The Slot Filling Challenge

- Hand annotation performance
  - Precision: 70%
  - Recall: 54%
  - F-measure: 61%
- Top systems rarely exceed 30% F-measure

Entry level is pretty high

Jim Parsons was born and raised in Houston ... He attended Klein Oak High School in ...
The Slot Filling Challenge

<table>
<thead>
<tr>
<th>Person</th>
<th>Organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>person's name</td>
<td>org's name</td>
</tr>
<tr>
<td>person's title</td>
<td>org's title</td>
</tr>
<tr>
<td>org's name</td>
<td>org's title</td>
</tr>
<tr>
<td>personal name</td>
<td>org's name</td>
</tr>
<tr>
<td>personal title</td>
<td>org's title</td>
</tr>
<tr>
<td>person's age</td>
<td>org's age</td>
</tr>
<tr>
<td>org's age</td>
<td>org's age</td>
</tr>
<tr>
<td>person's phone</td>
<td>org's phone</td>
</tr>
<tr>
<td>org's phone</td>
<td>org's phone</td>
</tr>
<tr>
<td>person's email</td>
<td>org's email</td>
</tr>
<tr>
<td>org's email</td>
<td>org's email</td>
</tr>
</tbody>
</table>

Overview of the NYU 2011 System

- Hand crafted patterns
- Pattern Filler

Pattern Filler

- Hand crafted patterns

<table>
<thead>
<tr>
<th>Pattern Set</th>
<th>Patterns</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local patterns for person queries</td>
<td>title of org, org title, org's title, title, employee_of</td>
<td>org's employee_of, org's employee_of, org's employee_of</td>
</tr>
<tr>
<td></td>
<td>org in GPE, employee_of</td>
<td>org in GPE, employee_of, org in GPE</td>
</tr>
<tr>
<td></td>
<td>employee_of</td>
<td>employee_of, employee_of, employee_of</td>
</tr>
<tr>
<td>Local patterns for org queries</td>
<td>title of org, org title, org's title, org's title, employee_of</td>
<td>org's employee_of, org's employee_of, org's employee_of</td>
</tr>
<tr>
<td></td>
<td>org in GPE, employee_of</td>
<td>org in GPE, employee_of, org in GPE</td>
</tr>
<tr>
<td></td>
<td>employee_of</td>
<td>employee_of, employee_of, employee_of</td>
</tr>
</tbody>
</table>

Hand crafted patterns

- Pattern Filler

Pattern Filler

- Hand crafted patterns

- Pattern Filler

http://cs.nyu.edu/grishman/jet/jet.html
Learned patterns (through bootstrapping)

Basic Idea:
It starts from some seed patterns which are used to extract named entity (NE) pairs, which in turn result in more semantic patterns learned from the corpus.

Learned patterns (through bootstrapping)

Problem: semantic drift
a pair of names may be connected by patterns belonging to multiple relations
Pattern Filler

- Learned patterns (through bootstrapping)
  - Problem: semantic drift
  - Solutions:
    - Manually review top ranked patterns
    - Guide bootstrapping with pattern clusters

Distant Learning Filler

- Distant Learning (the general algorithm)
  - Map relations in knowledge bases to KBP slots
  - Search corpora for sentences that contain name pairs
  - Generate positive and negative training examples
  - Train classifiers using generated examples
  - Fill slots using trained classifiers

Problems

- Problem 1: Class labels are noisy
  - Many False Positives because name pairs are often connected by non-relational contexts

Distant Learning Filler

- Distant Learning
  - Map 4.1M Freebase relation instances to 28 slots
  - Given a pair of names \( i,j \) occurring together in a sentence in the KBP corpus, treat it as:
    - positive example if it is a Freebase relation instance
    - negative example if \( i,j \) is not a Freebase instance but \( i,j' \) is an instance for some \( j' \neq j \)
    - Train classifiers using MaxEnt
    - Fill slots using trained classifiers, in parallel with other components of NYU system

Problems

- Problem 1: Class labels are noisy
  - Many False Negatives because of incompleteness of current knowledge bases

<table>
<thead>
<tr>
<th>Attribute of Person in Freebase</th>
<th>Incompleteness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area of work</td>
<td>0.970</td>
</tr>
<tr>
<td>Area of work</td>
<td>0.970</td>
</tr>
<tr>
<td>university</td>
<td>0.780</td>
</tr>
<tr>
<td>country</td>
<td>0.560</td>
</tr>
<tr>
<td>education</td>
<td>0.580</td>
</tr>
<tr>
<td>employment history</td>
<td>0.560</td>
</tr>
</tbody>
</table>

\[
\text{Incompleteness} (\text{Attr.}) = \frac{\# \text{Person without Attr.}}{\# \text{Person}}
\]
**Distant Learning Filler**

- **Problems**
  - **Problem 2**: Class distribution is extremely unbalanced
    - Treat as negative if \( i \neq j \) is NOT a Freebase relation instance
      - Positive VS negative: 2:37
    - Treat as negative if \( i \neq j \) is NOT a Freebase instance but \( i \neq j \) is an instance for some \( j \neq j \) AND \( i \neq j \) is separated by no more than 12 tokens
      - Positive VS negative: 2:13
    - Trained classifiers will have low recall, biased towards negative

- **Solutions to Problems**
  - **Problem 1**: Class labels are noisy
    - Refine class labels to reduce noise
  - **Problem 2**: Class distribution is extremely unbalanced
    - Undersample the majority classes
  - **Problem 3**: training ignores co-reference info
    - Incorporate coreference during testing

---

**Distant Learning Filler**

- **Problems**
  - **Problem 3**: training ignores co-reference info
    - Training relies on full name match between Freebase and text
    - But partial names (Bill, Mr. Gates ...) occur often in text
    - Use co-reference during training?
      - Co-reference module itself might be inaccurate and adds noise to training
    - But can it help during testing?

