Coreference Resolution using Web-Scale Statistics

Shane Bergsma
Johns Hopkins University

Stony Brook University
Stony Brook, NY

Fall, 2011
Pronoun Resolution

From Wikipedia:

“The first European to explore the California coast was... Juan Rodríguez Cabrillo... landing on September 28, 1542, on the shores of San Diego Bay. He claimed California for Spain.”

• Task: He → “Juan Rodríguez Cabrillo”

• Query (e.g. Google Search):
  “When did Cabrillo claim California for Spain?”
Spain began looking northward with the intention of increasing her empire. ... for a voyage up the California coast under the flag of Spain. ... Because of adverse winds Cabrillo turned back, harboring at San Miguel Island, and did not ...

www.sandiegohistory.org/bio/cabrillo/cabrillo.htm - Cached - Similar

In his days the Cabrillo expedition had no major impact. Spain did not make anything of its claims to California until the late 18th century, ...

www.win.tue.nl/~engels/discovery/cabrillo.html - Cached - Similar

Meanwhile, in 1532, Cabrillo traveled to Spain where he met Beatriz. As the Cabrillo family grew, so did his wealth and reputation as a ship builder. ... only a few hundred miles west of the coast of California. ...

www.nps.gov/cabr/juan.html - Cached - Similar

Cabrillo was then commissioned by the new Viceroy of New Spain, Antonio de Mendoza, ... The Cabrillo National Monument in San Diego, California ...

en.wikipedia.org/wiki/Juan_Rodriguez_Cabrillo - Cached - Similar

He was also instructed to discover and claim all new lands for Spain and, ... Cabrillo's ships sailed north, reaching the coast of southern California. ... sailed on to Northwest Cape beyond San Francisco Bay, which he did not find. ...

www.answers.com/topic/juan-rodriguez-cabrillo - Cached - Similar
Uses for Pronoun Resolution

• Text summarization (e.g. Google News):
  “[Juan Rodríguez Cabrillo] claimed California for Spain in 1542.”
Uses for Pronoun Resolution

• Text summarization (e.g. Google News):
  “[Juan Rodríguez Cabrillo] claimed California for Spain in 1542.”

• Machine translation (e.g. Google Translate):
  “the dog/cat doesn’t like its food anymore”

  “Der Hund/Die Katze mag sein/*sein[ihr] Futter nicht mehr”
Today’s Lecture

1. Introduction to Pronoun Resolution

2. Web-Scale Statistics for Pronoun Resolution: *Leveraging World Knowledge*
   A. Noun Gender
   B. Pattern Coreference
   C. Non-Referential Pronouns
1) Introduction to Pronoun Resolution

Coreference Resolution:

– “President Obama says he plans to discuss these issues with other leaders at the summit.”
– What do we know about these nouns? What do they refer to? Which are anaphora?
– PhD topics: NP coreference, other-anaphora, detecting anaphora, resolving pronouns, etc.
– Related topics: database de-duping, citation matching, cross-document coreference, and many others
Pronoun Resolution

• Scope: third-person anaphoric pronouns, including reflexives:
  – *He, his, him, himself* (masculine)
  – *She, her, herself* (feminine)
  – *It, its, itself* (neutral)
  – *They, their, them, themselves* (plural)
“In 2004, Exxon Mobil paid its chairman, Lee Raymond, a total of $38.1 million.”

Question: “Who is the chairman of Exxon Mobil?”

Goal:
- Resolve pronoun: “its”
- To antecedent: “Exxon Mobil”

Get: “Exxon Mobil’s Chairman, Lee Raymond”
Traditional Pronoun Resolution

“In 2004, Exxon Mobil paid its Chairman Lee Raymond a total of $38.1 million.”

Resolving a pronoun typically involves:

1. Parse text to determine noun phrases
2. Build list of preceding nouns as candidates
3. Filter candidates based on gender/number agreement, grammar violations, etc.
4. Select most likely candidate of remaining nouns based on measures of frequency, emphasis, etc.
Linguistic Knowledge

a) “John never saw the car. He arrived late.”
   “John never saw the car. It arrived late.”
Linguistic Knowledge

a) “John never saw the car. He arrived late.”
   “John never saw the car. It arrived late.”

