr>
<tr>
<td>Part Meronym</td>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td>table² → leg³</td>
</tr>
<tr>
<td>Part Holonym</td>
<td>Part-Of</td>
<td>From parts to wholes</td>
<td>course² → meal³</td>
</tr>
<tr>
<td>Antonym</td>
<td></td>
<td>Opposites</td>
<td>leader¹ → follower¹</td>
</tr>
</tbody>
</table>
WordNet 3.0

- Where it is:
  - http://wordnetweb.princeton.edu/perl/webwn

- Libraries
  - Python: WordNet from NLTK
    - http://www.nltk.org/Home
  - Java:
    - JWNL, extJWNL on sourceforge

WordNet: hypernymy/hyponymy

<table>
<thead>
<tr>
<th>Sense 1</th>
<th>door</th>
<th>(a swinging or sliding barrier that will close the entrance to a room or building; “he knocked on the door”; “he slammed the door as he left”)</th>
</tr>
</thead>
</table>
|         | movable | barrier | (a barrier that can be moved to allow passage)  
|         | barrier | (a structure or object that impedes free movement)  
|         | obstruction | impediment | (anything that makes progress difficult)  
|         | structure | construction | (a thing constructed; a complex construction or entity)  
|         | artifact | artefact | (a man-made object)  
|         | object | inanimate | object | physical object | (a nonliving entity)  
|         | entity | (something having concrete existence; living or nonliving)  

| Sense 2 | doorway, door, entree, entry, portal, room access | (the space in a wall through which you enter or leave a room or building; the space that a door can close; “he stuck his head in the doorway”)  
|---------|-------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|         | doorway | entrance | entranceway | (allows entry or exit)  
|         | access | approach | (a way of entering or leaving)  
|         | way | (any road or path affording passage from one place to another)  
|         | artifact | artefact | (a man-made object)  
|         | object | inanimate | object | physical object | (a nonliving entity)  
|         | entity | (something having concrete existence; living or nonliving)  

15383: txt proc
MeSH: Medical Subject Headings
thesaurus from the National Library of Medicine

- **MeSH (Medical Subject Headings)**
  - 177,000 entry terms that correspond to 26,142 biomedical “headings”

- **Hemoglobins**
  **Entry Terms:** Eryhem, Ferrous Hemoglobin, Hemoglobin
  **Definition:** The oxygen-carrying proteins of ERYTHROCYTES. They are found in all vertebrates and some invertebrates. The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements.

The MeSH Hierarchy

1. *Anatomy [A]*
2. *Organisms [B]*
3. *Diseases [C]*
4. - *Chemicals and Drugs [D]*
   - Inorganic Chemicals [D01] +
   - Organic Chemicals [D02] +
   - Heterocyclic Compounds [D03] +
   - Poly cyclic Compounds [D04] +
   - Macromolecular Substances [D05] +
   - Hormones, Hormone Substitutes, and Hormone /
     Enzymes and Coenzymes [D08] +
   - Carbohydrates [D09] +
   - L i p i d s [D10] +
   - Amino Acids, Peptides, and Proteins [D12] +
   - Nucleic Acids, Nucleotides, and Nucleosides [D13]
   - Complex Mixtures [D20] +
   - Biological Factors [D32] +
   - Biomedical and Dental Materials [D25] +
   - Pharmaceutical Preparations [D26] +
   - Chemical Actions and Uses [D27] +
5. *Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]*
6. *Psychiatry and Psychology [F]*
7. *Phenomena and Processes [G]*

- **Amino Acids, Peptides, and Proteins [D12]**
  - Proteins [D12.776]
    - Blood Proteins [D12.776.124]
      - Acute-Phase Proteins [D12.776.124.050] +
      - Anion Exchange Protein 1, Erythrocyte [D12.776.124.078]
    - Ankyrins [D12.776.124.080]
    - beta 2-Glycoprotein 1 [D12.776.124.117]
    - Blood Coagulation Factors [D12.776.124.125] +
    - Cholesterol Ester Transfer Proteins [D12.776.124.197]
  - Fibrin [D12.776.124.270] +
    - Glycophorin [D12.776.124.300]
    - Hemocyanin [D12.776.124.337]
  - Hemoglobins [D12.776.124.400]
    - Carboxyhemoglobin [D12.776.124.400.141]
    - Erythrocytoxins [D12.776.124.400.220]
Uses of the MeSH Ontology

- Provide synonyms (“entry terms”)
  - E.g., glucose and dextrose
- Provide hypernyms (from the hierarchy)
  - E.g., glucose ISA monosaccharide
- Indexing in MEDLINE/PubMED database
  - NLM’s bibliographic database:
    - 20 million journal articles
    - Each article hand-assigned 10-20 MeSH terms

Word Sense Disambiguation (WSD)

- Which synset does a given word (and its context) belong to?
  - In my office there are two tables and one chair.
  - This year the chair of conference is Prof. Ali.
How do we know when a word has more than one sense?

