Mining Unstructured Data
4. Lexical semantics
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Semantics

Semantics deals with the meaning:

- **Lexical semantics**: deals with the meaning of individual words
- **Compositional semantics**: deals with the construction of meaning usually in high concordance with syntax

This session focuses on lexical semantics
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Motivation of lexical semantics

Some examples of usefulness:

- Discovery of semantic patterns
  
  Ex: USA **bombed** Hiroshima
  They began to **bombard** the defenses
  → A *sense_12533* B

- Determine discourse relations
  
  Ex: [Anna will show up *later.*] [She has **missed the train.**] → explanation
  Ex: [Mathew is good cooking.] [Albert fails making every dish] → contrast

- Twitter sentiment analysis
  
  Ex: @vooda1: CNN Declines to Air White House Press Conference Live YES! THANK YOU @CNN FOR NOT LEGITIMI... positive
  Ex: @Slate: Donald Trump’s administration: “Government by the worst men.” negative
Outline

1 Semantics
   - Motivation of lexical semantics
   - Resources

2 WordNet
   - Definition
   - Similarities

3 SentiWordNet

4 Sentiment analysis
   - Definition
   - Examples of methods
Resources of lexical semantics

- **Knowledge-based resources: represented as graphs**
  - Ex: **WordNet** (English lexical ontology)
  - **SentiWordNet** (sentiment polarity into WordNet)
  - BabelNet (Wikipedia + WordNet)
  - VerbNet (syntactic/semantic verbal behaviour)
  - FrameNet (conceptual behaviour – fine-grained event representation –)
  - ConceptNet (common sense knowledge)

- **Corpus-based resources: contextual usage of words**
  - Ex: Latent Semantic Analysis (LSA)
  - Word embeddings (word2vect, glove, fasttext, ...)
  - Contextual word embeddings as compositional semantics (BERT, RoBERTA, GPT3, ...)

*We will study them from Session 8 on*
## Resources of lexical semantics

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td><a href="https://wordnet.princeton.edu/">https://wordnet.princeton.edu/</a></td>
</tr>
<tr>
<td>SentiWordNet</td>
<td><a href="https://github.com/aesuli/SentiWordNet">https://github.com/aesuli/SentiWordNet</a></td>
</tr>
<tr>
<td>BabelNet</td>
<td><a href="https://babelnet.org/">https://babelnet.org/</a></td>
</tr>
<tr>
<td>VerbNet</td>
<td><a href="https://verbs.colorado.edu/verbnet/">https://verbs.colorado.edu/verbnet/</a></td>
</tr>
<tr>
<td>FrameNet</td>
<td><a href="https://framenet.icsi.berkeley.edu/fndrupal/">https://framenet.icsi.berkeley.edu/fndrupal/</a></td>
</tr>
<tr>
<td>LSA</td>
<td>accessible from</td>
</tr>
<tr>
<td>Word embeddings</td>
<td><a href="https://radimrehurek.com/gensim/">https://radimrehurek.com/gensim/</a></td>
</tr>
</tbody>
</table>
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
WordNet

- Free large lexical database of English
- Contains only nouns, verbs, adjectives and adverbs
- Words are grouped into synonyms sets (synsets)
- Each synset has an associated gloss and some examples
- Synsets are interlinked by means of lexical relations

http://wordnetweb.princeton.edu/perl/webwn

Example:

Word to search for: age

Display Options: (Select option to change)  

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) age (how long something has existed) "it was replaced because of its age"
- S: (n) historic period, age (an era of history having some distinctive feature) "we live in a litigious age"
Lexical relations

Example of Lexical Relation Net

{conveyance; transport}

hyperonym

{vehicle}

hyperonym

{motor vehicle; automotive vehicle}

hyperonym

{car; auto; automobile; machine; motorcar}

meronym

{bumper}

meronym

{hinge; flexible joint}

meronym

{car door}

meronym

{doorlock}

meronym

{car window}

meronym

{armrest}

meronym

{car mirror}

{cruiser; squad car; patrol car; police car; prowl car}

{cab; taxi; hack; taxicab; }
Lexical relations

- **Synonym**: same meaning. Ex: age - historic_period
- **Antonym**: opposite meaning. Ex: dark - light
- **Homophone**: same sound. Ex: son - sun
- **Homograph**: same written form. Ex: lead (noun - verb)
- **Polysemy**: different related meaning. Ex: newspaper (paper - firm)
- **Homonymy**: different unrelated meaning. Ex: position (place - status)
- **Hypernym**: parent. Ex: cat - feline
- **Hyponym**: child. Ex: feline - cat
- **Holonymy**: group, whole. Ex: class - student
- **Meronymy**: member, part. Ex: student - class
- **Metonymy**: substitution of entity. Ex: We ordered many delicious dishes at the restaurant.
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Similarities in WordNet

- Shortest Path Length: \( \text{Sim}(s_1, s_2) = \frac{1}{1 + \text{SPL}(s_1, s_2)} \)
  where \( \text{SPL}(s_1, s_2) \) = Shortest Path Length from \( s_1 \) to \( s_2 \) as edge-countings

- Leacock & Chodorow: \( \text{Sim}(s_1, s_2) = -\log \frac{1 + \text{SPL}(s_1, s_2)}{2 \cdot \text{MaxDepth}} \)
  where \( \text{depth}(s) = \text{SPL}(\text{TopSynset}, s) \)
  \[ \text{MaxDepth} = \max_{s \in \text{WN}} \text{depth}(s) \]