Knowledge = Noun Gender
Linguistic Knowledge

a) “John never saw the car. He arrived late.”
   “John never saw the car. It arrived late.”

b) “John needs his friend.”
   “John needs his support.”
Linguistic Knowledge

a) “John never saw the car. He arrived late.”
   “John never saw the car. It arrived late.”

b) “John needs his friend.”
   “John needs his support.”

Knowledge =
You don’t need your own support
Linguistic Knowledge

a) “John never saw the car. He arrived late.”
   “John never saw the car. It arrived late.”

b) “John needs his friend.”
   “John needs his support.”

c) “You can make it in advance.”
   “You can make it in Hollywood.”
Linguistic Knowledge

a) “John never saw the car. He arrived late.”
   “John never saw the car. It arrived late.”

b) “John needs his friend.”
   “John needs his support.”

Knowledge = “make it” is an idiom

c) “You can make it in advance.”
   “You can make it in Hollywood.”
Linguistic Knowledge

a) “John never saw the car. **He** arrived late.”
   “John never saw the car. **It** arrived late.”

b) “John needs **his** friend.”
   “John needs **his** support.”

c) “You can make **it** in advance.”
   “You can make **it** in Hollywood.”
Knowledge via Web-Scale Counts

• Use an Internet search engine to harness all the linguistic data on the web

“Britney Spears” 64,100,000 pages
“Britany Spears” 434,000 pages
Deep Knowledge via Web-Scale Counts

• Positive or negative sentiment, connotation (Turney ‘02)
  – E.g. “unethical”
    unethical excellent         916,000 pages
    unethical poor            1,600,000 pages
A) Noun Gender Knowledge

“John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

Bergsma, Canadian AI’2005: Best Paper
Learning Noun Gender

• **Input:**
  – English noun (*John, car, Google*)

• **Output:**
  – Gender class: masculine (*he*), feminine (*she*), neutral (*it*), or plural (*they*)

• **Result:** improved pronoun resolution
If we had labeled data...

- “John never saw the car. **He** arrived late.”
  
  John.countMale++;

- “John loves his Honda.”
  
  John.countMale++;

- “Was that John? Has he lost weight?”
  
  John.countMale++;

- **John**: 100% male
Web-Mining Gender

• Count number of pages returned, e.g:
  “John * himself” “John * herself” “John * itself”
Web-Mining Gender

• Count number of pages returned, e.g:
  “John * himself” “John * herself” “John * itself”

• Other patterns:
  “John * his” “John * her” “John * its”
  “John * he” “John * she” “John * it”
  ...

25
## WordNet vs. Statistical Gender

<table>
<thead>
<tr>
<th>Noun</th>
<th>WordNet: Masculine acceptable?</th>
<th>Corpus Reflexives: P(Masculine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>OK</td>
<td>99.7%</td>
</tr>
<tr>
<td>Company</td>
<td>OK</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(93% neutral)</td>
</tr>
<tr>
<td>Computer</td>
<td>OK</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(99.2% neutral)</td>
</tr>
</tbody>
</table>
Supervised Machine Learning

- \( \mathbf{x} = (x_1, x_2, \ldots, x_n) \): Features (log-counts)
- \( \mathbf{w} = (w_1, w_2, \ldots, w_n) \): Learned weights
- Decision function:
  \[
  f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}
  \]
- Set the weights using a small number of labeled examples (SVM)
Gender Implementation

• Get counts using Google API
• Tested:
  – Gender assignment accuracy
  – within end-to-end pronoun resolution system
    (not as common as you might think)
Gender Results

• Gender classification F-Score
  – 85.4% using a corpus
  – 90.4% using the web
  – 92.2% using both
  – 88.8% for humans

• Pronoun resolution accuracy:
  – From 63.2% to 73.3% when adding learned gender information
B) Pattern Coreference

- “John needs **his** friend.”
- “John needs **his** support.”
- “John offered **his** support.”

