Snowball:
Extracting Relations from Large Plain-Text Collections

Eugene Agichtein
Luis Gravano

Department of Computer Science
Columbia University
Extracting Relations from Documents

Text documents hide valuable *structured* information.

If we manage to extract this information:
- We can answer user queries more accurately
- We can run data mining tasks (e.g., finding trends)
GOAL: Extract All Tuples “Hidden” in the Document Collection

System must:

• Require minimal training for each new task
• Recover from noise
• Exploit redundancy of information in documents
Example Task: Organization/Location

Redundancy

Microsoft’s central headquarters in Redmond is home to almost every product group and division.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>Redmond</td>
</tr>
<tr>
<td>Apple Computer</td>
<td>Cupertino</td>
</tr>
<tr>
<td>Nike</td>
<td>Portland</td>
</tr>
</tbody>
</table>

Brent Barlow, 27, a software analyst and beta-tester at Apple Computer headquarters in Cupertino, was fired Monday for "thinking a little too different."

Apple’s programmers "think different" on a "campus" in Cupertino, Cal. Nike employees "just do it" at what the company refers to as its "World Campus," near Portland, Ore.
Extracting Relations from Text Collections

- Related Work
- The *Snowball* System
- Evaluation Metrics
- Experimental Results
Related Work

- **Traditional Information Extraction**
  - MUCs (Message Understanding Conferences)
    - Significant (manual) training for each new task

- **Bootstrapping**
  - Riloff et al. ('99), Collins & Singer ('99)
    - (Named-entity recognition)

*Brin (DIPRE) ('98)*
Extracting Relations from Text: DIPRE

Initial Seed Tuples:

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>MICROSOFT</td>
<td>REDMOND</td>
</tr>
<tr>
<td>IBM</td>
<td>ARMONK</td>
</tr>
<tr>
<td>BOEING</td>
<td>SEATTLE</td>
</tr>
<tr>
<td>INTEL</td>
<td>SANTA CLARA</td>
</tr>
</tbody>
</table>

Initial Seed Tuples ➔ Occurrences of Seed Tuples ➔ Generate New Seed Tuples ➔ Augment Table ➔ Generate Extraction Patterns ➔ Initial Seed Tuples
Extracting Relations from Text: DIPRE

Occurrences of seed tuples:

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>MICROSOFT</td>
<td>REDMOND</td>
</tr>
<tr>
<td>IBM</td>
<td>ARMONK</td>
</tr>
<tr>
<td>BOEING</td>
<td>SEATTLE</td>
</tr>
<tr>
<td>INTEL</td>
<td>SANTA CLARA</td>
</tr>
</tbody>
</table>

Computer servers at Microsoft’s headquarters in Redmond...

In mid-afternoon trading, share of Redmond-based Microsoft fell...

The Armonk-based IBM introduced a new line...

The combined company will operate from Boeing’s headquarters in Seattle.

Intel, Santa Clara, cut prices of its Pentium processor.

Initial Seed Tuples

Occurrences of Seed Tuples

Generate New Seed Tuples

Augment Table

Generate Extraction Patterns
Extracting Relations from Text: DIPRE

DIPRE
Patterns:

1. `<STRING1>`’s headquarters in `<STRING2>`
2. `<STRING2>` -based `<STRING1>`
3. `<STRING1>`, `<STRING2>`

**Diagram:**
- Initial Seed Tuples
- Occurrences of Seed Tuples
- Generate New Seed Tuples
- Augment Table
- Generate Extraction Patterns
Extracting Relations from Text: DIPRE

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG EDWARDS</td>
<td>ST LUIS</td>
</tr>
<tr>
<td>157TH STREET</td>
<td>MANHATTAN</td>
</tr>
<tr>
<td>7TH LEVEL</td>
<td>RICHARDSON</td>
</tr>
<tr>
<td>3COM CORP</td>
<td>SANTA CLARA</td>
</tr>
<tr>
<td>3DO</td>
<td>REDWOOD CITY</td>
</tr>
<tr>
<td>JELLIES</td>
<td>APPLE</td>
</tr>
<tr>
<td>MACWEEK</td>
<td>SAN FRANCISCO</td>
</tr>
</tbody>
</table>

