Unsupervised Relation Extraction by Massive Clustering

Edgar Gonzàlez    Jordi Turmo

TALP Research Center

7 December 2009
Thousands of people were in the streets and in the basilica to pay tribute. Former president Jimmy Carter represented the United States.
Relation Detection

Thousands of people *were in the streets* and *in the basilica* to pay tribute. *Former president Jimmy Carter represented the United States.*

Entities
Thousands of people were in the streets and in the basilica to pay tribute. Former president Jimmy Carter represented the United States.

people $\leftrightarrow$ streets
Thousands of **people** were in **the streets** and in **the basilica** to pay **tribute**.

*Former president Jimmy Carter represented the United States.*

```
people ←→ basilica
```
Thousands of people were in the streets and in the basilica to pay tribute. Former president Jimmy Carter represented the United States.

Jimmy Carter $\leftrightarrow$ United States
Machine Learning for Information Extraction

- Relation Detection $\subset$ Information Extraction
  - Uses specific linguistic knowledge
  - Adaptation requires costly human effort
Machine Learning for Information Extraction

- Relation Detection ⊂ Information Extraction
  - Uses specific linguistic knowledge
  - Adaptation requires costly human effort
- Machine Learning → Adaptative IE
  - Supervised approaches
  - Weakly supervised approaches
Machine Learning for Information Extraction

- Relation Detection ⊂ Information Extraction
  - Uses specific linguistic knowledge
  - Adaptation requires costly human effort
- Machine Learning → Adaptative IE
  - Supervised approaches
  - Weakly supervised approaches
    - Supervision → Bias
Machine Learning for Information Extraction

- Relation Detection ⊂ Information Extraction
  - Uses specific linguistic knowledge
  - Adaptation requires costly human effort
- Machine Learning → Adaptative IE
  - Supervised approaches
  - Weakly supervised approaches
    - Supervision → Bias
  - Unsupervised approaches
    - Avoid biases
Machine Learning for Information Extraction

- Relation Detection ⊂ Information Extraction
  - Uses specific linguistic knowledge
  - Adaptation requires costly human effort
- Machine Learning $\rightarrow$ Adaptative IE
  - Supervised approaches
  - Weakly supervised approaches
    - Supervision $\rightarrow$ Bias
  - Unsupervised approaches
    - Avoid biases
    - Use clustering techniques
Our Proposal

- New unsupervised approach to learning for relation extraction
  - Using probabilistic clustering models
- Evaluation in ACE Relation Mention Detection task
  - Popular evaluation framework
Approach
Overview

Sentence \( x = \langle E_1, E_2 \rangle \) Instance Related / Unrelated Classifier Related / Unrelated
Overview

Approach

Scoring Filtering

\[ x = \langle E_1, E_2 \rangle \]

Instance

Score

Related / Unrelated

González & Turmo (TALP) Unsupervised Relation Extraction 7 December 2009 6 / 20
Assumptions

- Scoring:
  - Clustering $\rightarrow$ point of view
  - Cluster $\rightarrow$ shared sets of features $\rightarrow$ relatedness
  - Cluster $\rightarrow$ reliability $\rightarrow$ score
Scoring:

- Clustering → point of view
- Cluster → shared sets of features → relatedness
- Cluster → reliability → score
- $\sum$ clusterings with scored clusters ⇒ Scorer
Assumptions

- **Scoring:**
  - Clustering $\rightarrow$ point of view
  - Cluster $\rightarrow$ shared sets of features $\rightarrow$ relatedness
  - Cluster $\rightarrow$ reliability $\rightarrow$ score
  - $\sum$ clusterings with scored clusters $\Rightarrow$ Scorer

- **Filtering:**
  - Unsupervised learning $\rightarrow$ $\exists$ non-related instances
  - Highly scored instances $\rightarrow$ related pairs
Assumptions

- **Scoring:**
  - Clustering $\rightarrow$ point of view
  - Cluster $\rightarrow$ shared sets of features $\rightarrow$ relatedness
  - Cluster $\rightarrow$ reliability $\rightarrow$ score
  - $\sum$ clusterings with scored clusters $\Rightarrow$ Scorer

