(Computational) Lexical Semantics

MLP Course, winter term 11/12

based on chapters 19/12, Jurafsky and Martin

December 21, 2011
Outline

1. Lexical Semantics (Chapter 19, J+M)
   - Word senses
   - Relations between word senses
   - WordNet
   - Lexical semantics of verbs
   - Challenges

2. Computational Lexical Semantics (Chapter 20, J+M)
   - Word Sense Disambiguation
   - Word Similarity
   - Semantic Roles Labeling
   - Towards tracking semantic change by visual analytics (Rohrdantz et al 2011)
Outline

1. **Lexical Semantics (Chapter 19, J+M)**
   - Word senses
   - Relations between word senses
   - WordNet
   - Lexical semantics of verbs
   - Challenges

2. **Computational Lexical Semantics (Chapter 20, J+M)**
   - Word Sense Disambiguation
   - Word Similarity
   - Semantic Roles Labeling
   - Towards tracking semantic change by visual analytics (Rohrdantz et al 2011)
‘the bow’

“The bow should be tall enough to prevent water from washing over the ship.”

“The bow consists of a specially shaped stick and a ribbon stretched between its ends and is used to stroke the strings and create sound.”

“Robin Hood used bow and arrow to fight the rich.”

“The level and duration of the bow depends on status, age and other factors.”
Word senses

‘the bow’

“The **bow** should be tall enough to prevent water from washing over the ship.”
- a ship’s bow

“The **bow** consists of a specially shaped stick and a ribbon stretched between its ends and is used to stroke the strings and create sound.”
- the bow of a musical instrument

“Robin Hood used **bow** and arrow to fight the rich.”
- the bow as a weapon

“The level and duration of the **bow** depends on status, age and other factors.”
- the bow as a movement
Word senses

‘the bow’

“The bow should be tall enough to prevent water from washing over the ship.”
- a ship’s bow

“The bow consists of a specially shaped stick and a ribbon stretched between its ends and is used to stroke the strings and create sound.”
- the bow of a musical instrument

“Robin Hood used bow and arrow to fight the rich.”
- the bow as a weapon

“The level and duration of the bow depends on status, age and other factors.”
- the bow as a movement

The noun bow has at least four senses
Word Senses

- one word, but some of its senses are completely unrelated
  - e.g. *bank* as in money institute and seating possibility
  - homonyms → homonymy

- one word, its senses are semantically related
  - *bow* as in weapon and part of a musical instrument
  - polysems → polysemy

→ gradual distinction between homonymy and polysemy
Word Senses

- one word, but some of its senses are completely unrelated
  - e.g. *bank* as in money institute and seating possibility
  - homonyms $\rightarrow$ homonymy

- one word, its senses are semantically related
  - *bow* as in weapon and part of a musical instrument
  - polysems $\rightarrow$ polysemy

  $\rightarrow$ gradual distinction between homonymy and polysemy

- one aspect of a concept refers to another aspect of that concept
  - e.g. usage of *White House* when referred to the administration with offices in the White house
  - metonymy
Relations between word senses

- two words with (almost) identical senses
  - couch/sofa, to vomit/to throw up
  - synonymy
  - more formally: two words are synonymous if they are substitutable without changing the truth conditions of the sentence

- two words with opposed senses
  - short/long, rise/fall
  - antonymy
Relations between word senses

- one sense is more specific than another sense
  - *hyponymy*
- one sense is less specific than another sense
  - *hypernymy*

<table>
<thead>
<tr>
<th>hypernym</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyponym</td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>
Relations between word senses

- one sense is more specific than another sense
  - *hyponymy*
- one sense is less specific than another sense
  - *hypernymy*

<table>
<thead>
<tr>
<th>hypernym</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyponym</td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>

- senses are related by a part-whole relation
  - *leg/chair, wheel/car*
  - “part” = *leg* = meronym, “whole” = *chair* = holonym
Relations between word senses

- one sense is more specific than another sense
  - *hyponymy*
- one sense is less specific than another sense
  - *hypernymy*

<table>
<thead>
<tr>
<th>hypernym</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyponym</td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>

- senses are related by a part-whole relation
  - *leg/chair, wheel/car*
  - “part” = *leg* = meronym, “whole” = *chair* = holonym

→ these concepts are the building blocks of a taxonomy, i.e. a tree-like structure of senses
The taxonomy WordNet

- the most commonly used lexical resource for English words is WordNet (Fellbaum, 1998)
- based on the relations of senses as just discussed
- three separate databases for nouns, verbs and adjectives/adverbs
- WordNet 3.0 has 117097 nouns, 11488 verbs, 22141 adjectives and 4601 adverbs

Demo
Lexical semantics of verbs

Representation of an event in a neo-Davidsonian way:

Jane broke the window.

\[ \exists e, x, y \text{ Breaking}(e) \land \text{Jane}(x) \land \text{window}(y) \land \]

<table>
<thead>
<tr>
<th>Jane broke the window.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \exists e, x, y \text{ Breaking}(e) \land \text{Jane}(x) \land \text{window}(y) \land ]</td>
</tr>
</tbody>
</table>
Lexical semantics of verbs

Representation of an event in a neo-Davidsonian way:

Jane broke the window.

