COMP 790.139 (Fall 2016)  
Natural Language Processing  
(with some vision, robotics, and deep learning)  

Aug 31, 2016  

Mohit Bansal  

(various slides adapted/borrowed from courses by Dan Klein, Chris Manning)
Announcements

- Must have received system email from me reg. readings
- Tentative schedule and paper list decided
- Paper list on website
- Then start deciding which topic you would like to present
- Send me top choices and I will create an assignment ((I will notify you via email))
- Some topics will be in groups
NLP Basics and Core Tasks 1

- Part-of-Speech Tagging
- Syntactic Parsing: Constituent, Dependency, CCG, others
- Coreference Resolution
- Distributional Semantics: PMI, Neural, CCA
- Compositional Semantics: Logical-form, Semantic Parsing, Vector-form, Neural (RNNs/CNNs)

Note: we will be covering some of these briefly (so as to be able to reach the paper reading weeks quickly), so definitely follow up for more details in the prescribed readings and references, and talk to me in office hours!
Part-of-Speech Tagging

- Tag sequence of words with syntactic categories (noun, verb, preposition, …)

- Useful in itself:
  - Text-to-speech: *read, lead, record*
  - Lemmatization: *saw*[^v] → *see*, *saw*[^n] → *saw*
  - Shallow Chunking: grep `{JJ | NN}* {NN | NNS}`

- Useful for downstream tasks (e.g., in parsing, and as features in various word/text classification tasks)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>conjunction, coordinating</td>
</tr>
<tr>
<td>CD</td>
<td>numeral, cardinal</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
</tr>
<tr>
<td>IN</td>
<td>preposition or conjunction, subordinating</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective or numeral, ordinal</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
</tr>
<tr>
<td>JJS</td>
<td>adjective, superlative</td>
</tr>
<tr>
<td>MD</td>
<td>modal auxiliary</td>
</tr>
<tr>
<td>NN</td>
<td>noun, common, singular or mass</td>
</tr>
<tr>
<td>NNP</td>
<td>noun, proper, singular</td>
</tr>
<tr>
<td>NNPS</td>
<td>noun, proper, plural</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, common, plural</td>
</tr>
<tr>
<td>POS</td>
<td>genitive marker</td>
</tr>
<tr>
<td>PRP</td>
<td>pronoun, personal</td>
</tr>
<tr>
<td>PRPS</td>
<td>pronoun, possessive</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
</tr>
<tr>
<td>TO</td>
<td>&quot;to&quot; as preposition or infinitive marker</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, present participle or gerund</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, present tense, not 3rd person singular</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, present tense, 3rd person singular</td>
</tr>
<tr>
<td>WDT</td>
<td>WH-determiner</td>
</tr>
<tr>
<td>WP</td>
<td>WH-pronoun</td>
</tr>
<tr>
<td>WP$</td>
<td>WH-pronoun, possessive</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
</tr>
</tbody>
</table>

Examples:
- and both but either or
- mid-1890 nine-thirty 0.5 one
- a all an every no that the there
- gemeinschaft hund ich jeux among whether out on by if
- third ill-mannered regrettable bravest cheapest tallest can may might will would
- cabbage thermostat investment subhumanity Motown Cougar Yvette Liverpool Americans Materials States undergraduates bric-a-brac averages 's hers himself it we them her his mine my our ours their thy your occasionally maddeningly adventurously further gloomier heavier less-perfectly best biggest nearest worst aboard away back by on open through to huh howdy uh whammo shucks heck asked bring fire see take pleased swiped registered saw stirring focusing approaching erasing dilapidated imitated reunified unsettled twist appear comprise mold postpone bases reconstructs marks uses that what whatever which whichever that what whatever which who whom whose however whenever where why
Part-of-Speech Ambiguities

A word can have multiple parts of speech

VBD     VB
VBN     VBZ    VBP     VBZ
NNP     NNS    NN     NNS    CD    NN

Fed raises interest rates 0.5 percent

Mrs./NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN
Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD

Disambiguating features: lexical identity (word), context, morphology (suffixes, prefixes), capitalization, gazetteers (dictionaries), …
Classical Solution: HMMs

Trigram HMM: states = tag-pairs

Estimating Transitions: Standard smoothing w/ backoff

Estimating Emissions: Use unknown word classes (affixes, shapes) and estimate $P(t|w)$ and invert

Inference: choose most likely (Viterbi) sequence under model

$P(s, w) = \prod_i P(s_i|s_{i-1})P(w_i|s_i)$

[Brants, 2000]
POS Tagging: Other Models

- Discriminative sequence models with richer features: MEMMs, CRFs (SoA $\approx 97\%/90\%$ known/unknown)

- Universal POS tagset for multilingual and cross-lingual tagging and parsing [Petrov et al., 2012]

  12 tags: NOUN, VERB, ADJ, ADV, PRON, DET, ADP, NUM, CONJ, PRT, ., X

- Unsupervised tagging also works reasonably well!
  [Yarowsky et al., 2001; Xi and Hwa, 2005; Berg-Kirkpatrick et al., 2010; Christodoulopoulos et al., 2010; Das and Petrov, 2011]

[Brill, 1995; Ratnaparkhi, 1996; Toutanova and Manning, 2000; Toutanova et al., 2003]
Syntactic Parsing -- Constituent

- Phrase-structure parsing or Bracketing

```
S
  /\  /
 NP  VP
   /\  /
   NNP VBD NP
       /\  /
       John met PRP
           /\  /
           her  
```

- Demos: [http://tomato.banatao.berkeley.edu:8080/parser/parser.html](http://tomato.banatao.berkeley.edu:8080/parser/parser.html)
A context-free grammar is a tuple $<N, T, S, R>$

$N$ : the set of non-terminals
   Phrasal categories: $S$, $NP$, $VP$, $ADJP$, etc.
   Parts-of-speech (pre-terminals): $NN$, $JJ$, $DT$, $VB$

$T$ : the set of terminals (the words)

$S$ : the start symbol
   Often written as $ROOT$ or $TOP$
   *Not* usually the sentence non-terminal $S$

$R$ : the set of rules
   Of the form $X \rightarrow Y_1 Y_2 \ldots Y_k$, with $X, Y_i \in N$
   Examples: $S \rightarrow NP\ VP$, $VP \rightarrow VP\ CC\ VP$
   Also called rewrites, productions, or local trees
Probabilistic Context-free Grammars

A PCFG:

Adds a top-down production probability per rule $P(Y_1 Y_2 \ldots Y_k \mid X)$

Allows us to find the ‘most probable parse’ for a sentence

The probability of a parse is just the product of the probabilities of the individual rules
Treebank PCFG

- Need a PCFG for broad coverage parsing
- Extracting a grammar right off the trees is not effective:

```
ROOT
   |   
   S
   /   \
NP    VP
   |   |   |
PRP  VBD  ADJP
   |   |   |
He  was  JJ
   |   |   |
right
```

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT → S</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>S → NP VP .</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NP → PRP</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>VP → VBD ADJP</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model | F1
------|-----
Baseline | 72.0

[Charniak, 1996]
Grammar Refinement

Conditional independence assumptions often too strong! Not every NP expansion can fill every NP slot

Better results by enriching the grammar e.g.,

- Lexicalization [Collins, 1999; Charniak, 2000]
Grammar Refinement

- Conditional independence assumptions often too strong! Not every NP expansion can fill every NP slot

- Better results by enriching the grammar e.g.,
  - Lexicalization [Collins, 1999; Charniak, 2000]
Grammar Refinement

Conditional independence assumptions often too strong! Not every NP expansion can fill every NP slot