- The “zeugma” test: Two senses of serve?
  - Which flights serve breakfast?
  - Does Lufthansa serve Dubai?
  - ?Does Lufthansa serve breakfast and Dubai?
- Since this conjunction sounds weird,
  - we say that these are two different senses of “serve”

Word Sense Disambiguation (WSD)

- Which synset does a given word (and its context) belong to?
  - In my office there are two tables and one chair.
  - This year the chair of conference is Prof. Ali.

- Finding all word senses is a difficult problem:
  - We assume we have the collection of senses
    - WordNet
  - Given a sentence, choose the right sense

- Application:
  - Most areas of text processing
    - NER (e.g. Washington)
Approaches to WSD

- Dictionary-based:
  - Unsupervised (e.g. Lesk algorithm)
- Corpus-based approaches
  - Supervised algorithms (e.g. Ng & Lee 96)

Lesk Algorithm (1986)

- Only requires a dictionary
- The most likely sense for a word in a given context is decided by a measure of overlap between the definition of the target word and the definition of the words in the context of the target word.
Lesk algorithm

Given $W_1$ target word and $W_2$ in context of $W_1$

1. for each sense $i$ of $W_1$
2. for each sense $j$ of $W_2$
3. determine $\text{Overlap}_{ij}$ as the number of common occurrences between definitions of sense $i$ of $W_1$ and sense $j$ of $W_2$
4. find $i$ and $j$ for which $\text{Overlap}_{ij}$ is maximum
5. Assign sense $i$ to $W_1$ and sense $j$ to $W_2$

- This year the chair of the conference is Prof. Ali.
  - Chair: head or chief of an organization, meeting, ...
  - Conference: meeting, gathering, ...
LEXAS algorithm

• Uses supervised learning
  – A human-labeled dataset where all terms are disambiguated

• For each sense of word w, a set of features is extracted from the training data
• A test word t is compared with each of the training examples using a “distance” function
• The sense of w which minimizes the distance with the target word t is selected.

LEXAS: Training phase

• A window of words:
  [L3, L2, L1, W, R1, R2, R3]
• Feature extraction: A feature vector of
  – POS of the words to the left and right
  – Surrounding words
  – Keywords that co-occur frequently with w
  – Syntactic relationship
  ...

LEXAS: Distance estimation

- Distance between feature vector:
  - Sum of distance among features
- Distance between two values of feature $f$:
  \[
  \delta(v_1, v_2) = \sum_{i=1}^{n} \frac{|C_{1,i} - C_{2,i}|}{C_1 - C_2}
  \]
- $C_{1,i}$: number of training examples with value $v_1$ and sense $i$
- $C_1$: number of training examples with value 1
- $n$: total number of senses for the word

More supervised learning for WSD

- Availability of labeled data
  - Supervised algorithms:
    - Naïve Bayes
    - Decision List
    - ...
  - Semi-supervised Learning
WSD evaluation

• Intrinsic evaluation
  – Use gold-standard data to test and evaluate the actual algorithm

• Extrinsic evaluation
  – Employ the algorithm in a real-world application and see if it helps.

WSD evaluation

• SenseEval workshop: over 10 years of competition.

• Precision, Recall, Accuracy

• Baseline: Most frequent sense
  – Not always great performance by humans
    • On some polysemous words (e.g. title): ~ 70%

• Extrinsic evaluation: on MT, QA, IR
Word Similarity

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity** (or **distance**): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between **senses**
  - The word “bank” is not similar to the word “slope”
  - Bank\(^1\) is similar to fund\(^3\)
  - Bank\(^2\) is similar to slope\(^5\)

Why word similarity

- Information retrieval
- Question answering
- Machine translation
- Language modeling
- Automatic essay grading
- Plagiarism detection
Word similarity and word relatedness

