Need for NLP

- Vector space model limitations
  - Words in combination carry more/different meaning than isolation
  - President flew
    - to Washington
    - from the Revolution
- Words can mean different things
- Relative importance of different words
- Words vs. Concepts
Different meanings

- NLP Task: *Word Sense Disambiguation*
  - Given word, dictionary of multiple meanings
  - Determine from context which meaning applies
- Hard problem
  - SensEval 3 (2004): 65% accuracy

“Winner”: GAMBL
Decadt, Hoste, Daelemans, Bosch
“Winner”: GAMBL
Decadt, Hoste, Daelemans, Bosch

• Initial phase: Linguistic analysis
  – Tokenize
  – Part-of-speech
  – Grammatical relations

• Training data
  – Senseval-3 task (7860 words)
  – SemCor (WordNet), previous SenseEval (555,269 words)
“Winner”: GAMBL
Decadt, Hoste, Daelemans, Bosch

- Cascaded Classifiers
  - First stage: Broad context
    - Three sentences
    - Instance-based learning
  - Second stage: Narrow context
    - Seven words
    - Result of 1st classifier
    - Genetic algorithm

- NLP Task: *Word Sense Disambiguation*
  - Given word, dictionary of multiple meanings
  - Determine from context which meaning applies

- Hard problem
  - SensEval 3 (2004): 65% accuracy
    - “just choose most frequent sense” 60%
    - Inter-annotator agreement 72.5%
Words vs. Concepts

- Named Entity Recognition
  - People
  - Places
  - Organizations
  - Dates
  - ...
  *Success story – effective, learn new types of NER*
- Coreference Resolution
  - Different names for same entity in same document

Example: SpaCy
*(Sarkar, Intel)*

US GPE unveils world’s most powerful supercomputer, beats China GPE. The US GPE has unveiled the world’s most powerful supercomputer called ‘Summit’, beating the previous record-holder China GPE’s Sunway TaihuLight ORG. With a peak performance of 200,000 CARDINAL trillion calculations per second ORDINAL, it is over twice as fast as Sunway TaihuLight ORG, which is capable of 90,000 CARDINAL trillion calculations per second. Summit has 4,608 CARDINAL servers, which reportedly take up the size of two CARDINAL tennis courts.
## CoNLL-2003 Data
*(Tjong&Meulder’03)*

<table>
<thead>
<tr>
<th>Word</th>
<th>Part of Speech</th>
<th>Chunk tag</th>
<th>Named Entity tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.N.</td>
<td>NNP</td>
<td>I-NP</td>
<td>I-ORG</td>
</tr>
<tr>
<td>Official</td>
<td>NN</td>
<td>I-NP</td>
<td>O</td>
</tr>
<tr>
<td>Ekeus</td>
<td>NNP</td>
<td>I-NP</td>
<td>I-PER</td>
</tr>
<tr>
<td>Heads</td>
<td>VBZ</td>
<td>I-VP</td>
<td>O</td>
</tr>
<tr>
<td>For</td>
<td>IN</td>
<td>I-PP</td>
<td>O</td>
</tr>
<tr>
<td>Baghdad</td>
<td>NNP</td>
<td>I-NP</td>
<td>I-LOC</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

## Features Used
*(Tjong&Meulder’03)*

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lex</td>
<td>+</td>
</tr>
<tr>
<td>pos</td>
<td>+</td>
</tr>
<tr>
<td>aff</td>
<td>+</td>
</tr>
<tr>
<td>pre</td>
<td>+</td>
</tr>
<tr>
<td>ort</td>
<td>+</td>
</tr>
<tr>
<td>gaz</td>
<td>+</td>
</tr>
<tr>
<td>chu</td>
<td>-</td>
</tr>
<tr>
<td>pat</td>
<td>+</td>
</tr>
<tr>
<td>cas</td>
<td>-</td>
</tr>
<tr>
<td>tri</td>
<td>-</td>
</tr>
<tr>
<td>bag</td>
<td>-</td>
</tr>
<tr>
<td>quo</td>
<td>-</td>
</tr>
<tr>
<td>doc</td>
<td>-</td>
</tr>
</tbody>
</table>

**Aff:** affix information (n-grams); **bag:** bag of words; **cas:** global case information; **chu:** chunk tags; **doc:** global document information; **gaz:** gazetteers; **lex:** lexical features; **ort:** orthographic information; **pat:** orthographic patterns (like Aa0); **pos:** part-of-speech tags; **pre:** previously predicted NE tags; **quo:** flag signing that the word is between quotes; **tri:** trigger words.
NER – CoNLL 2003 Winner
*Florian, Ittycheriah, Jing, Zhang*