- **Filtering:**
  - Unsupervised learning $\rightarrow$ $\exists$ non-related instances
  - Highly scored instances $\rightarrow$ related pairs
  - Threshold value $\Rightarrow$ Filterer
Scorer Learning
Scorer Learning

- Corpus Pre-Processing
  - Tokenization, POS-Tagging, NERC
Scorer Learning

- Corpus Pre-Processing
  - Tokenization, POS-Tagging, NERC
- Instance Generation
  - $\mathcal{X} = \{x_i\}$
  - Pairs of entities co-occurring within a sentence
    - Distance threshold
  - Generation of binary features
    - Context window
    - Pattern-based $\rightarrow \text{dist}_d$, $\text{left}_d$...
    - Frequency threshold
Scorer Learning

- Instance Clustering
  - \( p(c_{pq}|x_i; \Theta_p) \)
  - Mixture of Bernoulli distributions
  - Expectation-Maximization algorithm
  - Massive repeated randomization \( \rightarrow \) Robustness
Scorer Learning

- **Instance Clustering**
  - \( p(c_{pq}|x; \Theta_p) \)
  - Mixture of Bernoulli distributions
  - Expectation-Maximization algorithm
  - Massive repeated randomization \(\rightarrow\) Robustness

- **Cluster Scoring**
  - \( z(c_{pq}) \)
  - Cluster Measures
    - Size
    - Homogeneousness \(\rightarrow\) Radius
Scorer Learning

- **Instance Clustering**
  - \( p(c_{pq}|x_i; \Theta_p) \)
  - Mixture of Bernoulli distributions
  - Expectation-Maximization algorithm
  - Massive repeated randomization → Robustness

- **Cluster Scoring**
  - \( z(c_{pq}) \)
  - Cluster Measures
    - Size (Normalized by number of non-empty clusters in clustering)
    - Homogeneousness → Radius
Scorer Learning

- **Instance Clustering**
  - $p(c_{pq}|x_i; \Theta_p)$
  - Mixture of Bernoulli distributions
  - Expectation-Maximization algorithm
  - Massive repeated randomization $\rightarrow$ Robustness

- **Cluster Scoring**
  - $z(c_{pq})$
  - Cluster Measures
    - Size (Normalized by number of non-empty clusters in clustering)
    - Homogeneousness $\rightarrow$ Radius

- **Formulae**
  - $\text{NSiz, Rad, NDns}$
Scoring

Corpus

Context

Features

Score

Models + Scores

Posterior Probabilities

\[ s(x_i) = \sum \hat{\Theta} p_k p_{\sum q=1} p(c_{pq} | x_i) \cdot z(c_{pq}) \]

González & Turmo (TALP) Unsupervised Relation Extraction 7 December 2009 11 / 20
Scoring

\[ s(x_i) = \sum_{\hat{\Theta}_p} \sum_{q=1}^{k_p} p(c_{pq} | x_i) \cdot z(c_{pq}) \]
Filterer Learning

- Determination of Threshold Score
  - $s_{th}$ such that $x_i \in R^+ \iff s(x_i) \geq s_{th}$
Filterer Learning

- Determination of Threshold Score
  - $s_{th}$ such that $x_i \in R^+ \leftrightarrow s(x_i) \geq s_{th}$
  - Heuristic-based
    1. Obtain scores of instances in training corpus
    2. Sort instances by score, obtaining a decreasing convex function
    3. Find a cut-off point
Filterer Learning

GPE-LOC - NSiz

Score

Instances

González & Turmo (TALP) Unsupervised Relation Extraction 7 December 2009 13 / 20
Approach

Filterer Learning

GPE-LOC - NSiz

Instances
Score
0.0
0.2
0.4
0.6
0.8
1.0
Evaluation
Evaluation Framework

- Corpora
  - AQUAINT (APW 2000) → 29Mw
  - ACE 2003–2008 → 500kw, 98k entities, 18k relations