$\exists e, x, y \text{ Breaking}(e) \land \text{Jane}(x) \land \text{window}(y) \land \text{Breaker}(e, x) \land \text{BrokenThing}(e, y)$
Lexical semantics of verbs

Representation of an event in a neo-Davidsonian way:

Jane broke the window.

\[ \exists e, x, y \text{ Breaking}(e) \land \text{Jane}(x) \land \text{window}(y) \land \text{Breaker}(e, x) \land \text{BrokenThing}(e, y) \]

- Breaker and BrokenThing are **deep** roles and are specific to each event
- **BUT:** in order to build computational systems we need to have a more general classification of arguments
- different approaches:
  - thematic roles (Fillmore 1968 and Gruber 1965)
  - proto roles as in PropBank
  - frame-specific roles as in FrameNet
Lexical semantics of verbs

Thematic roles (Fillmore 1968 and Gruber 1965)

<table>
<thead>
<tr>
<th>Thematic Role</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>The volitional causer of an event</td>
</tr>
<tr>
<td>EXPERIENCER</td>
<td>The experiencer of an event</td>
</tr>
<tr>
<td>FORCE</td>
<td>The non-volitional causer of the event</td>
</tr>
<tr>
<td>THEME</td>
<td>The participant most directly affected by an event</td>
</tr>
<tr>
<td>RESULT</td>
<td>The end product of an event</td>
</tr>
<tr>
<td>CONTENT</td>
<td>The proposition or content of a propositional event</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>The instrument used in an event</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>The beneficiary of an event</td>
</tr>
<tr>
<td>SOURCE</td>
<td>The origin of the objet of a transfer event</td>
</tr>
<tr>
<td>GOAL</td>
<td>The destination of an object of a transfer event</td>
</tr>
</tbody>
</table>
Lexical semantics of verbs

Representation of verb arguments with thematic roles:

Jane broke the window.

Jane = Agent, the window = Theme
Lexical semantics of verbs

Representation of verb arguments with thematic roles:

Jane broke the window.
Jane = Agent, the window = Theme
Lexical semantics of verbs

Representation of verb arguments with thematic roles:

Jane broke the window.
Jane = Agent, the window = Theme

Jane broke the window with a rock.

Lexical semantics of verbs

Representation of verb arguments with thematic roles:

Jane broke the window.
Jane = Agent, the window = Theme

Jane broke the window with a rock.
Jane = Agent, the window = Theme, the rock = Instrument
Lexical semantics of verbs

Representation of verb arguments with thematic roles:

Jane broke the window.
Jane = Agent, the window = Theme

Jane broke the window with a rock.
Jane = Agent, the window = Theme, the rock = Instrument

The window was broken by Jane.
Lexical semantics of verbs

Representation of verb arguments with thematic roles:

Jane broke the window.
Jane = Agent, the window = Theme

Jane broke the window with a rock.
Jane = Agent, the window = Theme, the rock = Instrument

The window was broken by Jane.
the window = Theme, Jane = Agent

Possible arguments of to break: AGENT, THEME, INSTRUMENT
Lexical semantics of verbs

But verbs can vary according to which thematic roles they assign in what position:

(1) a. Jane broke the window.
    b. The window broke.

(2) a. Jane cut the cake.
    b. *The cake cut. Conative alternation

Levin (1993) is a reference book that lists all verb alternations for English and detects semantic classes of verbs based on their syntactic behavior → basis for the English Verbnet (Demo)
Lexical semantics of verbs

But verbs can vary according to which thematic roles they assign in what position:

(4) a. Jane broke the window.
    b. The window broke.

(5) a. Jane cut the cake.
    b. *The cake cut. Conative alternation

(6) a. Jane gave the book to James.
    b. Jane gave James the book. Dative alternation

Levin (1993) is a reference book that lists all verb alternations for English and detects semantic classes of verbs based on their syntactic behavior → basis for the English Verbnet (Demo)
Lexical semantics of verbs

The Proposition Bank (PropBank)

- the PennTreebank annotated with semantic roles
- semantic roles are defined with respect to an individual verb sense
- roles in PropBank are numbered rather than labeled, e.g. Arg0, Arg1 etc.

Ex1: 
[Arg0] The group [Arg1] agreed [Arg2] it wouldn't make an offer unless it had Georgia Gulf's consent.