- Label each word
  - Start, continue, or end a named entity
- Key: good features
  - Words and part of speech, 5 word window
  - Prefix, suffixes of surrounding words
  - Word “flags” such as *firstCap, 2digit, allCaps*
  - Gazetteer – 130k known names
  - Output of existing NER systems trained for different output categories

Winner: Ensemble
*Florian, Ittycheriah, Jing, Zhang*

- Multiple classifiers
  - Robust risk minimization
  - Maximum entropy
  - Transformation-based learning
  - Hidden Markov model
- Weighted voting
- Results: 89% accuracy
  - Baseline 60%
Stanford NER
(Finkel, Grenager, Manning ‘05)

- Based on Conditional Random Fields (CRF)
  - Probabilistic Language Modeling Approach
- Can be trained on any corpus
  - Models tested for CoNLL in multiple languages, among others
- Available as Open Source (GPL)
  - https://nlp.stanford.edu/software/CRF-NER.shtml

Template Analysis

Named Entity Recognition on Steroids

- Given a “template” of desired structured information
  - Fill in fields of template from analysis of document
- Fields:
  - Entities (named entities)
  - Relationships
  - Time/date/order
Template Analysis: Example

<table>
<thead>
<tr>
<th>NAME:</th>
<th>Fletcher Maddox Maddox</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESCRIPTOR:</td>
<td>former Dean of the UCSD Business School his father the firm's CEO</td>
</tr>
<tr>
<td>CATEGORY:</td>
<td>PERSON</td>
</tr>
<tr>
<td>NAME:</td>
<td>Oliver</td>
</tr>
<tr>
<td>DESCRIPTOR:</td>
<td>His son Chief Scientist</td>
</tr>
<tr>
<td>CATEGORY:</td>
<td>PERSON</td>
</tr>
<tr>
<td>NAME:</td>
<td>Ambrose</td>
</tr>
<tr>
<td>DESCRIPTOR:</td>
<td>Oliver's brother the CFO of L.J.G.</td>
</tr>
<tr>
<td>CATEGORY:</td>
<td>PERSON</td>
</tr>
<tr>
<td>NAME:</td>
<td>UCSD Business School</td>
</tr>
<tr>
<td>DESCRIPTOR:</td>
<td></td>
</tr>
<tr>
<td>CATEGORY:</td>
<td>ORGANIZATION</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PERSON</th>
<th>Employee_of</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fletcher Maddox</td>
<td>Employee_of</td>
<td>UCSD Business School</td>
</tr>
<tr>
<td>Fletcher Maddox</td>
<td>Employee_of</td>
<td>La Jolla Genomics</td>
</tr>
<tr>
<td>Oliver Ambrose</td>
<td>Employee_of</td>
<td>La Jolla Genomics</td>
</tr>
<tr>
<td>Oliver</td>
<td>Employee_of</td>
<td>La Jolla Genomics</td>
</tr>
<tr>
<td>Employee_of</td>
<td>UCSD Business School</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARTIFACT</th>
<th>Product_of</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geninfo</td>
<td>Product_of</td>
<td>La Jolla Genomics</td>
</tr>
<tr>
<td>La Jolla</td>
<td>Product_of</td>
<td>La Jolla Genomics</td>
</tr>
<tr>
<td>Genomatics</td>
<td>Location_of</td>
<td>La Jolla Genomics</td>
</tr>
<tr>
<td>UCSD Business School</td>
<td>Location_of</td>
<td>La Jolla Genomics</td>
</tr>
</tbody>
</table>

Message Understanding Conferences

Nance, who is also a paid consultant to ABC News, said ...
SIFT: **Language model approach**
   - Uses Hidden Markov Models

SIFT: **Sentence-level model**
   - Part of speech
   - Named Entity
   - Parse (grammatical)
   - Relationships

SIFT: **Uses “outside” training data**
   - Penn Treebank, additional domain-specific text
SIFT: Additional semantics

- Further breakdown (e.g., distinguish title from name in Named Entity)
- Semantic labeling
- Co-reference
- *Probability labels for all of these*

SIFT: Sentence-level output
Cross-Sentence Model

- Similar approach
- Uses sentence parse/labeling as input

Basic Tools

- Part of Speech tagging
- Sentence diagramming