Generate new seed tuples; start new iteration

Augment Table

Generate Extraction Patterns

Generate New Seed Tuples

Initial Seed Tuples

Occurrences of Seed Tuples
Extracting Relations from Text: Potential Pitfalls

• Invalid tuples generated
  – Degrade quality of tuples on subsequent iterations
  – Must have automatic way to select high quality tuples to use as new seed

• Pattern representation
  – Patterns must generalize
Extracting Relations from Text Collections

- **Related Work**
  - DIPRE

- **The Snowball System**:
  - Pattern representation and generation
  - Tuple generation
  - Automatic pattern and tuple evaluation

- Evaluation Metrics

- Experimental Results
Extracting Relations from Text: *Snowball*

Initial Seed Tuples:

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>MICROSOFT</td>
<td>REDMOND</td>
</tr>
<tr>
<td>IBM</td>
<td>ARMONK</td>
</tr>
<tr>
<td>BOEING</td>
<td>SEATTLE</td>
</tr>
<tr>
<td>INTEL</td>
<td>SANTA CLARA</td>
</tr>
</tbody>
</table>

**Flowchart:***

1. **Initial Seed Tuples**
2. Generate New Seed Tuples
3. Tag Entities
4. Generate Extraction Patterns
5. Augment Table
6. Occurrences of Seed Tuples
Extracting Relations from Text: *Snowball*

Occurrences of seed tuples:

- **Microsoft**
- **Redmond**
- **IBM**
- **Armonk**
- **Boeing**
- **Seattle**
- **Intel**
- **Santa Clara**

Initial Seed Tuples

- Computer servers at Microsoft’s headquarters in Redmond...
- In mid-afternoon trading, share of Redmond-based Microsoft fell...
- The Armonk-based IBM introduced a new line...
- The combined company will operate from Boeing’s headquarters in Seattle.
- Intel, Santa Clara, cut prices of its Pentium processor.

**Tag Entities**

**Generate New Seed Tuples**

**Generate Extraction Patterns**

**Augment Table**
**Problem: Patterns Excessively General**

Pattern: `<STRING2>`-based `<STRING1>`

Today's merger with McDonnell Douglas positions **Seattle-based Boeing** to make major money in space.

..., a producer of **apple-based jelly**, ...

Incorrect!
Extracting Relations from Text: *Snowball*

Tag Entities

Use MITRE’s Alembic Named Entity tagger

[Himanshu] + use of types

Initial Seed Tuples

Occurrences of Seed Tuples

Generate New Seed Tuples

Tag Entities

Augment Table

Generate Extraction Patterns

Computer servers at Microsoft’s headquarters in Redmond...

In mid-afternoon trading, share of Redmond-based Microsoft fell...

The Armonk-based IBM introduced a new line...

The combined company will operate from Boeing’s headquarters in Seattle.

Intel, Santa Clara, cut prices of its Pentium processor.
Extracting Relations from Text

- \(<\text{ORGANIZATION}>\)'s headquarters in \(<\text{LOCATION}>\)
- \(<\text{LOCATION}>\) -based \(<\text{ORGANIZATION}>\)
- \(<\text{ORGANIZATION}>\) , \(<\text{LOCATION}>\)

**PROBLEM:** Patterns too specific: have to match text exactly.

Initial Seed Tuples \(\rightarrow\) Occurrences of Seed Tuples

Generate New Seed Tuples \(\rightarrow\) Tag Entities

Augment Table

Generate Extraction Patterns
**Snowball: Pattern Representation**

A *Snowball* pattern vector is a 5-tuple

<left, tag1, middle, tag2, right>,

- *tag1, tag2* are named-entity tags
- *left, middle, and right* are vectors of weighed terms.
**Snowball: Pattern Generation**

Tagged Occurrences of seed tuples:

- Computer servers at **Microsoft**’s central headquarters in **Redmond**.
- In mid-afternoon trading, share of **Redmond**-based **Microsoft** fell...
- The **Armonk**-based **IBM** introduced a new line...
- The combined company will operate from **Boeing**’s headquarters in **Seattle**.
Snowball Pattern Generation: Cluster Similar Occurrences