Ex2: 
Lexical semantics of verbs

The Proposition Bank (PropBank)

- the PennTreebank annotated with semantic roles
- semantic roles are defined with respect to an individual verb sense
- roles in PropBank are numbered rather than labeled, e.g. Arg0, Arg1 etc.

agree.01

Agr0: Agreeer
Agr1: Proposition
Agr2: Other entity agreeing

Ex1: \([Agr_0 \text{ The group }] \text{ agreed } [Agr_1 \text{ it wouldn’t make an offer unless it had Georgia Gulf’s consent }]\).

Ex2: \([ArgM-TMP \text{ Usually }] [Arg_0 \text{ John }] \text{ agrees } [Arg_2 \text{ with Mary}] [Arg_1 \text{ on everything}]\).
Lexical semantics of verbs

Problems with PropBank

✓ \([Agr_0 \ \text{The group}] \ \text{agreed} \ [Agr_1 \ \text{it wouldn’t make an offer unless it had Georgia Gulf’s consent}].\)

✓ \([ArgM-TMP \ \text{Usually}] \ [Arg_0 \ \text{John}] \ \text{agrees} \ [Arg_2 \ \text{with Mary}] \ [Arg_1 \ \text{on everything}].\)
Lexical semantics of verbs

Problems with PropBank

✓ \[Agr_0 \ \text{The group } \] agreed \[Agr_1 \ \text{it wouldn’t make an offer unless it had Georgia Gulf’s consent } \].

✓ \[ArgM-TMP \ \text{Usually } \] \[Arg_0 \ \text{John } \] agrees \[Arg_2 \ \text{with Mary } \] \[Arg_1 \ \text{on everything} \].

? \[ArgM-TMP \ \text{Usually } \] \[Arg_0 \ \text{John } \] consents \[Arg_2 \ \text{with Mary } \] \[Arg_1 \ \text{on everything} \].

? There is an agreement of \[Arg_0 \ \text{John} \] with \[Arg_2 \ \text{with Mary} \].

We would like to represent these roles in a uniform way, across different verbs and also across nouns and verbs → FrameNet
Lexical semantics of verbs

FrameNet

- semantic role labeling project that attempts to address the problems of thematic roles and PropBank (Baker et al. 1998, Lowe et al. 1997 and Ruppenhofer et al. 2006)
- verbs are grouped in frames where specific roles hold
- e.g. frame make_agreement_on_action

Demo
Two main challenges in the computational treatment of lexical semantics:

- **selectional restrictions**
  - semantic constraint that the verb imposes on the concepts that are allowed to fill its argument structure

- **metaphors**
  - relation between two completely different domains of meaning - generating an independent meaning
Challenges

Selectional restrictions:

(7) a. I want to eat Malaysian food.
    b. I want to eat somewhere.

How do we know that *somewhere* is not the direct object of the sentence?

- intransitive and transitive version of *to eat*
- the direct object of *to eat* must be an edible entity
- *somewhere* is a location and not edible
Challenges

(8) a. Does American Airlines still serve a hot meal?  
   b. Does American Airlines still serve Denver?

Senses of serve:
- cooking/providing food
- providing a commercial service
- and probably other senses, too

→ the set of concepts needed to represent selectional restrictions is almost open-ended
→ no resource available that encodes a full range of these concepts (does a finite set of these concepts exist at all?)
Challenges

Can we get around the problem of selectional restrictions?

1. Usage of WordNet?
   - for the case of *to eat* we could refer to the synset *food, nutrient* for its direct object
   - but then we also need to account for cases like *I ate rabbit the other day* item include the synset *animal* as well?

2. Decomposing the meaning of words into their primitive semantic elements?
   - What would these elements be for *cow, bull, calf*?
Challenges

A further problem for computers: metaphors

(9) It doesn’t scare Microsoft that Apple’s new IPad is out.

- here, the company is viewed as a person that can experience fear
- problem for the computer: when is an expression metaphorically used and when is it ill-formed?
  - Apple is scared of mice.
Quick recap

- **Word senses:**
  - homonymy
  - polysemy
  - metonymy

- **Relations between word senses:**
  - synonymy
  - antonymy
  - hyponymy/hypernymy
  - meronymy

- **verb lexical semantics**
  - thematic roles
  - proto-roles
  - frame roles
Outline

1 Lexical Semantics (Chapter 19, J+M)
   - Word senses
   - Relations between word senses
   - WordNet
   - Lexical semantics of verbs
   - Challenges

2 Computational Lexical Semantics (Chapter 20, J+M)
   - Word Sense Disambiguation
   - Word Similarity
   - Semantic Roles Labeling
   - Towards tracking semantic change by visual analytics (Rohrdantz et al 2011)
Computational Lexical Semantics