Better results by enriching the grammar e.g.,

- Lexicalization [Collins, 1999; Charniak, 2000]
- Latent Tag-splitting [Matsuzaki et al., 2005; Petrov et al., 2006]
bestScore(s)
for (i : [0,n-1])
    for (X : tags[s[i]])
        score[X][i][i+1] = tagScore(X,s[i])
for (diff : [2,n])
    for (i : [0,n-diff])
        j = i + diff
        for (X->YZ : rule)
            for (k : [i+1, j-1])
                score[X][i][j] = max{score[X][i][j], score(X->YZ) * score[Y][i][k] * score[Z][k][j]}

[CKY Parsing Algorithm (Bottom-up)]

[Cocke, 1970; Kasami, 1965; Younger, 1967]
Some Results

- Collins, 1999 → 88.6 F1 (generative lexical)
- Charniak and Johnson, 2005 → 89.7 / 91.3 F1 (generative lexical / reranking)
- Petrov et al., 2006 → 90.7 F1 (generative unlexical)
- McClosky et al., 2006 – 92.1 F1 (generative + reranking + self-training)
Syntactic Parsing -- Dependency

- Predicting directed head-modifier relationship pairs

Syntactic Parsing -- Dependency

- Pure (projective, 1\textsuperscript{st} order) dependency parsing is only cubic [Eisner, 1996]

- Non-projective dependency parsing useful for Czech & other languages – MST algorithms [McDonald et al., 2005]

---

root John saw a dog yesterday which was a Yorkshire Terrier
Parsing: Other Models and Methods

- Transition-based Dependency Parsing [Yamada and Matsumoto, 2003; Nivre, 2003]
- Tree-Insertion Grammar, DOP [Schabes and Waters, 1995; Hwa, 1998; Scha, 1990; Bod, 1993; Goodman, 1996; Bansal and Klein, 2010]
- Tree-Adjoining Grammar [Resnik, 1992; Joshi and Schabes, 1998; Chiang, 2000]
- Shift-Reduce Parser [Nivre and Scholz, 2004; Sagae and Lavie, 2005]
- Other: Reranking, A*, K-Best, Self-training, Co-training, System Combination, Cross-lingual Transfer [Sarkar, 2001; Steedman et al., 2003; Charniak and Johnson, 2005; Hwa et al., 2005; Huang and Chiang, 2005; McClosky et al., 2006; Fossum and Knight, 2009; Pauls and Klein, 2009; McDonald et al., 2011]

Clean the dishes in the sink.
They considered running the ad during the Super Bowl.

[Nakov and Hearst 2005; Pitler et al., 2010; Bansal and Klein, 2011]
Web Features for Syntactic Parsing

They considered running the ad during the Super Bowl.

Web Ngrams

\[ \text{count(} \text{running it during}\text{)} > \text{count(} \text{considered it during}\text{)} \]

- 7-10% relative error reduction over 90-92% parsers

[Bansal and Klein, 2011]
Unsup. Representations for Parsing

- Discrete or continuous, trained on large amounts of context

**BROWN (Brown et al., 1992):**

![Brown word cluster hierarchy diagram]

**SKIPGRAM (Mikolov et al., 2013):**

![Skipgram representation diagram]
Unsup. Representations for Parsing

- Condition on dependency context instead of linear, then convert each dependency to a tuple:

\[
\text{dep label} \quad \text{grandparent} \quad \text{parent} \quad \text{child} \quad \text{dep label}
\]

\[
[ PMOD_{<L>} \quad \text{regulation}_{<G>} \quad \text{of} \quad \text{safety} \quad PMOD_{<L>} ]
\]

- 10% rel. error reduction over 90-92% parsers

[Mrs., Mrs., Ms., Prof., III, Jr., Dr.]
[Jeffrey, William, Dan, Robert, Stephen, Peter, John, Richard, ...]
[Portugal, Iran, Cuba, Ecuador, Greece, Thailand, Indonesia, ...]