Occurrences of seed tuples converted to Snowball representation:

{<servers 0.75> <at 0.75>}

ORGANIZATION

{<’s 0.5> <central 0.5> <headquarters 0.5> <in 0.5>}

LOCATION

{<shares 0.75> <of 0.75>}

LOCATION

{<- 0.75> <based 0.75>}

ORGANIZATION

{<fell 1>}

{<the 1>}

LOCATION

{<- 0.75> <based 0.75>}

ORGANIZATION

{<introduced 0.75> <a 0.75>}

{<operate 0.75> <from 0.75>}

ORGANIZATION

{<'s 0.7> <headquarters 0.7> <in 0.7>}

LOCATION
Similarity Metric

\[ P = \langle L_p, \text{tag}1, M_p, \text{tag}2, R_p \rangle \]

\[ S = \langle L_s, \text{tag}1, M_s, \text{tag}2, R_s \rangle \]

\[ \text{Match}(P, S) = \begin{cases} 
L_p \cdot L_s + M_p \cdot M_s + R_p \cdot R_s \\
0 
\end{cases} \]

if the tags match

otherwise

[Ankit]
- Could be better?

[Yash]
- Semantic eq of context missing?
Snowball Pattern Generation: Clustering

Cluster 1

{<servers 0.75> <at 0.75>}

ORGANIZATION

{<’s 0.5> <central 0.5> <headquarters 0.5> <in 0.5>}

LOCATION

{<operate 0.75> <from 0.75>}

ORGANIZATION

{<’s 0.7> <headquarters 0.7> <in 0.7>}

LOCATION

Cluster 2

{<shares 0.75> <of 0.75>}

LOCATION

{<– 0.75> <based 0.75>}

ORGANIZATION

{<fell 1>}

{<the 1>}

LOCATION

{<– 0.75> <based 0.75>}

ORGANIZATION

{<introduced 0.75> <a 0.75>}

Eugene Agichtein
Columbia University
Snowball: Pattern Generation

Patterns are formed as **centroids** of the clusters. Filtered by minimum number of supporting tuples.

Pattern1

| **ORGANIZATION** | {‘s 0.7} <in 0.7> <headquarters 0.7> | **LOCATION** |

Pattern2

| **LOCATION** | {← 0.75} <based 0.75> | **ORGANIZATION** |
Snowball: Tuple Extraction

Using the patterns, scan the collection to generate new seed tuples:

- Initial Seed Tuples
- Occurrences of Seed Tuples
- Generate New Seed Tuples
- Tag Entities
- Generate Extraction Patterns
- Augment Table
Represent each new text segment in the collection as the context 5-tuple:

\[
\text{Netscape} \quad \text{'s flashy headquarters in Mountain View is near}
\]

\[
\text{ORGANIZATION} \quad \{<'s 0.5>, <\text{flashy} 0.5>, <\text{headquarters} 0.5>, <\text{in} 0.5>\}
\]

\[
\text{LOCATION} \quad \{<\text{is} 0.75>, <\text{near} 0.75>\}
\]

Find most similar pattern (if any)

\[
\text{ORGANIZATION} \quad \{<'s 0.7>, <\text{headquarters} 0.7>, <\text{in} 0.7>\}
\]

\[
\text{LOCATION}
\]
**Snowball: Automatic Pattern Evaluation**

Pattern “ORGANIZATION, LOCATION” in action:

- **Positive**
  - Boeing, Seattle, said...
  - Intel, Santa Clara, cut prices...
- **Negative**
  - invest in Microsoft, New York-based analyst Jane Smith said

**Pattern Confidence:**

\[
\text{Conf}(\text{Pattern}) = \frac{\text{Positive}}{\text{Positive} + \text{Negative}}
\]

E.g., \(\text{Conf}(\text{Pattern}) = \frac{2}{3} = 66\%\)
Snowball: Automatic Tuple Evaluation

Conf(Tuple) = 1 - \( \prod (1 - \text{Conf}(P_i)) \)

- Estimation of Probability (Correct (Tuple))
- A tuple will have high confidence if generated by multiple high-confidence patterns (P_i).

