LING/C SC 581:
Advanced Computational Linguistics

Lecture Notes
Feb 3rd
Administrivia

• Homework 3 is out today
• No lecture next week
• Therefore due date for homework 3 will be the following Tuesday by midnight
• You are welcome to listen to Prof. Bryan Heidorn of the School of Information this afternoon in my seminar class (Douglass 211; 3:30pm–6pm)
Today's Topics

• Revisiting *Colorless Green Ideas* and the Penn Treebank (WSJ)
• tregex lecture
• Homework 3
• examples
  – (1) colorless green ideas sleep furiously
  – (2) furiously sleep ideas green colorless

• Chomsky (1957):
  – . . . It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally `remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.

• idea
  – (1) is syntactically valid,
  – (2) is word salad

One piece of supporting evidence:
(1) pronounced with normal intonation
(2) pronounced like a list of words ...
Colorless green ideas

• **Sentences:**
  - (1) **colorless green ideas** sleep furiously
  - (2) furiously sleep ideas green colorless

• **Statistical Experiment (Pereira 2002)**

\[
p(w_1 \cdots w_n) = p(w_1) \prod_{i=2}^{n} p(w_i | w_{i-1}) .
\]

Using this estimate for the probability of a string and an aggregate model with \( C = 16 \) trained on newspaper text using the expectation-maximization (EM) method (Dempster, Laird, & Rubin, 1977), we find that

\[
\frac{p(\text{Colorless green ideas sleep furiously})}{p(\text{Furiously sleep ideas green colorless})} \approx 2 \times 10^5 .
\]

Thus, a suitably constrained statistical model, even a very simple one, can meet Chomsky’s particular challenge.
Part-of-Speech (POS) Tag Sequence

• Chomsky's example:
  – colorless green ideas sleep furiously
  – JJ JJ NNS VBP RB (POS Tags)

• Similar but grammatical example:
  – revolutionary new ideas appear infrequently
  – JJ JJ NNS VBP RB

11    colorless green ideas sleep furiously
This distinction can be made (in this case, but not in many others that will concern us) by demonstrating that 11 is an instance of the sentence form Adjective-Adjective-Noun-Verb-Adverb, which is grammatical by virtue of such sentences as

13    revolutionary new ideas appear infrequently
that might well occur in normal English.
Penn Treebank (PTB) Corpus

How is it used?
- One million words of 1989 Wall Street Journal (WSJ) material
- nearly 50,000 sentences (25 sections)
- Training: (2–21) 39,832 sentences
- Evaluation: (23) 2,416 sentences

- Standard practice

Penn Treebank (PTB) Corpus

How is it used?
- One million words of 1989 Wall Street Journal (WSJ) material
- nearly 50,000 sentences (25 sections)
- Training: (2–21) 39,832 sentences
- Evaluation: (23) 2,416 sentences

- Standard practice
Stanford Parser

- Stanford Parser:
  
  - a probabilistic PS parser trained on the Penn Treebank

**Tagging**

Colorless/JJ green/JJ ideas/NNS sleep/VBP furiously/RB

**Tagging**

Revolutionary/NNP new/JJ ideas/NNS appear/VBP infrequently/RB
Stanford Parser

• Stanford Parser: a probabilistic PS parser trained on the Penn Treebank
Penn Treebank (PTB)

- **Corpus**: word frequencies:

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>colorless</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>green</td>
<td>NNP</td>
<td>33</td>
</tr>
<tr>
<td>JJ</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>ideas</td>
<td>NNS</td>
<td>32</td>
</tr>
<tr>
<td>sleep</td>
<td>VB</td>
<td>5</td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>VBP</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>NNP</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>furiously</td>
<td>RB</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>revolutionary</td>
<td>JJ</td>
<td>6</td>
</tr>
<tr>
<td>NNP</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>new</td>
<td>JJ</td>
<td>1795</td>
</tr>
<tr>
<td>NNP</td>
<td></td>
<td>1459</td>
</tr>
<tr>
<td>NNPS</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ideas</td>
<td>NNS</td>
<td>32</td>
</tr>
<tr>
<td>appear</td>
<td>VB</td>
<td>55</td>
</tr>
<tr>
<td>VBP</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>infrequently</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Stanford Parser