- introduction to computational tasks associated with
  - word sense disambiguation
  - extract of relations among words
  - automatic assignment of thematic structure to predicate-bearing words
Word Sense Disambiguation (WSD)

Two main approaches:

1. lexical sample approach
   - a small pre-selected set of target words to be disambiguated
   - set of senses for each word from a lexicon
   - corpus instances of the target words are hand-labelled with the correct senses
     - e.g. line-hard-serve corpus (Leacock et al. 1993), interest corpus (Bruce and Wiebe 1994) and SENSEVAL corpora
   - classifier systems are trained on these instances
   - unlabeled instances are then tagged with the classifier
Word Sense Disambiguation (WSD)

Two main approaches:

1. lexical sample approach
   - a small pre-selected set of target words to be disambiguated
   - set of senses for each word from a lexicon
   - corpus instances of the target words are hand-labelled with the correct senses
     - e.g. line-hard-serve corpus (Leacock et al. 1993), interest corpus (Bruce and Wiebe 1994) and SENSEVAL corpora
   - classifier systems are trained on these instances
   - unlabeled instances are then tagged with the classifier

2. all-words approach
   - a system is given a text and a lexicon with senses of the words of the text
     - e.g. SemCor (Miller et al. 1993, Landes et al. 1998) and SENSEVAL-3 (Palmer et al. 2001)
   - then every content word of the text is disambiguated
Word Sense Disambiguation (WSD)

Approach 1: Supervised learning (lexical sample task):

1. extraction of features that are predictive of word senses
   - collocational features: position-specific relation to the target word
   - bag-of-words features: unordered set of words, exact position is ignored
Word Sense Disambiguation (WSD)

Approach 1: Supervised learning (lexical sample task):

1. extraction of features that are predictive of word senses
   ▶ collocational features: position-specific relation to the target word
   ▶ bag-of-words features: unordered set of words, exact position is ignored

An electric guitar and **bass** player stand off to one side, just as a sort of nod to gringo expectation perhaps.

Collocational feature vector of target word $w_i$:

$[w_{i-2}, \text{POS}_{i-2}, w_{i-1}, \text{POS}_{i-1}, w_{i+1}, \text{POS}_{i+1}, w_{i+2}, \text{POS}_{i+2}]$
Word Sense Disambiguation (WSD)

**Approach 1: Supervised learning (lexical sample task):**

1. extraction of features that are predictive of word senses
   - collocational features: position-specific relation to the target word
   - bag-of-words features: unordered set of words, exact position is ignored

   An electric guitar and **bass** player stand off to one side, just as a sort of nod to gringo expectation perhaps.

Collocational feature vector of target word $w_i$:

$$[w_{i-2}, \text{POS}_{i-2}, w_{i-1}, \text{POS}_{i-1}, w_{i+1}, \text{POS}_{i+1}, w_{i+2}, \text{POS}_{i+2}]$$

[guitar, NN, and, CC, player, NN, stand, VB]
Vocabulary vector of the 10 most frequent content words in bass sentences:

\[ \text{fishing, sound, player, fly, rod, double, runs, playing, guitar, band} \]

An electric guitar and bass player stand off to one side, just as a sort of nod to gringo expectation perhaps.

Bag-of-words feature vector with binary features:
Word Sense Disambiguation (WSD)

Vocabulary vector of the 10 most frequent content words in bass sentences:
[fishing, sound, player, fly, rod, double, runs, playing, guitar, band]

An electric guitar and bass player stand off to one side, just as a sort of nod to gringo expectation perhaps.

Bag-of-words feature vector with binary features:
[0,0,1,0,0,0,0,0,1,0]
Word Sense Disambiguation (WSD)

Vocabulary vector of the 10 most frequent content words in bass sentences:

\[\text{fishing, sound, player, fly, rod, double, runs, playing, guitar, band}\]

An electric guitar and bass player stand off to one side, just as a sort of nod to gringo expectation perhaps.

Bag-of-words feature vector with binary features:

\[0,0,1,0,0,0,0,0,1,0,0,0,1,0]\]

These vectors are then input to machine learning algorithms.
Word Sense Disambiguation (WSD)

2. Training of a classifier: the Naive Bayes classifier:

\[ \hat{s} = \arg\max P(s) \prod_{j=1}^{n} P(f_j|s) \]

- training a naive Bayes classifier means estimating each of these probabilities:
  - \( P(s_i) = \frac{\text{count}(s_i, w_j)}{\text{count}(w_j)} \) = prior probability of each sense
    - counting the number of times sense \( s_i \) occurs, divided by the total number of target word \( w_j \)
    - If the target word \( \text{bass} \) appears 150 times in the corpus and it has sense \( \text{bass}^1 \) in 60 cases, what is the prior probability of the sense?
2. Training of a classifier: the Naive Bayes classifier:

- \( \hat{s} = \operatorname{argmax} P(s) \prod_{j=1}^{n} P(f_j|s) \)

training a naive Bayes classifier means estimating each of these probabilities

- \( P(s_i) = \frac{\text{count}(s_i,w_j)}{\text{count}(w_j)} = \text{prior probability of each sense} \)
  
  - counting the number of times sense \( s_i \) occurs, divided by the total number of target word \( w_j \)
  
  - If the target word \( \text{bass} \) appears 150 times in the corpus and it has sense \( \text{bass}^1 \) in 60 cases, what is the prior probability of the sense?