Brent Barlow, 27, a software analyst and beta-tester at Apple Computer headquarters in Cupertino, was fired Monday for "thinking a little too different."

Apple's programmers "think different" on a "campus" in Cupertino, Cal.

[Barun] - Not a good metric??

<Apple Computer, Cupertino>
Snowball: Filtering Seed Tuples

Initial Seed Tuples

- ORGANIZATION
- LOCATION
- CONF

Generate new seed tuples:

- AG EDWARDS
- ST LUIS
- 0.93
- AIR CANADA
- MONTREAL
- 0.89
- 7TH LEVEL
- RICHARDSON
- 0.88
- 3COM CORP
- SANTA CLARA
- 0.8
- 3DO
- REDWOOD CITY
- 0.8
- 3M
- MINNEAPOLIS
- 0.8
- MACWORLD
- SAN FRANCISCO
- 0.7
- 157TH STREET
- MANHATTAN
- 0.52
- 15TH CENTURY EUROPE
- NAPOLEON
- 0.3
- 15TH PARTY CONGRESS
- CHINA
- 0.3
- MAD
- SMITH
- 0.3

Occurrences of Seed Tuples

Generate New Seed Tuples

Tag Entities

Augment Table

Generate Extraction Patterns
Extracting Relations from Text Collections

- Related Work
- The **Snowball** System:
  - Pattern representation and generation
  - Tuple generation
  - Automatic pattern and tuple evaluation
- Evaluation Metrics
- Experimental Results

[Kuldeep]
- Many thresholds
Task Evaluation Methodology

• Data: Large collection, extracted tables contain many tuples (> 80,000)

• Need scalable methodology:
  – *Ideal* set of tuples
  – Automatic recall/precision estimation

• Estimated precision using sampling
Collections used in Experiments

More than 300,000 real newspaper articles

<table>
<thead>
<tr>
<th>Collection</th>
<th>Source</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Wall Street Journal</td>
<td>1996</td>
</tr>
<tr>
<td></td>
<td>The Los Angeles Times</td>
<td>1996</td>
</tr>
<tr>
<td>Test</td>
<td>The New York Times</td>
<td>1995</td>
</tr>
<tr>
<td></td>
<td>The Wall Street Journal</td>
<td>1995</td>
</tr>
<tr>
<td></td>
<td>The Los Angeles Times</td>
<td>1995,’97</td>
</tr>
</tbody>
</table>
The *Ideal* Metric (1)

Creating the *Ideal* set of tuples

* A perfect, *(ideal)* system would be able to extract all these tuples.
The Ideal Metric (2)

- Precision:
  \[
  \frac{|\text{Correct (Extracted } \cap \text{ Ideal)}|}{|\text{Extracted } \cap \text{ Ideal}|}
  \]

- Recall:
  \[
  \frac{|\text{Correct (Extracted } \cap \text{ Ideal)}|}{|\text{Ideal}|}
  \]

[Danish]
+ great way to compute precision
Estimate Precision by Sampling

- Sample extracted table
  - Random samples, each 100 tuples

- Manually check validity of tuples in each sample

[Haroun] - fishy?
Extracting Relations from Text Collections

- **Related Work**

- **The *Snowball* System:**
  - Pattern representation and generation
  - Tuple generation
  - Automatic pattern and tuple validation

- **Evaluation Metrics**

- **Experimental Results**
Experimental results: Test Collection

Recall (a) and precision (a) using the *Ideal* metric, plotted against the minimal number of occurrences of test tuples in the collection.
Experimental results: Sample and Check

Recall (a) and precision (b) for varying minimum confidence threshold $T_t$.

**NOTE**: Recall is estimated using the *Ideal* metric, precision is estimated by *manually checking random samples* of result table.
Approximate Matching of Organizations

- Use *Whirl* (W. Cohen @ AT&T) to match similar organization names
- Self-join the Extracted table on the *Organization* attribute
- Join resulting table with the *Test* table, and compare values of *Location* attributes
Conclusions

We presented

• Our \textit{Snowball} system:
  – Requires minimal training (handful of seed tuples)
  – Uses a flexible pattern representation
  – Achieves high recall/precision
    > 80% of test tuples extracted

• Scalable evaluation methodology

[Haroun] + qual analysis
Critique

• Negation

[Himanshu] Algorithm does not take into account the semantic meaning of words in a given pattern, which may lead to inaccurate results. For eg. It may extract tuples that follow the pattern <organization> is not located in <headquarters>

• Orgs with multiple offices //gen principle?