• Structure of NPs:
  – colorless green ideas
  – revolutionary new ideas

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>( [_{NP} \text{JJ JJ NNS}] )</td>
<td>1073</td>
</tr>
<tr>
<td>( [_{NP} \text{NNP JJ NNS}] )</td>
<td>61</td>
</tr>
</tbody>
</table>
An experiment

• examples
  – (1) colorless green ideas sleep furiously
  – (2) furiously sleep ideas green colorless

• Question:
  – Is (1) the most likely permutation of these particular five words?
Parsing Data

• All 5! (=120) permutations of
  – colorless green ideas sleep furiously .
• The winning sentence was:
  ①**furiously ideas sleep colorless green**.

- after training on sections 02-21 (approx. 40,000 sentences)

- **sleep** selects for ADJP object with 2 heads
- adverb (RB) **furiously** modifies noun
The next two highest scoring permutations were:

② *Furiously green ideas sleep colorless.*

③ *Green ideas sleep furiously colorless.*

*sleep* takes NP object  
*sleep* takes ADJP object
(Pereira 2002) compared Chomsky’s original minimal pair:

23. colorless green ideas sleep furiously
36. furiously sleep ideas green colorless

Ranked #23 and #36 respectively out of 120
But ...

- graph (next slide) shows how arbitrary these rankings are
  - when trained on randomly chosen sections covering 14K-31K sentences

Example:

#36 furiously sleep ideas green colorless
outranks
#23 colorless green ideas sleep furiously (and the top 3)
over much of the training space

Example: Chomsky's original sentence

#23 colorless green ideas sleep furiously
outranks both the top 3 and #36 just briefly at one data point
Sentence Rank vs. Amount of Training Data

Best three sentences

<table>
<thead>
<tr>
<th>Amount of training data</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>110</td>
<td>105</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>105</td>
<td>110</td>
<td>105</td>
</tr>
<tr>
<td>4</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>7</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>8</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>9</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>10</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>11</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>12</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>13</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>14</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>15</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>16</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>17</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>18</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>19</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
<tr>
<td>20</td>
<td>110</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>21</td>
<td>115</td>
<td>110</td>
<td>115</td>
</tr>
</tbody>
</table>
Sentence Rank vs. Amount of Training Data

#23 colorless green ideas sleep furiously
#36 furiously sleep ideas green colorless

Amount of training data
Sentence Rank vs. Amount of Training Data

#23 colorless green ideas sleep furiously
#36 furiously sleep ideas green colorless
Break
tregex

• The best introduction to Tregex is the brief powerpoint tutorial for Tregex by Galen Andrew.
  — The_Wonderful_World_of_Tregex.ppt
Homework Discussion

- useful command line tool:
  - `diff <file1> <file2>`

```bash
dhcp-10-142-182-95:cleft searches sandiway$ diff whclf-0 whclf
4a5,6
> wsj_0415.mrg-5   Who that winner will be *T*-1 is highly uncertain .
> wsj_0415.mrg-22  `` And where we are *T*-1 is bad .``
```

![regexGUI](image)
Homework 3

• Consider small clauses, e.g.
  – *John considers himself smart*
  – *John considers [s himself smart]*
  – *himself smart* is a "small clause" (no verb)
  – small clause has a subject: *himself*
  – small clause has a predicate: *smart*
Homework 3

• docs/prsguid1.pdf (page 252)

15 Small Clauses and their near relatives

This section is concerned with the closely related complements of certain verbs. The verbs discussed here generally have a noun phrase complement that is the logical subject of a second complement that appears to be predicated of the first complement—a construction often referred to as a "small clause". However, since we cannot hope to capture all the subtleties of the small clause-type structures described in syntactic literature, we therefore annotate alike all complement pairs that show a predicative relationship. As a result, this section covers many constructions that are not "small clauses" in the technical sense, including constructions which upon closer inspection may in fact not merit the predicative analysis we give them here.