- \( P(f_j|s) = \frac{\text{count}(f_j,s)}{\text{count}(s)} = \text{individual feature probabilities} \)
  
  - If a feature such as \([w_{i-2} = \text{guitar}]\) occurs three times for sense \( \text{bass}^1 \), and sense \( \text{bass}^1 \) occurs 60 times in the corpus, what is its individual feature probability?
Word Sense Disambiguation (WSD)

Naive Bayes classifier:

\[ \hat{s} = \text{argmax} \ P(s) \prod_{j=1}^{n} P(f_j | s) \]

- putting in the values computed before for
  - \( P(s_i) = \frac{\text{count}(s_i, w_j)}{\text{count}(w_j)} \) = prior probability of each sense
  - \( P(f_j | s) = \frac{\text{count}(f_j, s)}{\text{count}(s)} \) = individual feature probabilities

- What is the probability of guitar occurring with sense \( \text{bass}^1 \)?
3. Evaluation of WSD systems:

- A WSD system can be evaluated with respect to sense accuracy
  - The percentage of words that are tagged identically to the hand-labeled sense tags in the test set
- Usually compared to two measures:
  - Baseline
    - E.g. simply take the most frequent sense for each word
  - Ceiling
    - E.g. human inter-annotator agreement
Word Sense Disambiguation (WSD)

Approach 2: Dictionary and Thesaurus Methods (all-words approach)

- The Lesk algorithm: family of algorithms for dictionary-based sense disambiguation
- Simplified Lesk algorithm (Kilgarriff and Rosenzweig 2000):
  - which sense gloss shares the most words with the target word’s neighbourhood?

The bank can guarantee deposits that will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.
Word Sense Disambiguation (WSD)

Approach 2: Dictionary and Thesaurus Methods (all-words approach)

- The Lesk algorithm: family of algorithms for dictionary-based sense disambiguation
- Simplified Lesk algorithm (Kilgarriff and Rosenzweig 2000):
  ▶ which sense gloss shares the most words with the target word’s neighbourhood?

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

<table>
<thead>
<tr>
<th>bank¹</th>
<th>Gloss: a financial institution that accepts deposits and channels the money into lending activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank²</td>
<td>Gloss: sloping land (especially the slope beside a body of water)</td>
</tr>
</tbody>
</table>

Which sense is taken?
Word Sense Disambiguation (WSD)

2. Dictionary and Thesaurus Methods

- **Original Lesk algorithm (Lesk 1986):**
  - the gloss of the target word is compared to the glosses of the surrounding words
  - the sense with the most overlapping words is chosen
Word Sense Disambiguation (WSD)

2. Dictionary and Thesaurus Methods

- Original Lesk algorithm (Lesk 1986):
  - the gloss of the target word is compared to the glosses of the surrounding words
  - the sense with the most overlapping words is chosen

\[
\text{pine cone}
\]

<table>
<thead>
<tr>
<th>pine$^1$</th>
<th>Gloss: kinds of evergreen trees with needle-shaped leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>pine$^2$</td>
<td>Gloss: waste away through sorrow or illness</td>
</tr>
<tr>
<td>cone$^1$</td>
<td>Gloss: solid body which narrows to a point</td>
</tr>
<tr>
<td>cone$^2$</td>
<td>Gloss: something of this shape whether solid or hollow</td>
</tr>
<tr>
<td>cone$^3$</td>
<td>Gloss: fruit of certain evergreen trees</td>
</tr>
</tbody>
</table>

Which sense is taken?
Word Sense Disambiguation (WSD)

The caveat of large hand-built resources

- both the supervised approach and the dictionary-based approach require large amounts of labeled data
- what can be done if these resources are not available?

