Content Selection

• Choosing the right information to communicate
  – Domain dependent complex task

• Content Selection Example
  – Input: Set of Attribute Value Pairs

| (name first) | John | (name last) | Doe | (weight) | 150Kg | (height) | 160cm |
| (occupation) | c-writer | (occupation) | c-producer | (award title) | BAFTA | (award year) | 1999 |
| (relative type) | c-grandson | (rel. firstN) | Dashiel | (rel. lastN) | Doe | (rel. birthD) | 1990 |

  – Output: Selected Attribute-Value Pairs

| (name first) | John | (name last) | Doe | (occupation) | c-writer | (occupation) | c-producer |

John Doe is a writer, producer, . . .

• Our focus
  – Descriptive texts (single, informative, communicative goal)
  – High-level content selection rules, to filter out the input
Our Approach: Learning of Content Selection Rules

- **Input to Our Learning System**
  - A set of associated knowledge base and text pairs

<table>
<thead>
<tr>
<th>name first</th>
<th>John</th>
<th>name last</th>
<th>Doe</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
<td>150Kg</td>
<td>height</td>
<td>160cm</td>
</tr>
</tbody>
</table>

$\Rightarrow \ldots \Rightarrow$ John Doe, American writer, born in Maryland in 1967, famous for his strong prose and ...
Our Approach: Learning of Content Selection Rules

- **Input to Our Learning System**
  - A set of text and associated knowledge base pairs

  | name first | John | name last | Doe |
  | weight     | 150Kg| height    | 160cm |

  vs.

  | name first | John | name last | Doe |
  | weight     | 150Kg| height    | 160cm |

  John Doe, American writer, born in Maryland in 1967, famous for his strong prose and...

  vs.

  | name first | John | name last | Doe |
  | weight     | 150Kg| height    | 160cm |
Our Approach: Learning of Content Selection Rules

• Input to Our Learning System
  – A set of text and associated knowledge base pairs

  | name first | John | name last | Doe |
  | weight     | 150Kg| height    | 160cm|

  ↓...↓

  John Doe, American writer, born in Maryland in 1967, famous for his strong prose and ...

• Output
  – Content Selection rules, constrained by what is in the data

• Methods
  – Analyze how variation on the data influence variations in the text
    * Compare the cross entropy of cluster of text induced by clusters on the data
Our Domain

- Generate immediate up-to-date information about individuals of interest

- PROGENIE: Automatic Biographical Description

- Columbia University—University of Colorado AQUAINT project
  - Open Question Answering
Availability of Input Material

- **PROGENIE** has three major components
  1. Knowledge Component
  2. Generation Component
  3. Learning Component

- The Knowledge Component provides structured knowledge for the Generation Component
  - Noisy input

- The Learning Component trains off-line major parts of the Generation Component
  - Using cleaner data, in the form of text and associated knowledge (Text and Knowledge Resource, TKR)
Example of the Approach

- **Given:**
  - $(KB_1, Bio_1), (KB_2, Bio_2), (KB_3, Bio_3), (KB_4, Bio_4)$

- **If:**
  - $(KB_1, KB_2)$ contain $(\langle birth place state \rangle, 'MD')$
  - $(KB_3, KB_4)$ contain $(\langle birth place state \rangle, 'NY')$

- **Then:**
  - Compare the language models of $\{Bio_1, Bio_2\}$ against $\{Bio_3, Bio_4\}$.
  - If the models differ (cross entropy), select $\langle birth place state \rangle$.

- $Bio_1 \Rightarrow \ldots born in Maryland \ldots$
- $Bio_2 \Rightarrow \ldots from Maryland \ldots$
- $Bio_3 \Rightarrow \ldots native from New York \ldots$
- $Bio_4 \Rightarrow \ldots born in New York \ldots$
Bio_k \leftrightarrow K B_k

Harris, Ed. (1950–).

Actor.