[his, your, her, its, their, my, our]
[Your, Our, Its, My, His, Their, Her]
[truly, wildly, politically, financially, completely, potentially, ...]
Visual Recognition Cues

- Joint parsing and image recognition

*the mug on the table with a crack*

[Christie et al., 2016]
Visual Recognition Cues

- Joint parsing and image recognition

the mug on the table with a crack

red chair and table
light green table

[Christie et al., 2016]
10-min Break?
President Barack Obama received the Serve America Act after congress’ vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

- Mentions to entity/event clusters
President Barack Obama received the Serve America Act after congress' vote. He signed the bill ...

Mention-pair Models

Pair-wise classification approach:

\[ \text{Features } f \rightarrow \text{Pair-wise classifier} \rightarrow \text{coref}(a_1, m) \]

\[ (a_1, m) \]

[Soon et al. 2001, Ng and Cardie 2002; Bengtson and Roth, 2008; Stoyanov et al., 2010]
Mention-pair Model

For each mention $m$, $\hat{a}_m = \operatorname{argmax}_{a_i \in A(m)} \text{coref}(a_i, m)$.

[Soon et al. 2001, Ng and Cardie 2002; Bengtson and Roth, 2008; Stoyanov et al., 2010]
# Standard features

<table>
<thead>
<tr>
<th>Type</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEXICAL</td>
<td>SOON_STR</td>
<td>Do the strings match after removing determiners ?</td>
</tr>
<tr>
<td>GRAMMATICAL</td>
<td>NUMBER</td>
<td>Do NP(_i) and NP(_j) agree in number ?</td>
</tr>
<tr>
<td></td>
<td>GENDER</td>
<td>Do NP(_i) and NP(_j) agree in gender ?</td>
</tr>
<tr>
<td></td>
<td>APPOSITIVE</td>
<td>Are the NPs in an appositive relationship ?</td>
</tr>
<tr>
<td>SEMANTIC</td>
<td>WORDNET_CLASS</td>
<td>Do NP(_i) and NP(_j) have the same WordNet class ?</td>
</tr>
<tr>
<td></td>
<td>ALIAS</td>
<td>Is one NP an alias of the other ?</td>
</tr>
<tr>
<td>POSITIONAL</td>
<td>SENTNUM</td>
<td>Distance between the NPs in terms of # of sentences</td>
</tr>
</tbody>
</table>

- **Weaknesses:** All pairs, Transitivity/Independence errors (He – Obama – She), Insufficient information

[Soon et al. 2001, Ng and Cardie 2002; Bengtson and Roth, 2008; Stoyanov et al., 2010]
Entity-centric Models

- Each coreference decision is globally informed by previously clustered mentions and their shared attributes.

- Lee et al., 2013’s deterministic (rule-based) system: multiple, cautious sieves from high to low precision.

- Durrett et al., 2013’s entity-level model is discriminative, probabilistic using factor graphs and BP.

[Hayighi and Klein, 2009; Lee et al., 2013; Durrett et al., 2013]
Mention-Ranking Models (Learned)

- Log-linear model to select at most 1 antecedent for each mention or determine that it begins a new cluster.

\[ Pr(A_i = a | x) \propto \exp(w^\top f(i, a, x)) \]

- Recent work (Wiseman et al., 2016, Clark & Manning, 2016) has used NNs for non-linear and vector-space coreference features to achieve SoA!

*Recent work* (Wiseman et al., 2016, Clark & Manning, 2016) has used NNs for non-linear and vector-space coreference features to achieve SoA!
Adding Knowledge to Coref

- **External corpora:** Web, Wikipedia, YAGO, FrameNet, Gender/Number/Person lists/classifiers, 3D Images, Videos

- **Methods:**
  - Self-training, Bootstrapping
  - Co-occurrence, Distributional, and Pattern-based Features
  - Entity Linking
  - Visual Cues from 3D Images and Videos

- Daumé III and Marcu, 2005; Markert and Nissim, 2005; Bergsma and Lin, 2006; Ponzetto and Strube, 2006; Haghighi and Klein, 2009; Kobdani et al., 2011; Rahman and Ng, 2011; Bansal and Klein, 2012; Durrett and Klein, 2014; Kong et al., 2014; Ramanathan et al., 2014
Web Features for Coreference

count(Obama * president) vs count(Jobs * president)