- We often distinguish **word similarity** from **word relatedness**
  - **Similar words**: near-synonyms
  - **Related words**: can be related any way
    - car, bicycle: **similar**
    - car, gasoline: **related**, not similar

Two classes of similarity algorithms

- Thesaurus-based algorithms
  - Are words “nearby” in hyponym hierarchy?
  - Do words have similar glosses (definitions)?
- Distributional algorithms
  - Do words have similar distributional contexts?
Path based similarity

• Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  – have a short path between them
  – concepts have path 1 to themselves

Refinements to path-based similarity

• $\text{pathlen}(c_1, c_2) = 1 + \text{number of edges in the shortest path in the hypernym graph between sense nodes } c_1 \text{ and } c_2$
• ranges from 0 to 1 (identity)
• $\text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}$
• $\text{wordsim}(w_1, w_2) = \max \sim(c_1, c_2)$

$c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)$
Example: path-based similarity
\[ \text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)} \]

\[ \text{simpath}(\text{nickel}, \text{coin}) = \frac{1}{2} = .5 \]
\[ \text{simpath}(\text{fund}, \text{budget}) = \frac{1}{2} = .5 \]
\[ \text{simpath}(\text{nickel}, \text{currency}) = \frac{1}{4} = .25 \]
\[ \text{simpath}(\text{nickel}, \text{money}) = \frac{1}{6} = .17 \]
\[ \text{simpath}(\text{coinage}, \text{Richter scale}) = \frac{1}{6} = .17 \]

Problem with basic path-based similarity
- Assumes each link represents a uniform distance
  - But nickel to money seems to us to be closer than nickel to standard
  - Nodes high in the hierarchy are very abstract
- We instead want a metric that
  - Represents the cost of each edge independently
  - Words connected only through abstract nodes
    - are less similar
**Information content similarity metrics**

- Let’s define $P(c)$ as:
  - The probability that a randomly selected word in a corpus is an instance of concept $c$
  - Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
    - For a given concept, each observed noun is either
      - A member of that concept with probability $P(c)$
      - Not a member of that concept with probability $1-P(c)$
  - All words are members of the root node (Entity)
    - $P(\text{root})=1$
  - The lower a node in hierarchy, the lower its probability

**Information content similarity**

- Train by counting in a corpus
  - Each instance of hill counts toward frequency of natural elevation, geological formation, entity, etc
  - Let $\text{words}(c)$ be the set of all words that are children of node $c$
    - $\text{words}(\text{"geo-formation"}) = \{\text{hill, ridge, grotto, coast, cave, shore, natural elevation}\}$
    - $\text{words}(\text{"natural elevation"}) = \{\text{hill, ridge}\}$

\[
P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}
\]
Information content similarity

- WordNet hierarchy augmented with probabilities $P(c)$

The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at glosses
- Two concepts are similar if their glosses contain similar words
  - *Drawing paper*: paper that is *specially prepared* for use in drafting
  - *Decal*: the art of transferring designs from *specially prepared paper* to a wood or glass or metal surface
- For each $n$-word phrase that’s in both glosses
  - Add a score of $n^2$
  - *Paper* and *specially prepared* for $1 + 2^2 = 5$
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms
Libraries for computing thesaurus-based similarity

- NLTK
- WordNet::Similarity
  - Web-based interface:
    - [http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi](http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi)

Problems with thesaurus-based meaning

- We don’t have a thesaurus for every language
- Even if we do, they have problems with recall
  - Many words are missing
  - Most (if not all) phrases are missing
  - Some connections between senses are missing
  - Thesauri work less well for verbs, adjectives
    - Adjectives and verbs have less structured hyponymy relations
Distributional models of meaning

• Zellig Harris (1954): “oculist and eye-doctor ... occur in almost the same environments.... If A and B have almost identical environments we say that they are synonyms.

• Firth (1957): “You shall know a word by the company it keeps!”

• Offer much higher recall than hand-built thesauri
  – Although they tend to have lower precision

Intuition of distributional word similarity

A bottle of Laban is on the table
She likes laban
Laban helps your digestion
We make laban out of milk.

• From context words humans can guess laban means
  – A beverage like milk
• Intuition for algorithm:
  – Two words are similar if they have similar word contexts.
Clustering related words

Summary

• We need to process and get the meaning of the text
  – Lexical semantics: word-level meaning
    • Usually in the context
    • Challenge ambiguity (polysemy)
  – Semantic resources: WordNet
  – Word sense disambiguation
  – Word similarity