Bergsma & Lin, COLING-ACL 2006
Dependency Path/Pattern

- Sequence of dependency links between terminal entities in a typed dependency tree

- Short form: **Noun** needs **pronoun’s** support
Learning Pattern Coreference

• Input:

\[ p = \text{“Noun needs pronoun’s support”} \]

• Output:

  – Are \textit{Noun} and \textit{pronoun} coreferent in \( p \)?
Learning Pattern Coreference

• If we had labeled data...
  – Count how often noun and pronoun co-refer in each labeled $p$

• Instead, learn from unambiguous examples:
  “We need your support”
  » not coreferent
## Pronoun Class Agreement

<table>
<thead>
<tr>
<th>Subject</th>
<th>Pronoun</th>
<th>Object</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>need</td>
<td><strong>his</strong></td>
<td>support</td>
</tr>
<tr>
<td>They</td>
<td>need</td>
<td><strong>my</strong></td>
<td>support</td>
</tr>
<tr>
<td>We</td>
<td>need</td>
<td><strong>her</strong></td>
<td>support</td>
</tr>
<tr>
<td>He</td>
<td>needs</td>
<td><strong>his</strong></td>
<td>support</td>
</tr>
<tr>
<td>I</td>
<td>need</td>
<td><strong>your</strong></td>
<td>support</td>
</tr>
</tbody>
</table>

Agreement = 20% → non-coreferent
Learned Patterns

• Example non-coreferent patterns:
  – “John married his mother.”
  – “Sue wrote her obituary.”

• Example coreferent patterns:
  – “Google says it intends to...”
  – “The revolutionaries consolidated their power.”
Using Pattern Coreference

• Use it directly as a feature in your system
• Enhancing a semantic compatibility model
• Use it to bootstrap probabilistic noun gender/number information:
  – “The newspaper says it intends to...”
  – Assume coreference, count as instance of “newspaper” being neutral
Results

• Adding path coreference:
  – From 1.3% to 1.8% improvement on three datasets (significant, p=0.05)

• Share large corpus of noun gender collected using the patterns
  – Useful for many people (e.g., used in NLP course at Stanford, 2011 CoNLL shared task)
  – Described in Jurafsky & Martin 2nd Edition
C) Non-Referential Pronouns

E.g. the word “it” in English

• “You can make it in advance.”
  – referential (50-75%)

• “You can make it in Hollywood.”
  – non-referential (25-50%)
Non-Referential Pronouns

• [Hirst, 1981]: detect non-referential pronouns, “lest precious hours be lost in bootless searches for textual referents.”

• Most existing pronoun/coreference systems just ignore the problem

• A common ambiguity:
  – “it” comprises 1% of English tokens
Non-Referential Pronouns

• Not just an English phenomenon:
  – “Wie geht es Ihnen?” (German)
  – “S’il vous plaît.” (French)

• Non-ref pronouns also in pro-drop languages:
  – “<> Es importante.” (Spanish, referential)
  – “<> Es important que ...” (Spanish, non-referential)
Non-Ref Detection as Classification

- **Input:**
  \[ s = “You can make \textit{it} in advance” \]

- **Output:**
  Is \textit{it} a non-referential pronoun in \( s \)?

Method: train a supervised classifier to make this decision on the basis of some features

A Machine Learning Approach

\[ h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} \quad \text{(predict non-ref if } h(\mathbf{x}) > 0) \]

• Typical ‘lexical’ features: binary indicators of context:
  \[ \mathbf{x} = (\text{previous-word} = \text{make}, \text{next-word} = \text{in}, \text{previous-two-words} = \text{can}+\text{make}, \ldots) \]

• Use training data to learn good values for the weights, \( \mathbf{w} \)
  – Classifier learns, e.g., to give negative weight to PPs immediately preceding ‘it’ (e.g. ... from it)
Better: Features from the Web

[Bergsma, Lin, Goebel, ACL 2008, IJCAI 2009]

• Convert sentence to a context pattern: “make ____ in advance”

• Collect counts from the web:
  – “make it/Them in advance”
    • 442 vs. 449 occurrences in Google N-gram Data
  – “make it/Them in Hollywood”
    • 3421 vs. 0 occurrences in Google N-gram Data
DANGERS

Indexed by the number of Google results for "Died in a ____ accident"

