Natural Language Processing
CS 6320
Lecture 2
NLP Resources

Instructor: Sanda Harabagiu
NLP Resources

- **WordNet**
  - [http://wordnet.princeton.edu](http://wordnet.princeton.edu)
- **Extended WordNet (XWN)**
  - [http://xwn.hlt.utdallas.edu](http://xwn.hlt.utdallas.edu)
- **SemCor**
- **FrameNet**
  - [http://framenet.icsi.berkeley.edu](http://framenet.icsi.berkeley.edu)
- **POS Taggers**
- **Syntactic Parsers**
- **Treebank**
WordNet 1/7

- **WordNet web page:**
  [http://wordnet.princeton.edu/~wn](http://wordnet.princeton.edu/~wn)

- A lexical database for the English language

- Developed by the Cognitive Science Laboratory at Princeton University (professor George A. Miller)

- Its design was inspired by current psycholinguistic theories of human lexical memory

- User friendly interface

  *Library of C functions: allows you to access the synsets and the relations between them directly from your programs*
Parts of speech covered by WordNet:
- Nouns
- Verbs
- Adjectives
- Adverbs

The fundamental unit of WordNet is the SYNSET (synonym set)
- each synset represents one underlying lexical concept
- Example of SYNSET: {car, auto, automobile, machine, motorcar}

Different relations link the synonym sets
WordNet 3/7

- **WordNet 2.1 (latest release)**
- "Car" has 5 senses in WN.
- Each synset has at least one definition and there could be some sentences using the words of the synset.
- For the previous synset, we have:
  - **Definition:** 4-wheeled motor vehicle; usually propelled by an internal combustion engine.
  - **Sentence:** "He needs a car to go to work."
- Other senses of "car" belong to different synsets, and have different definitions:
  1. {car, railcar, railroad car, railway car}
  2. {car, cable car}
  3. {car, gondola}
  4. {car, elevator car}
WordNet 4/7

- Relations between synsets:
  - Synonymy
  - Hypernymy (superordination)
  - Hyponymy (subordination)
  - Holonymy (whole to part relation)
  - Meronymy (part to whole relation)
  - Antonymy
  - Troponymy (particular way to do something)
• **Synonymy relation:**
  - (motor vehicle, automotive vehicle)
  - *Definition*: a self propelled wheeled vehicle that does not run on rails.

• **Hypernymy relation:**
  - (vehicle)
  - *Definition*: a conveyance that transports people or objects.

• **Hyponymy relation:**
  - (ambulance)
  - *Definition*: a vehicle that takes people to and from hospitals.
**Holonymy relation:**
- \( (\text{bicycle wheel}) \)
- **Definition:** the wheel of a bicycle
- **Has the holonym:**
  - \( (\text{bicycle, bike, wheel}) \)
  - **Definition:** has two wheels; moved by foot pedals

**Meronymy relation:**
- \( (\text{bicycle wheel}) \)
- **Definition:** the wheel of a bicycle
- **Has the meronym:**
  - \( (\text{spoke, radius}) \)
  - **Definition:** a radial member of a wheel joining the hub to the rim.
• **Antonymy relation:**
  • *(sweet)*
  • *Definition: having a pleasant taste (as of sugar)*
  • *Has the antonym:*
  • *(sour)*
  • *Definition: having a sharp biting taste.*

• **Troponymy relation:**
  • *(dream)*
  • *Definition: experience while sleeping.*
  • *Has the troponym:*
  • *(fantasize)*
  • *Definition: have fantasies.*
eXtended WordNet 1/6

- **eXtended WordNet web page:** [http://xwn.hlt.utdallas.edu](http://xwn.hlt.utdallas.edu)
  - Provides several important enhancements (over WordNet 2.0) intended to remedy the present limitations of WordNet
  - WordNet 2.0 glosses are syntactically parsed, transformed into logic forms and content words are semantically disambiguated
  - eXtended WordNet is an ongoing project at the Human Language Technology Research Institute ([http://www.hlt.utdallas.edu](http://www.hlt.utdallas.edu)), University of Texas at Dallas
  - second release- the next release scheduled for the end of 2004
eXtended WordNet 2/6