Note also that the policy described here turns out to be a bit too complex to follow entirely consistently. Users of the corpus should therefore expect a little roughness around the edges.

15.1 Bracketing

In general, non-finite clausal complements are labeled S. The "subject" of the clause is marked -SBJ, and the "predicate" is marked -PRD (unless the predicate is a VP, which never bears the -PRD tag).

(S (NP-SBJ I)
  (VP consider
    (S (NP-SBJ Kris)
      (NP-PRD a fool))))
Homework 3

- docs/prsguid1.pdf (page 252)

If the verb is passive, the null passive object is shown as the subject of a clause:

(S (NP-SBJ-1 Kris)
 (VP is
  (VP considered
   (S (NP-SBJ *-1)
    (NP-PRD a fool))
   (PP by
    (NP-LGS most people)))))

Small clauses may be structurally distinguished from ordinary main clauses by the fact that they are immediately dominated by a VP and lack a tensed verb or modal (POS-tagged VBP, VBZ, VBD, or MD) in an S-level VP.
15.2 Small clause criteria

The main difficulty lies in distinguishing whether a particular verb has a clausal complement or two independent complements that are not so closely related (the latter exemplified by *give Calvin a comic book* or *persuade Hobbes to eat Calvin*, neither of which are given the small clause analysis). In all cases, for a pair of complements to be eligible for "small clause" bracketing, the NP complement must be the logical subject of the second complement.

In addition, it should meet certain requirements (in the form of syntactic and semantic tests which are specific to the nature of the predicate) as described in the following sections. The tests in a given section are only intended to apply to the sort of predicate that the section is concerned with. For example, the criteria for adjectival predicates do not apply to adverbial predicates, and vice-versa. The intention of this policy is to make bracketing decisions much faster and easier, by allowing the annotator to decide most cases by quick, easy rules, only resorting to Tests in a few cases. (Note that despite this intention, some annotators occasionally confused the sections and applied incorrect criteria to different predicates, especially among the non-verbal predicates. Again, users of the corpus should expect a little roughness around the edges.)
Homework 3

• Double Object Examples:
  – verbs *give* and *persuade* may take two objects
  – *give Calvin a comic book*
  – \[ [\text{VP} [\text{VB} \text{give}] \ [\text{NP} [\text{NNP Calvin}]] \ [\text{NP} [\text{DT a} \text{ comic book}]]]\]
  – *persuade Hobbes to eat Calvin*
  – \[ [\text{VP} [\text{VB persuade}] \ [\text{NP} [\text{NNP Hobbes}]] \ [\text{S} [\text{NP *}] [\text{VP to eat} \ [\text{NP} [\text{NNP Calvin}]]]]]]\]
Homework 3

• Wikipedia on small clauses and constituency testing:

a. She proved **him guilty**.

b. *Him guilty* she proved. - Small clause fails the topicalization diagnostic for identifying constituents.

c. *It is him guilty* that she proved. - Small clause fails the clefting diagnostic for identifying constituents.

d. *What she proved was him guilty*. - Small clause fails the pseudoclefting diagnostic for identifying constituents.

e. *What did she prove?* - ??**Him guilty**. - Small clause fails the answer fragment diagnostic for identifying constituents.
Homework 3

• Using tregex find all the small clauses in the WSJ corpus
• Question 1: How many are there in total?
• Question 2: Would you expect the NP-PRD type to be the most numerous? Is it?
• Question 3: List all the different types and their frequencies (in descending order)
• Show your work:
  – i.e. show the tregex search terms you used