  → e.g. Yarovsky algorithm (1995)

  ▶ small seedset of labeled instances of each sense and a much larger unlabeled corpus
  ▶ first training of an initial classifier on the seedset
  ▶ then parsing of the unlabeled data with this classifier
  ▶ selection of the most confident labeled instance and addition to the training set
  ▶ with each iteration, the training set grows and the unlabeled corpus shrinks
Word Similarity

- to compute word similarity is useful for many natural language applications
Word Similarity

- to compute word similarity is useful for many natural language applications
  - machine translation
  - information retrieval
  - question answering
  - text summarization
Word Similarity

- to compute word similarity is useful for many natural language applications
  - machine translation
  - information retrieval
  - question answering
  - text summarization

- two classes of algorithms: thesaurus-based algorithms and distributional algorithms
Word Similarity

Approach 1: Thesaurus-based algorithms:

- usage of the structure of a thesaurus to define word similarity
- word similarity $\neq$ word relatedness
  - word relatedness characterizes a larger set of potential relationship between words
  - e.g. antonyms are related but not similar
- Path-length-based similarity: measuring the edges between two concepts
  \[
  \text{sim}_{path}(c_1, c_2) = \text{pathlen}(c_1, c_2)
  \]
- Log transform of path-length-based similarity
  \[
  \text{sim}_{path}(c_1, c_2) = -\log \text{pathlen}(c_1, c_2)
  \]
Word Similarity

- problem with path-length algorithms:
  - assumption that each link in the thesaurus represents a uniform distance

→ information-content word-similarity algorithms (following Resnik 1995)
  - the lower a concept in a hierarchy, the lower its probability
  - $P(c)$ is the probability that a randomly selected word in a corpus is an instance of concept $c$
  - $P(\text{root}) = 1$ (any word is subsumed by the root concept)

$$P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}$$
Word Similarity

- two additional definitions are needed:
  - information content (IC) of a concept: \( \text{IC}(c) = -\log P(c) \)
  - lowest common subsumer (LCS) of two concepts: \( \text{LCS}(c_1, c_2) \)
    - the lowest node in the hierarchy that subsumes (is a hypernym of) both \( c_1 \) and \( c_2 \).

- Resnik similarity measure:
  \[
  \text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2))
  \]
  - information content of the lowest common subsumer of the two nodes
Word Similarity

Approach 2: Distributional Algorithms

- Intuition: the meaning of a word is related to the distribution of words around it
  - “You shall know a word by the company it keeps.” (Firth 1957)

A bottle of warzyku is on the table
Everybody likes warzyku
Warzyku makes you drunk
We make warzyku out of corn.

- “word meaning” as a feature vector $\vec{w}$ with a binary features $f_n$
- the words in the context are $v_n$
- if $v_1$ is present, then feature $f_1$ is 1
- here: $w =$ warzyku, $v_1 =$ bottle, $v_2 =$ like, $v_3 =$ drunk, $v_4 =$ corn, $v_5 =$ matrix
Word Similarity

Approach 2: Distributional Algorithms

- Intuition: the meaning of a word is related to the distribution of words around it
  - “You shall know a word by the company it keeps.” (Firth 1957)
    
    A bottle of warzyku is on the table
    Everybody likes warzyku
    Warzyku makes you drunk
    We make warzyku out of corn.

- “word meaning” as a feature vector $\vec{w}$ with a binary features $f_n$
- the words in the context are $v_n$
- if $v_1$ is present, then feature $f_1$ is 1
- here: $w = \text{warzyku}$, $v_1 = \text{bottle}$, $v_2 = \text{like}$, $v_3 = \text{drunk}$, $v_4 = \text{corn}$, $v_5 = \text{matrix}$
- word vector: $\vec{w} = (1, 1, 1, 1, 0)$
Word Similarity

applying distributional algorithms for word similarity measure means deciding about the following facts:

1. how are the co-occurrence terms defined (i.e. what counts as a neighbor)?
2. how are these terms weighted?
3. what vector distance metrics are used?
1. What counts as a neighbor?

- neighborhoods range from small windows (2 words before and after the target word) to very large context windows (500 words)
- Schütze (2001)’s experiments show that a context window of 50 words is enough for word sense disambiguation
- usually, stop words are removed
- grammatical dependencies and relations can also be used for context vectors
2. How are the terms weighted?

<table>
<thead>
<tr>
<th>target word w</th>
<th>warzyku</th>
<th>f</th>
<th>2</th>
<th>4</th>
<th>1</th>
</tr>
</thead>
</table>

- vector of N x R features, where R is the number of possible relations
- here: feature f are frequencies (a better indicator than binary values)
- \( f = (r, w') \)
- \( P(f \mid w) = \frac{\text{count}(f,w)}{\text{count}(w)} \) (the probability of feature f given a target word w)
Word Similarity

- $P(f, w) = \frac{\text{count}(f, w)}{\sum_{w'} \text{count}(w')} \text{ (the joint probability of feature } f \text{ given a target word } w \text{ and a context word } w')$