[Gagan] Google’s NY office
Recent and Future Work

• Recent (presented in DMKD’00 workshop)
  – Alternative pattern representation
  – Combining representations

• Future Work
  – Evaluation on other extraction tasks
  – Extensions:
    • Non-binary relations
    • Relations with no key

[Dinesh]
- Only binary
- Seed per relation

HTML documents
Snowball: Extracting Relations from Large Plain-Text Collections

Eugene Agichtein (eugene@cs.columbia.edu)
Luis Gravano

Department of Computer Science
Columbia University
Backup Slides
Snowball Solutions

- Flexible pattern representation
- Pattern generation

Automatic pattern and tuple evaluation
- Able to recover from noise
- Keeps only high quality tuples as new seed
Experimental Results: Training

Recall (a) and precision (b) using the Ideal metric (training collection)
The tuples in the random samples were checked by hand to pinpoint the “culprits” responsible for incorrect tuples. Sample size is 100.

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Location</th>
<th>Organization</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIPRE</td>
<td>74</td>
<td>26</td>
<td>3</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>Snowball (all tuples)</td>
<td>52</td>
<td>48</td>
<td>6</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>Snowball (t = 0.8)</td>
<td>93</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Baseline</td>
<td>25</td>
<td>75</td>
<td>8</td>
<td>62</td>
<td>5</td>
</tr>
</tbody>
</table>
## Sample Discovered Patterns

<table>
<thead>
<tr>
<th>Left</th>
<th>Middle</th>
<th>Right</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;NEAR 0.01&gt;</td>
<td>&lt;IN 0.79&gt;</td>
<td>&lt;, 0.20&gt;</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>&lt;HEADQUARTERS 0.03&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt; OF 0.61&gt; &lt;, 0.61&gt;</td>
<td>&lt;, 0.15)</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>&lt; - 0.53&gt;</td>
<td>&lt;SAID 0.1&gt;</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>&lt; BASED 0.53&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;, 0.25 &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;WHILE 0.01&gt;</td>
<td>&lt;BASED 0.52&gt;</td>
<td>&lt;, 0.28&gt;</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>&lt;IN 0.52&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;, 0.43&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt; - 0.70&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;, 0.08&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FROM 0.01</td>
<td>&lt;S 0.52&gt;</td>
<td>&lt;AND 0.01&gt;</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>'&lt; 0.52&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;IN 0.24&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;HEADQUARTERS 0.22&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Convergence of *Snowball* and DIPRE

(a) Precision vs. Iterations

(b) Recall vs. Iterations

Precision (a) and Recall (b) of the DIPRE and *Snowball* with increased iterations
Approximate Matching of Organizations

- Use *Whirl* (W. Cohen @ AT&T) to match similar organization names
- Self-join the Extracted table on the *Organization* attribute
- Join resulting table with the *Test* table, and compare values of *Location* attributes

<table>
<thead>
<tr>
<th>Location</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redmond</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Armonk</td>
<td>IBM</td>
</tr>
<tr>
<td>Santa Clara</td>
<td>Intel</td>
</tr>
<tr>
<td>Mountain View</td>
<td>Netscape</td>
</tr>
<tr>
<td>Cupertino</td>
<td>Apple</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Corp.</td>
<td>Wash.</td>
</tr>
<tr>
<td>Microsoft Corporation</td>
<td>Redmond</td>
</tr>
<tr>
<td>Microsoft Corp.</td>
<td>WA</td>
</tr>
<tr>
<td>Apple Computer</td>
<td>Calif.</td>
</tr>
<tr>
<td>Apple Corp</td>
<td>Cupertino</td>
</tr>
<tr>
<td>Apple Computer Corp.</td>
<td>US</td>
</tr>
</tbody>
</table>
References


- Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. ACL’95.