- For each WordNet 2.0 gloss, eXtended WordNet associates three types of information:
  - its parse tree
  - its logic form
  - each noun, verb, adjective and adverb of the gloss is semantically disambiguated (with respect to WordNet 2.0)

- Exploits the rich information contained in the definitional glosses

- Increases the connectivity between synsets

- Provides computer access to a broader context for each concept
eXtended WordNet 3/6

- Consists of four XML files--one for each part of speech:
  - Noun
  - Verb
  - Adjective
  - Adverb

- The eXtended WordNet may be used as a Core Knowledge Base for applications such as:
  1. Question Answering
  2. Information Retrieval
  3. Information Extraction
  4. Summarization
  5. Natural Language Generation
  6. Inferences
  7. other knowledge intensive applications
• The glosses contain a part of the world knowledge since they define the most common concepts of the English language.

• Problem: they are written in natural language (English), and the word are lexically and semantically ambiguous

  o Solution: Disambiguate the content words (nouns, verbs, adjectives and adverbs) from each gloss

  o How?

  • Three different methods, described in “WordNet 2- A Morphologically and Semantically Enhanced Resource”, by S. Harabagiu, G. Miller and D. Moldovan
### Method 1: Classes of heuristics for word sense disambiguation
- Method 2: Conceptual Density
- Method 3: Statistics on large corpora

#### Results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>87%</td>
</tr>
<tr>
<td>Method 2</td>
<td>80%</td>
</tr>
<tr>
<td>Method 3</td>
<td>68%</td>
</tr>
<tr>
<td>Methods 1+2</td>
<td>92%</td>
</tr>
<tr>
<td>Methods 1+2+3</td>
<td>94%</td>
</tr>
</tbody>
</table>

#### Example: `{tennis, lawn tennis}` has the GLOSS: (a game played with rackets by two or four players who hit a ball back and forth over a net that divides a tennis court)
- The gloss is disambiguated as:

  (a *game:*n#2 *played:*v#2 with *rackets:*n#4 by *two:*n#1 or *four:*n#1 *players:*n#1 who *hit:*v#1 a *ball:*n#1 *back and forth:*r#1 over a *net:*n#5 that *divides:*v#5 a *tennis court:*n#1)
• Glosses can also be transformed into logical formulae (axioms)
• A predicate is generated from every noun, verb, adjective and adverb encountered in every gloss.
• The name of the predicate is a concatenation of:
  • The base form of the word
  • The part-of-speech of the word (n,v,a,r)
  • The WordNet semantic sense of the word
• The predicates have arguments
  • Examples: e (eventuality), X1: subject; X2: object
• **Example:** *The Logical form of the gloss of tennis is:*

```
game:n#2(x2) & play:v#2(e1,x1,x2) & with(e1,x3) & racket:n#4(x3) & by(e1,x1) &
or(x1,x3,x4) & two:n#1(x3) & four:n#1(x4) & players:n#1(x1) & hit:v#1(e2,x1,x5) &
ball:n#1(x5) & back and forth:r#1(e2) & over(e2,x6) & net:n#5(x6) &
divides:v#5(e3,x6,x7) & tennis court:n#1(x7)
```
SemCor 1/6

- **SemCor’s webpage:**
  [http://multisemcor.itc.it/semcor.html](http://multisemcor.itc.it/semcor.html)

- **Developed by Professor George Miller at Princeton University in 1993.**

- **Contains 37176 sentences from 352 newspaper articles on different topics.**
  - A corpus tagged with WordNet senses.
  - An XML-like format: for each word has attached its POS, its lemma and its WordNet sense.
  - The different versions of SemCor depend on the WordNet version used for the senses.
The SemCor corpus, created by the Princeton University, is a subset of the English Brown corpus containing almost 700,000 running words.

In SemCor all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged according to Princeton WordNet 1.6.

More in detail, the SemCor corpus is composed of 352 texts.