When Obama met Jobs, the president discussed the …

[Bansal and Klein, 2012]
When Obama met Jobs, the … He signed bills that …

\[
\text{count}(\text{Obama signed bills}) \text{ vs } \text{count}(\text{Jobs signed bills})
\]

<table>
<thead>
<tr>
<th></th>
<th>MUC F1</th>
<th>B3 F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>69.5</td>
<td>77.0</td>
</tr>
<tr>
<td>+Coocc</td>
<td>69.8</td>
<td>78.0</td>
</tr>
<tr>
<td>+Hearst</td>
<td>70.0</td>
<td>80.0</td>
</tr>
<tr>
<td>+Entity</td>
<td>70.4</td>
<td>82.0</td>
</tr>
<tr>
<td>+Cluster</td>
<td>70.7</td>
<td></td>
</tr>
<tr>
<td>+Pronoun</td>
<td>71.3</td>
<td></td>
</tr>
</tbody>
</table>

Results:

- **Haghighi & Klein, 2010**
  - MUC F1: 67.0
  - B3 F1: 77.0

- **Us**
  - MUC F1: 69.1
  - B3 F1: 80.0
Visual Cues for Coreference

- Joint coreference and 3D image recognition

<table>
<thead>
<tr>
<th></th>
<th>MUC</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>Stanford</td>
<td>61.56</td>
<td>62.59</td>
</tr>
<tr>
<td>Ours</td>
<td>83.69</td>
<td>51.08</td>
</tr>
</tbody>
</table>

Words occurring in similar context have similar linguistic behavior (meaning) [Harris, 1954; Firth, 1957]

Traditional approach: context-counting vectors
- Count left and right context in window
- Reweight with PMI or LLR
- Reduce dimensionality with SVD or NNMF


More word representations: hierarchical clustering based on bigram LM LL [Brown et al., 1992]
Unsupervised Embeddings

Vector space representations learned on unlabeled linear context (i.e., left/right words): distributional semantics (Harris, 1954; Firth, 1957)
Newer approach: context-predicting vectors (NNs)

SENNA [Collobert and Weston, 2008; Collobert et al., 2011]: Multi-layer DNN w/ ranking-loss objective; BoW and sentence-level feature layers, followed by std. NN layers. Similar to [Bengio et al., 2003].

\[
i\text{-th output} = P(w_t = i \mid \text{context})
\]
Distributional Semantics -- NNs

HUANG [Huang et al., 2012]: Add global, document-level context

<table>
<thead>
<tr>
<th>Local Context</th>
<th>Global Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>sum</td>
</tr>
<tr>
<td>score(_l)</td>
<td>score(_g)</td>
</tr>
</tbody>
</table>

Document
- river
- play
- shore
- water

global semantic vector

weighted average
CBOW, SKIP, word2vec [Mikolov et al., 2013]: Simple, super-fast NN w/ no hidden layer. Continuous BoW model predicts word given context, skip-gram model predicts surrounding context words given current word

Other: [Mnih and Hinton, 2007; Turian et al., 2010]

Distributional Semantics

- Other approaches: spectral methods, e.g., CCA
  - Word-context correlation [Dhillon et al., 2011, 2012]
  - Multilingual correlation [Faruqui and Dyer, 2014; Lu et al., 2015]

- Some recent directions: Train task-tailored embeddings to capture specific types of similarity/semantics, e.g.,
  - Dependency context [Bansal et al., 2014, Levy and Goldberg, 2014]
  - Predicate-argument structures [Hashimoto et al., 2014; Madhyastha et al., 2014]
  - Lexicon evidence (PPDB, WordNet, FrameNet) [Xu et al., 2014; Yu and Dredze, 2014; Faruqui et al., 2014; Wieting et al., 2015]