Low-Pass Semantics

Fernando Pereira

with

Rahul Gupta, Ni Lao, Amar Subramanya
Rest of the Google language understanding team
William Cohen, CMU
Let’s go: Mount Shasta, California
How do I climb Shasta?

... way down ...
I'd like to ski from the summit... way down...
Lessons so far

- “Things, not strings”
- Semantic index
- Join keys for composing relations
- But most information is still in text
  - Link “things” and text
Low-Pass Semantics

• Learn to *translate* and *link*

• Natural-language text

• Knowledge bases (*key spaces*)

• (Semi-)structured data

• Accept lack of finer semantic details

• *Links* ⇒ *inference* ⇒ *meaning*
Layers of meaning

- Grammatical structure (parsing)
- Referring expression classification
- Within-document coreference
- Entity resolution
  - Entity co-occurrences $\Rightarrow$ relations
Erik Wesner will talk about his research tomorrow. Though Wesner’s parents are from Poland, he was born and raised in Raleigh, NC.
## Linguistic analysis matters

### Relation extraction quality

<table>
<thead>
<tr>
<th></th>
<th>text slow parse</th>
<th>text slow parse coref</th>
<th>text fast parse</th>
<th>text fast parse coref</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ examples</td>
<td>580k</td>
<td>580k</td>
<td>935k</td>
<td>941k</td>
</tr>
<tr>
<td>patterns</td>
<td>187</td>
<td>478</td>
<td>1023</td>
<td>1013</td>
</tr>
<tr>
<td>relations</td>
<td>24</td>
<td>32</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>known triples</td>
<td>69k</td>
<td>191k</td>
<td>281k</td>
<td>270k</td>
</tr>
<tr>
<td>(likely) correct triples</td>
<td>67k</td>
<td>198k</td>
<td>478k</td>
<td>454k</td>
</tr>
<tr>
<td>parsing acc.</td>
<td>86.49</td>
<td>86.49</td>
<td>82.72</td>
<td></td>
</tr>
</tbody>
</table>
Detour: Research vs Product

- Research explores what is *possible*
- Product is about what is *effective*
- Enough impact to matter to users
- Quality does not disappoint users
- Dominates incumbent technology
- Cost-benefit tradeoff: impact/resources
- Most research results fail these tests
Case Study: Entities ⇒ relations
Distributed large-scale learning

N. Lao, A. Subramanya, F. Pereira, W. Cohen
EMNLP 2012
Joint KB+text inference

- Path types: relation chains
- Random walks on graph
- Learn type weights to favor desired walks
Path Ranking Algorithm (PRA)

Given a directed, edge-labeled graph.

Is the source $s$ connected to the target $t$ by an edge labeled $r$?

PRA (II)

Path type: $\pi = \langle r_1, r_2, \ldots, r_l \rangle$

Train by regularized logistic regression

$$P(r(s, t)) \propto \exp \sum_{\pi \in B} P(s \rightarrow t; \pi) \theta_{\pi, r}$$
Case study: extend Freebase

• Freebase: 21M concepts, 70M edges
• Study relations: profession, nationality, parent
• Profession stats:
  • 2M people in Freebase
  • 0.3M have a recorded profession
  • Biased statistics (0.24M politicians, actors)
Training data for relation \( r \)

- **Positive examples**: sample \( r \) edges with frequent targets
  \[ \frac{1}{\sqrt{\text{count}}} \]

**Negative examples**: sample pairs of concepts of types compatible with \( r \) but not in \( r \) in the knowledge base
Text Processing

**POS Tagging & Dependency Parsing**

**Noun-phrase Extraction**

**Coreference Resolution** (Haghighi & Klein 2009)

**Concept Resolution**

most likely concept for named mentions in a coreference cluster

Concept Resolution

Miles Davis

Male

Album

Kind of Blue

Mention

Mention
Text graph construction

- 60M Web pages with relevant concepts
- Dominated by frequent concepts
  - *Barack Obama* is mentioned about 8.9M times in our corpus
- Sample their sentences $\propto 1/\sqrt{\text{count}}$
Freebase and text graph

Miles Davis
Gender
Male
Album
Kind of Blue
Cheju, South Korea
South Korea

Davis
dobj

Jeju Island
nsubj

Kind of Blue

Text Graph

21M nodes 70M edges

2B nodes 5B edges
Scaling up PRA

• Path type pruning
  • Distribute the queries and explore in parallel in the Map step
  • Aggregate path types in the Reduce step
• Index text graph by entity mentions
A learned path for profession

\langle \text{Mention, conj, Mention}^{-1}, \text{Profession} \rangle
Relation extraction results

**Known triples**

<table>
<thead>
<tr>
<th>Task</th>
<th>KB</th>
<th>Text</th>
<th>KB+Text</th>
<th>KB+Text[b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>0.532</td>
<td>0.516</td>
<td>0.583</td>
<td>0.453</td>
</tr>
<tr>
<td>Nationality</td>
<td>0.734</td>
<td>0.729</td>
<td>0.812</td>
<td>0.693</td>
</tr>
<tr>
<td>Parents</td>
<td>0.329</td>
<td>0.332</td>
<td>0.392</td>
<td>0.319</td>
</tr>
</tbody>
</table>

**Human evaluation**

<table>
<thead>
<tr>
<th>Task</th>
<th>p@100</th>
<th>p@1k</th>
<th>p@10k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>0.97</td>
<td>0.92</td>
<td>0.84</td>
</tr>
<tr>
<td>Nationality</td>
<td>0.98</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td>Parents</td>
<td>0.86</td>
<td>0.81</td>
<td>0.79</td>
</tr>
</tbody>
</table>

**Coverage**

<table>
<thead>
<tr>
<th>Profession triples</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>970</td>
</tr>
<tr>
<td>10,000</td>
<td>8,726</td>
</tr>
<tr>
<td>100,000</td>
<td>79,885</td>
</tr>
</tbody>
</table>
A glimpse of meaning

- Capable language analysis components enable interpreting the Web as a huge, noisy knowledge base that combines weak sources from text and structured data to support useful inferences.