- maximum likelihood estimate replaces the frequencies in the table

<table>
<thead>
<tr>
<th>relation, $w'$</th>
<th>subj-of, make</th>
<th>obj-of, like</th>
<th>obj-of, make</th>
</tr>
</thead>
<tbody>
<tr>
<td>target word $w$</td>
<td>warzyku</td>
<td>$f$</td>
<td>MLE</td>
</tr>
</tbody>
</table>
Word Similarity

3. What vector distance metrics are used?

- measure for taking two such vectors and giving a measure of vector similarity

- Levensthein distance: \( \text{dist}_L (\vec{v}, \vec{w}) = \sum_{n=1}^{N} |v_i - w_i| \)
Word Similarity

3. What vector distance metrics are used?

- measure for taking two such vectors and giving a measure of vector similarity

- Levensthein distance: \( \text{dist}_L (\vec{v}, \vec{w}) = \sum_{n=1}^{N} |v_i - w_i| \)

- Euclidean distance: \( \text{dist}_E (\vec{v}, \vec{w}) = \sqrt{\sum_{n=1}^{N} (v_i - w_i)^2} \)
Word Similarity

3. What vector distance metrics are used?

- measure for taking two such vectors and giving a measure of vector similarity
  - Levensthein distance: $\text{dist}_L (\vec{v}, \vec{w}) = \sum_{n=1}^{N} |v_i - w_i|$
  - Euclidean distance: $\text{dist}_E (\vec{v}, \vec{w}) = \sqrt{\sum_{n=1}^{N} (v_i - w_i)^2}$
- both measures are rarely used for word similarity, because extreme values change the measure significantly
Word Similarity

- dot product as similarity measure: \( \text{dist}_L (\vec{v}, \vec{w}) = \sum_{n=1}^{N} v_i - w_i \)
Word Similarity

- dot product as similarity measure: \( \text{dist}_L (\vec{v}, \vec{w}) = \sum_{n=1}^{N} v_i - w_i \)
- BUT: we normalize for the vector length
Word Similarity

- dot product as similarity measure: $\text{dist}_L (\vec{v}, \vec{w}) = \sum_{n=1}^{N} v_i - w_i$
- BUT: we normalize for the vector length
- vector length: $|\vec{v}| = \sqrt{\sum_{n=1}^{N} v_i^2}$
- normalized dot product:

$$\text{sim}_{\text{norm-dot-product}} (\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| |\vec{w}|} = \frac{\sum_{n=1}^{N} v_i - w_i}{\sqrt{\sum_{n=1}^{N} v_i^2} \sqrt{\sum_{n=1}^{N} w_i^2}}$$
Semantic Role Labeling

- current approaches rely on adequate amounts of training and testing data
- General (simplified) approach:
  1. parsing the sentence
  2. finding all predicates (here: verbs)
  3. traversing the tree to determine the roles of the constituents with respect to that predicate
     → feature vector
Semantic Role Labeling

- these observations (feature vectors) are then divided in test and training set
- training of classifier which then yields good results on unlabeled data
- training is mostly done in different stages
  - elimination of some possible role constituents based on simple heuristics (pruning) → speeds up training
  - binary identification of each node as being either ARG or NONE
  - classification of the ARG labeled constituents
Towards tracking semantic change by visual analytics

**Motivation**

1. increasing amount of diachronic data electronically available
2. demand of historical linguists to process these corpora and see developments and patterns over time at-a-glance
Towards tracking semantic change by visual analytics

Motivation

1. increasing amount of diachronic data electronically available
2. demand of historical linguists to process these corpora and see developments and patterns over time at-a-glance

Challenge

Tracking of overall developments of language and also allowing to delve into the details of the data.
Towards tracking semantic change by visual analytics

Motivation

1. increasing amount of diachronic data electronically available
2. demand of historical linguists to process these corpora and see developments and patterns over time at-a-glance

Challenge

Tracking of overall developments of language and also allowing to delve into the details of the data.

Research question

Can we create tools that aid during the analysis of language change, can they test existing hypotheses of change and can they even generate new ones?
Towards tracking semantic change by visual analytics

The object under investigation is semantic change (here: in English)

But what is semantic change?

- if a word changes its meaning over time, it has undergone semantic change.
- some types of semantic change:
  - narrowing (the meaning of a word becomes restricted), e.g. skyline
  - widening (the meaning of a word widens), e.g. horn
- semantic change in the last 20 years: words related to the computer and the internet
Towards tracking semantic change by visual analytics

Methodology

- search New York Times corpus
  - 1.8 million newspaper articles from 1987 to 2007
  - each article has a specific time stamp
- extract context of 25 words before and after the lexical item under investigation
- use statistics to model word senses on the basis of word contexts
  - Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
    - not applied on documents but on contexts
    - we predefine the number of senses, each context is assigned to one sense
- add a visualization layer that graphically interprets the information from the statistical analysis and makes it accessible to historical linguists
Towards tracking semantic change by visual analytics