---

**Distant Learning Filler--Class Label Refinement**

- The refinement algorithm
  1. Represent a training instance by its dependency pattern, the shortest path connecting the two names in the dependency tree representation of the sentence.
  2. Estimate precision of the pattern
    
    $$\text{prec}(p,c) = \frac{\text{count}(p,c)}{\sum \text{count}(p,c)}$$

    Precision of a pattern \( p \) for the class \( C \) is defined as the number of occurrences of \( p \) in the class \( C \) divided by the number of occurrences of \( p \) in any of the classes \( C \).
  3. Assign the instance the class that its dependency pattern is most precise about.

- **The refinement algorithm (cont)**
  - **Examples**

    | Example Sentence | Class       | Co-reference | Freebase | Person | Employee of | Organization | Founded by |
    |------------------|-------------|--------------|----------|--------|-------------|-------------|-----------|
    | Jon Corin        | PERSON      | appos chairman prep_of | Goldman Sachs |        |             |             |           |
    | William S. Paley | PERSON      | appos chairman prep_of | CBS       |       |             |             |           |
    |                   |             | prec(appos chairman prep_of, PERSON:Employee_of) = 0.756 | prec(appos chairman prep_of, ORG:Founded by) = 0.002 |

---

**Distant Learning Filler--Class Label Refinement**

- **Effort 1**: multiple n-way instead of single n-way classification
  - **single n-way**: an n-way classifier for all classes
    - Biased towards majority classes
  - **multiple n-way**: an n-way classifier for each pair of name types
    - A classifier for PERSON and PERSON
    - Another one for PERSON and ORGANIZATION
    - ... ...
  - On average (10 runs on 2011 evaluation data)
    - **single n-way**: 180 fills for 8 slots
    - **multiple n-way**: 240 fills for 15 slots
**Distant Learning Filler--Undersampling the Majority Classes**

- **Effort 2:**
  - Even with multiple n-way classification approach
  - OTHER (not a defined KBP slot) is still the majority class for each such n-way classifier
  - Downsize OTHER by randomly selecting a subset of them

**Distant Learning Filler--Contribution of Coreference**

- No use of co-reference during training
- Run Jet (NYU IE toolkit) to get co-referred names of the query
- Use these names when filling slots for the query
- Co-reference is beneficial to our official system
  - P/R/F of the distant filler itself
    - With co-reference: 36.4/12.4/17.4
    - Without co-reference: 28.8/10.0/14.3

**Distant Learning Filler--Experimental Results**

- Multiple n-way outperformed single n-way
- Models with refinement:
  - higher performance
  - curves are much flatter
  - less sensitive to undersampling ratio
  - more robust to noise

**Undersampling Ratio**

- MNR := Multiple n-way classifier without refinement
- MR := Multiple n-way classifier with refinement
- SR := Single n-way classifier with refinement
- SNR := Single n-way classifier without refinement

**Distant Learning Filler--Experimental Results**

- Models with refinement have better P, R
- Multiple n-way outperforms single n-way mainly through improved recall

Thanks!
Overview of 2011 System

- Baseline: 2010 System (three basic components)
  1) Document Retrieval
     - Use Lucene to retrieve a maximum of 300 documents
     - Query: the query name and some minor name variants
  2) Answer Extraction
     - Begins with text analysis: POS tagging, chunking, name tagging, time expression tagging, and coreference
     - Coreference is used to fill alternate_names slots
     - Other slots are filled using patterns (hand-coded and created semi-automatically using bootstrapping)
  3) Merging
     - Combines answers from different documents and passages, and from different answer extraction procedures

Overview of 2011 System

- Passage Retrieval (QA)
  - For each slot, a set of index terms is generated using distant supervision (using Freebase)
  - Terms are used to retrieve and rank passages for a specific slot
  - An answer is then selected based on name type and distance from the query name
  - Due to limitations of time, this procedure was only implemented for a few slots and was used as a fall-back strategy, if the other answer extraction components did not find any slot fill.

Overview of 2011 System

- Result Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYU1 (With QA)</td>
<td>25.5</td>
<td>33.6</td>
<td>29.1</td>
<td>35.4</td>
<td>51.0</td>
<td>42.1</td>
</tr>
<tr>
<td>NYU2 (Without QA)</td>
<td>25.5</td>
<td>35.0</td>
<td>29.1</td>
<td>38.1</td>
<td>57.0</td>
<td>46.4</td>
</tr>
<tr>
<td>Implicit correlation</td>
<td></td>
<td></td>
<td></td>
<td>35.4</td>
<td>51.0</td>
<td>42.1</td>
</tr>
<tr>
<td>Functional norms</td>
<td>6.5</td>
<td>28.8</td>
<td>10.0</td>
<td>25.3</td>
<td>40.5</td>
<td>30.3</td>
</tr>
<tr>
<td>Bootstrapped linear patterns</td>
<td>3.5</td>
<td>14.1</td>
<td>6.6</td>
<td>24.8</td>
<td>34.6</td>
<td>26.8</td>
</tr>
<tr>
<td>Bootstrapped dependency patterns</td>
<td>7.8</td>
<td>10.2</td>
<td>7.5</td>
<td>20.0</td>
<td>33.3</td>
<td>25.2</td>
</tr>
</tbody>
</table>

(NYU2 R/P/F 25.5/33.6/42.1)