- In 186 texts all the open class words (nouns, verbs, adjectives, and adverbs) are annotated with PoS, lemma and sense, while in the remaining 166 texts only verbs are annotated with lemma and sense.
- The "all-words" component of SemCor has 359,732 tokens among which 192,639 are semantically annotated, while the "only-verbs" component has 316,814 tokens among which 41,497 verb occurrences are semantically annotated.
An example from SemCor 2.0:
A Texas halfback who doesn't even know the team's plays, Eldon Moritz, ranks fourth in Southwest Conference scoring after three games.

```xml
<p pnum=1>
<s snum=1>
<wf cmd=done pos=DT ot=notag>A</wf>
<wf cmd=done pos=NN lemma=texas wnsn=1 lexsn=1:15:00::>Texas</wf>
<wf cmd=done pos=NN lemma=halfback wnsn=1 lexsn=1:18:00::>halfback</wf>
<wf cmd=ignore pos=WP>who</wf>
<wf cmd=done pos=VBZ ot=notag>does</wf>
<wf cmd=done pos=RB lemma=n't wnsn=0 lexsn=4:02:00::>n't</wf>
```
even
know
the
team
's
plays
Eldon_Moritz
ranks
Southwest Conference scoring after three games.
• **Word Sense Disambiguation (WSD).**

• **Learning algorithms for WSD could use this corpus as a training text.**

• **Projects that need data annotated with WordNet senses may use this corpus.**
• **FrameNet website:**
  http://framenet.icsi.berkeley.edu

• **Frames and Understanding**
  *Hypothesis:* People understand things by performing mental operations on what they already know. Such knowledge is describable in terms of information about known situations, and such packaged information is called **frames.**
• A Frame is a schematic representation of situations involving various participants, props and other conceptual roles, called **Frame Elements (FEs)**

• **Example: The frame THEFT**
  - it involves situations in which a PERPETRATOR takes GOODS that belonged to the VICTIM. The MEANS by which that is accomplished can also be expressed.
  - There are also annotations of these FEs from sentences selected from the British National Corpus (BNC)
The core work of FrameNet

- Characterized frames
- Find words that fit the frames (TARGETs of the frames) – also known as LEXICAL UNITS
- Develop descriptive terminology
- Extract sample sentences
- Annotate selected examples
- Derive “valence” descriptions
Sample Event Frame:

Commercial Transaction

**Initial state:**
- Vendor has Goods, wants Money
- Customer wants Goods, has Money

**Transition:**
- Vendor transmits Goods to Customer
- Customer transmits Money to Vendor

**Final State:**
- Vendor has Money
- Customer has Goods
**Meaning and Syntax**

- The various verbs that evoke this frame introduce the elements of the frame in different ways.
- The identities of the buyer, seller, goods and money (FEs)
- Information expressed in sentences containing these verbs, namely the FEs, occurs in different places in the sentence depending on the verb.
She **bought** some carrots from the greengrocer for a dollar.

She **paid** a dollar to the greengrocer for some carrots.

She **paid** the greengrocer a dollar for the carrots.
FrameNet Project:

- For every target word,
- *describe the frames or conceptual structures which underlie them*,
- and annotate example sentences that cover the ways in which information from the associated frames are expressed in these sentences
FrameNet Entities and Relations

- **Frames**
  - Background descriptions
  - Lexical units associated with each frame
- **Frame Elements (Roles)***
- **Binding Constraints**
  - **Identify**
- **ISA (x:Frame, y:Frame)**
- **SubframeOf (x:Frame, y:Frame)**
- **Subframe Ordering**
  - **precedes**
- **Annotations**
TreeBank web page:
http://www.cis.upenn.edu/~treebank

Treebank = a bank of linguistic trees

- Large corpus of texts annotated with syntactic information
- Penn Treebank Project has been developed at University of Pennsylvania
- POS (Part Of Speech) tagged
- Parsed trees
- Corpora:
  - Wall Street Journal
  - The Brown Corpus
  - Switchboard
  - ATIS
Large corpora:

Penn Treebank –

www.cis.upenn.edu/~treebank

All Treebank data is released through The Language Data Consortium (LDC)
• Corpus of 4.5 million words of American English
• Part_of_Speech tagged
• Syntactic Bracketing
An example of POS tagged text:

[ Mr./NNP Volk/NNP ]

/,

[ 55/CD years/NNS ]

old/JJ ,/ succeeds/VBZ

[ Duncan/NP Dwight/NP ]

/,

[ who/WP ]

retired/VBD in/IN

[ September/NP ]