- Task
  - Relation Mention Detection
  - Recall, Precision, F1

- Approaches
  - GRAMS-Ub
  - Single
  - Mass
## Average Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grams-Ub</td>
<td>-</td>
<td>43.5</td>
<td>65.6</td>
<td>51.0</td>
</tr>
<tr>
<td>Single</td>
<td>NSiz</td>
<td>52.8</td>
<td>54.3</td>
<td>52.3</td>
</tr>
<tr>
<td>Single</td>
<td>Rad</td>
<td>52.1</td>
<td>54.2</td>
<td>50.3</td>
</tr>
<tr>
<td>Single</td>
<td>NDNs</td>
<td>53.4</td>
<td>54.1</td>
<td>52.5</td>
</tr>
<tr>
<td>Mass</td>
<td>NSiz</td>
<td>59.5</td>
<td>53.7</td>
<td>55.8</td>
</tr>
<tr>
<td>Mass</td>
<td>Rad</td>
<td>62.8</td>
<td>51.7</td>
<td>56.0</td>
</tr>
<tr>
<td>Mass</td>
<td>NDNs</td>
<td>59.1</td>
<td>54.2</td>
<td>55.9</td>
</tr>
</tbody>
</table>
## Average Results

<table>
<thead>
<tr>
<th></th>
<th>Rec</th>
<th>Prc</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grams-Ub</strong></td>
<td>-</td>
<td>43.5</td>
<td>65.6</td>
</tr>
<tr>
<td><strong>Single</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSiz</td>
<td>52.8</td>
<td>54.3</td>
<td>52.3</td>
</tr>
<tr>
<td>Rad</td>
<td>52.1</td>
<td>54.2</td>
<td>50.3</td>
</tr>
<tr>
<td>NDns</td>
<td>53.4</td>
<td>54.1</td>
<td>52.5</td>
</tr>
<tr>
<td><strong>Mass</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSiz</td>
<td>59.5</td>
<td>53.7</td>
<td>55.8</td>
</tr>
<tr>
<td>Rad</td>
<td>62.8</td>
<td>51.7</td>
<td>56.0</td>
</tr>
<tr>
<td>NDns</td>
<td>59.1</td>
<td>54.2</td>
<td>55.9</td>
</tr>
</tbody>
</table>
## Average Results

<table>
<thead>
<tr>
<th></th>
<th>Rec</th>
<th>Prc</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grams-Ub</strong></td>
<td>43.5</td>
<td>65.6</td>
<td>51.0</td>
</tr>
<tr>
<td><strong>Single</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSiz</td>
<td>52.8</td>
<td>54.3</td>
<td>52.3</td>
</tr>
<tr>
<td>Rad</td>
<td>52.1</td>
<td>54.2</td>
<td>50.3</td>
</tr>
<tr>
<td>NDns</td>
<td>53.4</td>
<td>54.1</td>
<td>52.5</td>
</tr>
<tr>
<td><strong>Mass</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSiz</td>
<td>59.5</td>
<td>53.7</td>
<td>55.8</td>
</tr>
<tr>
<td>Rad</td>
<td>62.8</td>
<td>51.7</td>
<td>56.0</td>
</tr>
<tr>
<td>NDns</td>
<td>59.1</td>
<td>54.2</td>
<td>55.9</td>
</tr>
</tbody>
</table>
Filtering

**GPE-Loc - NSiz**

![Graph showing GPE-Loc - NSiz](image-url)
Filtering

GPE-Loc - NSiz

Score / F1 vs. Instances

Train, Test, F1
Filtering

GPE-LOC - NSIZ

Score / F1

Instances

González & Turmo (TALP) Unsupervised Relation Extraction 7 December 2009 17 / 20
Conclusions
Conclusions

- New unsupervised approach to learning for relation extraction
  - Using probabilistic clustering models
- Evaluation in ACE Relation Mention Detection task
  - Popular evaluation framework
Conclusions

- New unsupervised approach to learning for relation extraction
  - Using probabilistic clustering models
- Evaluation in ACE Relation Mention Detection task
  - Popular evaluation framework
  - 4-point F1 increase above state-of-the-art upper bound
  - Inclusion of richer features $\rightarrow$ Greater flexibility
  - Benefits of massive combination
  - Robustness to cluster score function
Thank you!