First visualization approach

- aggregated view on the data

<table>
<thead>
<tr>
<th>to browse</th>
<th>to surf</th>
</tr>
</thead>
<tbody>
<tr>
<td>time, library,</td>
<td>sport, wind,</td>
</tr>
<tr>
<td>student, music,</td>
<td>water, ski, offer</td>
</tr>
<tr>
<td>people</td>
<td>wave, surfer,</td>
</tr>
<tr>
<td>shop, street,</td>
<td>board, year,</td>
</tr>
<tr>
<td>book, store, art</td>
<td>sport</td>
</tr>
<tr>
<td>book, read,</td>
<td>channel,</td>
</tr>
<tr>
<td>bookstore, find,</td>
<td>television,</td>
</tr>
<tr>
<td>year</td>
<td>show, watch, tv</td>
</tr>
<tr>
<td>deer, plant,</td>
<td>web, internet,</td>
</tr>
<tr>
<td>tree, garden,</td>
<td>site, computer,</td>
</tr>
<tr>
<td>animal</td>
<td>company</td>
</tr>
<tr>
<td>software, microsoft, internet, netscape, windows</td>
<td>film, boy, movie, show, ride</td>
</tr>
<tr>
<td>web, internet,</td>
<td>year, day, time,</td>
</tr>
<tr>
<td>site, mail,</td>
<td>school, friend</td>
</tr>
<tr>
<td>computer</td>
<td>beach, wave,</td>
</tr>
<tr>
<td>store, shop,</td>
<td>surfer, long,</td>
</tr>
<tr>
<td>buy, day,</td>
<td>coast</td>
</tr>
<tr>
<td>customer</td>
<td></td>
</tr>
</tbody>
</table>
Towards tracking semantic change by visual analytics

Second visualization approach

- **individual plotting of the contexts of** *to browse*

```
d
deer, plant, tree, garden, animal
```

```
e
software, microsoft, internet, netscape, windows
```

```
d
2007
```

```
e
1987
```
Towards tracking semantic change by visual analytics

Second visualization approach

- individual plotting of the contexts of *to browse*

![Visual Analytics Diagram]

Sat Dec 13 1997 --- system to personal computer makers. The consens agreement was signed just as use of the Internet was beginning to soar, fueled by easy-to-use **browsing** programs for using the World Wide Web. The first major commercial browser was the Netscape Communications Corporation’s Navigator. Netscape remains the leader with more ---

software, microsoft, internet, netscape, windows

1987

2007
Towards tracking semantic change by visual analytics

Second visualization approach

- individual plotting of the contexts of *to browse*

---

Sun Oct 06 1991 --- defensive landscaping is an almost impossible achievement. But there are some plants that deer prefer to eat, and these species could be avoided where deer *browsing* has been a recurrent problem. At the top of the animal’s feeding list is the yew *Taxus*, which they devour with abandon and nibble right ---

---

---
Towards tracking semantic change by visual analytics

Second visualization approach

- individual plotting of the contexts of *to browse*

Thu May 08 2003 --- a computer programmer has used correct language syntax and rules in writing the code. Runtime errors can be caused by a variety of factors, like *browsing* Web pages that use coding that your browser program cannot understand. When a program encounters a runtime error, it may produce an alert box or ---
Towards tracking semantic change by visual analytics

Evaluation

- generally difficult (if not impossible) to fully evaluate statistical approaches to meaning change
- one attempt: compare the findings from the visualization with information from dictionaries from different time periods
  - Longman Dictionary from 1987 (LONG)
  - WordNet from 1998 (WN)
  - Collins dictionary from 2007 (COLL)
Towards tracking semantic change by visual analytics

Evaluation

<table>
<thead>
<tr>
<th></th>
<th>to browse # of word senses</th>
<th>to surf # of word senses</th>
<th>messenger # of word senses</th>
<th>bookmark # of word senses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIC</td>
<td>VIS</td>
<td>DIC</td>
<td>VIS</td>
</tr>
<tr>
<td>1987 (LONG)</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1998 (WN)</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2007 (COLL)</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table: Evaluation of visualized senses against dictionary senses

- in general, the number of our senses corresponds to the information coming from the dictionary
- in the case of “messenger” the visualization proves to be even more detailed
# Towards tracking semantic change by visual analytics

## Evaluation

<table>
<thead>
<tr>
<th>Year</th>
<th>Definition</th>
<th>Long</th>
<th>Vis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>LONG: a person who brings a message</td>
<td>bike messenger, messenger (genetics)</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>WN: a person who carries a message</td>
<td>bike messenger, messenger (genetics), religious messenger</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>COLL: a person who brings a message</td>
<td>bike messenger, messenger (genetics), religious messenger, instant messenger</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Sense development of *messenger* from 1987 to 2007