./.
## Tagsets for English $\frac{1}{2}$ (POS)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordin. Conjunction</td>
<td>and, but, or</td>
<td>NNS</td>
<td>Noun, plural</td>
<td>llamas</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td>one, two, three</td>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>IBM</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>a, the</td>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Carolinas</td>
</tr>
<tr>
<td>EX</td>
<td>Existential ‘there’</td>
<td>there</td>
<td>PDT</td>
<td>Predeterminer</td>
<td>all, both</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td>mea culpa</td>
<td>POS</td>
<td>Possesive ending</td>
<td>‘s</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/sub-conj</td>
<td>of, in, by</td>
<td>PP</td>
<td>Personal pronoun</td>
<td>I, you, he</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>yellow</td>
<td>PP$</td>
<td>Possesive pronoun</td>
<td>your, one’s</td>
</tr>
<tr>
<td>JJR</td>
<td>Adjective, comparative</td>
<td>bigger</td>
<td>RB</td>
<td>Adverb</td>
<td>quickly, never</td>
</tr>
<tr>
<td>JJS</td>
<td>Adjective, superlative</td>
<td>wildest</td>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>faster</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td>1, 2, One</td>
<td>RBS</td>
<td>Adverb, superlative</td>
<td>fastest</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>can, should</td>
<td>RP</td>
<td>Particle</td>
<td>up, off</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, singular or mass</td>
<td>llama</td>
<td>SYM</td>
<td>Symbol</td>
<td>+, %, &amp;</td>
</tr>
</tbody>
</table>
## Tagsets for English 2/2 (POS)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>TO</td>
<td>“to”</td>
<td>to</td>
<td>WRB</td>
<td>Wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
<td>ah, oops</td>
<td>$</td>
<td>Dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>VB</td>
<td>Verb, base form</td>
<td>eat</td>
<td>#</td>
<td>Pound sign</td>
<td></td>
</tr>
<tr>
<td>VBD</td>
<td>Verb, past tense</td>
<td>ate</td>
<td>“</td>
<td>Left quote</td>
<td>‘ or “</td>
</tr>
<tr>
<td>VBG</td>
<td>Verb, gerund</td>
<td>eating</td>
<td>”</td>
<td>Right quote</td>
<td>’ or ”</td>
</tr>
<tr>
<td>VBN</td>
<td>Verb, past participle</td>
<td>eaten</td>
<td>(</td>
<td>Left parenthesis</td>
<td>[, (, {, &lt;</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, non-3sg. present</td>
<td>eat</td>
<td>)</td>
<td>Right parenthesis</td>
<td>}, ), }, &gt;</td>
</tr>
<tr>
<td>VBZ</td>
<td>Verb, 3sg. present</td>
<td>eats</td>
<td>,</td>
<td>Comma</td>
<td>,</td>
</tr>
<tr>
<td>WDT</td>
<td>Wh-determiner</td>
<td>which, that</td>
<td>.</td>
<td>Sentence-final punc</td>
<td>. ! ?</td>
</tr>
<tr>
<td>WP</td>
<td>Wh-pronoun</td>
<td>what, who</td>
<td>:</td>
<td>Mid-sentence punc</td>
<td>: ; ... -- -</td>
</tr>
<tr>
<td>WP$</td>
<td>Possessive wh-pronoun</td>
<td>whose</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The parse tree for previous text:

(S
  (NP (NP Mr. Volk)
    ,
    (ADJP (NP 55 years)
      old)
    ,)
  (VP succeeds
    (NP (NP Duncan Dwight)
      ,
      (SBAR
        (WHNP who)
        (S (NP T)
          (VP retired
            (PP in
              (NP September))))))))) .)
## The Syntactic Tag Set