<table>
<thead>
<tr>
<th>Type of Accident</th>
<th>Google Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skydiving</td>
<td>710</td>
</tr>
<tr>
<td>Elevator</td>
<td>575</td>
</tr>
<tr>
<td>Surfing</td>
<td>496</td>
</tr>
<tr>
<td>Skateboarding</td>
<td>473</td>
</tr>
<tr>
<td>Camping</td>
<td>166</td>
</tr>
<tr>
<td>Gardening</td>
<td>100</td>
</tr>
<tr>
<td>Ice Skating</td>
<td>94</td>
</tr>
<tr>
<td>Knitting</td>
<td>7</td>
</tr>
<tr>
<td>Blogging</td>
<td>12</td>
</tr>
</tbody>
</table>
Applying the Web Counts

• How wide should the patterns span?
  – We can use all that Google N-gram Data allows:

    You can make __
    can make __ in
    make __ in advance
    __ in advance .

  – Five 5-grams, four 4-grams, three 3-grams and two bigrams

• What fillers to use? (e.g. *it, they/them, any NP*?)
Web Count Features

“it”:

log-cnt("You can make \textit{it} in") 5-grams
log-cnt("can make \textit{it} in advance")
log-cnt("make \textit{it} in advance.")
...
log-cnt("You can make \textit{it}") 4-grams
log-cnt("can make \textit{it} in")
...

“them”:

log-cnt("You can make \textit{them} in") 5-grams
...

A Machine Learning Approach Revisited

\[ h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} \quad \text{(predict non-ref if } h(\mathbf{x}) > 0) \]

- Typical features: binary indicators of context:
  \[ \mathbf{x} = (\text{previous-word} = \text{make}, \text{next-word} = \text{in}, \text{previous-two-words} = \text{can+make}, \ldots) \]

- New features: real-valued counts in web text:
  \[ \mathbf{x} = (\log\text{-cnt}(\text{"make it in advance"}), \log\text{-cnt}(\text{"make them in advance"}), \log\text{-cnt}(\text{"make * in advance"}), \ldots) \]

- Key conclusion: classifiers with web features are robust on new domains! [Bergsma, Pitler, Lin, ACL 2010]
NADA [Bergsma & Yarowsky, DAARC 2011]

• Non-Anaphoric Detection Algorithm:
  – a system for identifying non-referential pronouns

http://code.google.com/p/nada-nonref-pronoun-detector/

• Works on raw sentences; no parsing/tagging of input needed

• Classifies ‘it’ in up to 20,000 sentences/second

• It works well when used out-of-domain
  – Because it’s got those Web count features
Using web counts works great... but is it practical?

<table>
<thead>
<tr>
<th>Description</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>All N-grams in the Google N-gram corpus</td>
<td>93 GB</td>
</tr>
<tr>
<td>Extract N-grams of length-4 only</td>
<td>33 GB</td>
</tr>
<tr>
<td>Extract N-grams containing <em>it, they, them only</em></td>
<td>500 MB</td>
</tr>
<tr>
<td>Lower-case, truncate tokens to four characters, replace special tokens (e.g. named entities, pronouns, digits) with symbols, etc.</td>
<td>189 MB</td>
</tr>
<tr>
<td>Encode tokens (6 bytes) and values (2 bytes), store only changes from previous line</td>
<td>44 MB</td>
</tr>
<tr>
<td>gzip resulting file</td>
<td>33 MB</td>
</tr>
</tbody>
</table>
Accuracy with Different Features

- BBN-test
- WSJ-2
- ItBank

Majority Class
Lexical
Web-Count
NADA: Lexical+Web
NADA versus Other Systems

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paice &amp; Husk</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charniak &amp; Elsner</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NADA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Paice & Husk: Precision and Accuracy are significantly lower compared to NADA.
- Charniak & Elsner: Recall is higher than Precision and F-Score.
- NADA: Highest scores across all metrics.
Conclusion

• Web-derived knowledge is very useful for pronoun resolution:
  – Gender
  – Pattern coreference
  – Non-referential pronoun patterns
• Useful to share data and software that leverage web-scale knowledge
  – always the best way to have an impact on the field
Thanks