- Wu & Palmer:
  \[ \text{Sim}(s_1, s_2) = \frac{2 \cdot \text{depth}(\text{LCS}(s_1, s_2))}{\text{depth}_{\text{LCS}(s_1, s_2)}(s_1) + \text{depth}_{\text{LCS}(s_1, s_2)}(s_2)} \]
  where \( \text{LCS}(s_1, s_2) \) = Lowest Common Subsumer of \( s_1 \) and \( s_2 \)
  \[ \text{depth}_{s'}(s) = \text{SPL}(\text{TopSynset}, s) \text{ throw } s' \]

- Lin: \( \text{Sim}(s_1, s_2) = \frac{2 \cdot \text{IC}(\text{LCS}(s_1, s_2))}{\text{IC}(s_1) + \text{IC}(s_2)} \)
  where \( \text{IC}(s) = -\log_2 P(s) \) = information content of \( s \) (from frequencies in a corpus)
Example / exercise

\[ \text{spl}(\text{beer, milk}) = 4 \]
\[ \text{Sim}_{\text{spl}}(\text{beer, milk}) = 0.2 \]
\[ \text{Sim}_{\text{wp}}(\text{beer, milk}) = 0.71 \]

\[ \text{Sim}_{\text{spl}}(\text{drug, milk})? \]
\[ \text{Sim}_{\text{wp}}(\text{drug, milk})? \]
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Definition

Extension of wordnet that adds for each synset 3 measures:

- positive_score
- negative_score
- objective_score = 1 - positive_score - negative_score

<table>
<thead>
<tr>
<th>Antonym Synsets</th>
<th>Gloss</th>
<th>Wordnet</th>
<th>SentiWordnet</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad.a.01</td>
<td>having undesirable or negative qualities</td>
<td>0.375</td>
<td>0.0</td>
</tr>
<tr>
<td>good.a.01</td>
<td>having desirable or positive qualities</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>bad.n.01</td>
<td>that which is below standard or expectations as of ethics or decency</td>
<td>0.125</td>
<td>0.0</td>
</tr>
<tr>
<td>good.n.03</td>
<td>that which is pleasing, valuable, useful</td>
<td>0.375</td>
<td>0.625</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gloss</th>
<th>obj</th>
<th>pos</th>
<th>neg</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad.a.01</td>
<td>0.375</td>
<td>0.0</td>
<td>0.625</td>
</tr>
<tr>
<td>good.a.01</td>
<td>0.25</td>
<td>0.75</td>
<td>0.0</td>
</tr>
<tr>
<td>bad.n.01</td>
<td>0.125</td>
<td>0.0</td>
<td>0.875</td>
</tr>
<tr>
<td>good.n.03</td>
<td>0.375</td>
<td>0.625</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Sentiment analysis

Different subtasks:

- **Opinion detection**: given a piece of text (document or sentence), is it an objective text or a subjective one?

- **Polarity classification**: given a subjective piece of text, is it a positive opinion or a negative one?

- Opinion extraction: given a subjective piece of text, recognise the focuses of the opinion (templates $<$entity, aspect, polarity$>$).
Outline

1. Semantics
   - Motivation of lexical semantics
   - Resources

2. WordNet
   - Definition
   - Similarities

3. SentiWordNet

4. Sentiment analysis
   - Definition
   - Examples of methods
Unsupervised sentiment analysis

Possible simple solution with lexical information:

\[ h(D) = \sum_{w \in \hat{D}} \text{word\_score}(w) \]

\[ \text{word\_score}(w) = 1/|S(w)| \times \sum_{s \in S(w)} \text{score}(s) \]

\( \hat{D} \) is usually the set of adjectives, or nouns and adjectives, or nouns, verbs, adjectives and adverbs. \( S(w) \) is the set of synsets for word \( w \).

- **Opinion detection:**
  \[ \text{score}(s) = 1 - obj_s \quad \text{or} \quad \text{score}(s) = obj_s \]

- **Polarity classification:**
  \[ \text{score}(s) = pos_s - neg_s \]

**Pros:**
- no need for training corpora

**Cons:**
- low results
- need for POS tagger
Supervised sentiment analysis

Possible simple solution with lexical information:

Bag of words with Naïve Bayes

\[ h(D) = h(w_1, \ldots, w_n) = \arg \max_y P(y) \prod_{i=1}^{n} P(w_i|y) \]

where \( y \) is the category (positive/negative, subjective/objective), and \( w_1, \ldots, w_n \) is the bag of words related to \( D \)

- Given a training corpus \( C = \{d_i\} \) partitioned into subsets \( Y_1 \) and \( Y_2 \)
  - \( P(y) \approx P_{MLE}(y) = \frac{|Y_i|}{|C|} \)
  - \( P(w_i|y) \approx P_{MLE}(w_i|Y_j) = \frac{c(w_i, Y_j)}{\sum_{w_i \in Y_j} c(w_i, Y_j)} \)

Pros:
- higher results
- no need for POS tagger

Cons:
- need for training corpora
Hybrid approach for sentiment analysis

Possible solution with lexical information:

- Combine two supervised methods with SentiWordnet method
- I.e., consensuate the output of the three methods, using *voting*, for instance:
  
  if at least 2 of the methods answer *y* then output *y*
  
  else output the answer of the method with better accuracy in the training corpus

The combination improves the results of the isolated methods

Possible solution with compositional semantics: fine-tuning contextual embeddings (from Session 8 on). Better results, but contextual embeddings do not exist for every language or sublanguage!!
Annex

- Base on the Bayes’ theorem:

\[
P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}
\]

- Naïve assumption of independence between features:

\[
P(y|x_1, \ldots, x_n) \approx P(y) \prod_{i=1}^{n} P(x_i|y)
\]

- Maximum likelihood estimation of \(P(y)\) and \(P(x_i|y)\) as training model

- Test prediction as:

\[
h(x_1, \ldots, x_n) = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i|y)
\]

- Need a smoothing technique to avoid zero counts: in NLTK never seen features are discarded