<table>
<thead>
<tr>
<th>Tags</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ADJP</td>
<td>Adjective phrase</td>
</tr>
<tr>
<td>2. ADVP</td>
<td>Adverb phrase</td>
</tr>
<tr>
<td>3. NP</td>
<td>Noun phrase</td>
</tr>
<tr>
<td>4. PP</td>
<td>Prepositional phrase</td>
</tr>
<tr>
<td>5. S</td>
<td>Simple declarative clause</td>
</tr>
<tr>
<td>6. SBAR</td>
<td>Clause introduced by subordinating conjunction or 0 (see below)</td>
</tr>
<tr>
<td>7. SBARQ</td>
<td>Direct question introduced by wh-word or wh-phrase</td>
</tr>
<tr>
<td>8. SINV</td>
<td>Declarative sentence with subject-aux inversion</td>
</tr>
<tr>
<td>9. SQ</td>
<td>Subconstituent of SBARQ excluding wh-word or wh-phrase</td>
</tr>
<tr>
<td>10. VP</td>
<td>Verb phrase</td>
</tr>
<tr>
<td>11. WHADVP</td>
<td>wh-adverb phrase</td>
</tr>
<tr>
<td>12. WHNP</td>
<td>wh-noun phrase</td>
</tr>
<tr>
<td>13. WHPP</td>
<td>wh-prepositional phrase</td>
</tr>
<tr>
<td>14. X</td>
<td>Constituent of unknown or uncertain category</td>
</tr>
<tr>
<td>Null elements</td>
<td></td>
</tr>
<tr>
<td>1. *</td>
<td>“Understood” subject of infinitive or imperative</td>
</tr>
<tr>
<td>2. 0</td>
<td>Zero variant of that in subordinate clauses</td>
</tr>
<tr>
<td>3. T</td>
<td>Trace-marks position where moved wh-constituent is interpreted</td>
</tr>
<tr>
<td>4. NIL</td>
<td>Marks position where preposition is interpreted in pied-piping contexts</td>
</tr>
</tbody>
</table>
This information is very useful for learning POS and parsing algorithms.
The hand tagged text is a good source of correct examples for training the algorithms.
The parsing algorithms use TreeBank for training and measuring their accuracy.
Brill's tagger was trained on TreeBank.
Charniak's parser was trained on TreeBank.
Statistical processes.
Treebank 6/6: The Penn Treebank Statistics

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Tagged for Part-of-Speech</th>
<th>Skeletal Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dept. of Energy abstracts</td>
<td>231,404</td>
<td>231,404</td>
</tr>
<tr>
<td>Dow Jones Newswire stories</td>
<td>3,065,776</td>
<td>1,061,166</td>
</tr>
<tr>
<td>Dept. of Agriculture bulletins</td>
<td>78,555</td>
<td>78,555</td>
</tr>
<tr>
<td>Library of America texts</td>
<td>105,652</td>
<td>105,652</td>
</tr>
<tr>
<td>MUC-3 messages</td>
<td>111,828</td>
<td>111,828</td>
</tr>
<tr>
<td>IBM Manual sentences</td>
<td>89,121</td>
<td>89,121</td>
</tr>
<tr>
<td>WBUR radio transcripts</td>
<td>11,589</td>
<td>11,589</td>
</tr>
<tr>
<td>ATIS sentences</td>
<td>19,832</td>
<td>19,832</td>
</tr>
<tr>
<td>Brown Corpus, retagged</td>
<td>1,172,041</td>
<td>1,172,041</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>4,885,798</strong></td>
<td><strong>2,881,188</strong></td>
</tr>
</tbody>
</table>
The most popular Part-Of-Speech tagger is Brill's tagger

Implements a simple rule based tagger using transformation-based learning

**INPUT:** tokenized text -> **OUTPUT:** POS tagged text
(each word gets attached its POS)

Accuracy in the upper 90s percents

Can be trained on your own corpus

Can be augmented with new rules, making possible its tuning for a certain domain

Has been used in eXtended WordNet development (an enhanced version)

State-of-the-art POS tagger
**INPUT:** I saw the boy in the park with the telescope.

**OUTPUT:** I/PRP saw/VBD the/DT boy/NN in/IN the/DT park/NN with/IN the/DT telescope/NN ./.
Syntactic Parsers

• The most popular syntactic parser in NLP community is Charniak's parser
• A maximum-entropy-inspired parser
• INPUT: plain text -> OUTPUT: parse trees (each sentence gets its parse tree)
• Accuracy around 90%
• Has been trained on TreeBank
• Has been used in eXtended WordNet development
• State-of-the-art syntactic parser
Syntactic Parser (example)

**INPUT:**

<s> I saw the boy in the park with the telescope . </s>

**OUTPUT:**

(S1 (S (NP (PRP I) )
   (VP (VBD saw)
   (NP (DT the) (NN boy) )
   (PP (IN in)
   (NP (NP (DT the) (NN park) )
   (PP (IN with)
   (NP (DT the) (NN telescope) ) ) ) ) ) ) (. .) )