Born November 28, 1950

in Tenafly, New Jersey

Harris’ first acting role came at the age of eight when he appeared in The Third Miracle a made for television movie. After studying acting at Oklahoma University
(A) Baseline Content Selection Rules

- Obtained directly from the exact matching step
- Useful as a baseline for comparison

**Induction Algorithm**
- Count the number of times a data path appears matched in the texts
- Select the data path if above some fixed threshold

**Example**
- Always select \( \langle \text{name last} \rangle \)
- Never select \( \langle \text{height} \rangle \)
System

\[ \{ KB_1, KB_2, KB_3, KB_4 \} \]

\[ (\text{birth place state}, \text{'MD'}) \Rightarrow \{ KB_1, KB_2 \} \]

\[ (\text{birth place state}, \text{'NY'}) \Rightarrow \{ KB_3, KB_4 \} \]
System

MATCHING

CLUSTERING

Semantic inputs

Target texts

Semantic clusters

Matched texts

Counting and thresholding

Baseline rules

Class-based rules

Statistical selector

Rule-mixing logic
(B) Class-based Content Selection Rules

- Augment the baseline rules

- Select or unselect each and every instance of a given data path

- Example
  - Will add to the baseline rules like \( \text{birth place state} \)

- Impact
  - Include datapaths where no exact match between data and text can be found (e.g., “MD” \(\rightarrow\) “Maryland”).
(3) Statistical Selector Module

Find a change in word choice correlated with a change in data

SAMPLING  LANGUAGE MODELLING  CROSS ENTROPY

null hypothesis

5 random  $W_1$  $\sum_i P_{LM_1}(i) \log P_{LM_2}(i)$  2475.24

5 random

5 from cluster  $LM_2$  

5 from cluster

null hypothesis

5 random  $W_1$  $\sum_i P_{LM_1}(i) \log P_{LM_2}(i)$  3183.42

2132.22

2913.81

1451.42

..............

(20 times)

test hypothesis

5 random  $LM_1$  $\sum_i P_{LM_1}(i) \log P_{LM_2}(i)$  3670.00

2970.66

2780.76

3720.38

2429.74

..............

(20 times)
(3) Statistical Selector Module

Find a change in word choice correlated with a change in data

| null hypothesis | 2475.24  
|                | 3183.42 
|                | 2132.22  
|                | 2913.81  
|                | 1451.42  
|                | ..........  
|                | (20 times) |

| test hypothesis | 3670.00 
|                | 2970.66 
|                | 2780.76 
|                | 3720.38 
|                | 2429.74 
|                | ..........  
|                | (20 times) |

larger? Mann–Whitney U test
System
(C) Content Selection Rules

- **Rules so far**
  - Always include \(\text{birth date day}\) (baseline)
  - Always include \(\text{birth place state}\) (class-based)

- **We want constrained rules**
  - *Include the name of the award, if it is an Oscar.*

- **Example**
  - It appears ... *won an Oscar* ...
  - It does not appears ... *won an Actors Association Award* ...

- **Approach: look for \(n\)-grams in the text**
  - As a **signal** for selection
  - *won an* \(\text{award name}\)
(4) \(n\)-gram Distiller Module

**Obtaining finer grained information**

- The most significant \(n\)-grams were picked by looking at their overall contribution to the CE term

\[
CE(M_1, M_2) = - \sum_{n\text{-gram}} P_{M_1}(n\text{-gram}) \log P_{M_2}(n\text{-gram})
\]

- Re-sample and measure the impact of each \(n\)-gram to the cross-entropy formula

- Different strategies evaluated to select appropriate \(n\)-grams from the sampling
  - Top \(n\)-grams
  - Global discounting based on \(n\)-gram frequency
(5) Example Extractor Module

Extract training examples

- Training data for each data path is generated.
- Select the classification label (selected or unselected)
  - Via direct extraction from the exact match; or
  - Via the signaling $n$-grams.
- Transform the weak evidence to direct evidence

<table>
<thead>
<tr>
<th>(name first)</th>
<th>John</th>
<th>(name last)</th>
<th>Doe</th>
</tr>
</thead>
<tbody>
<tr>
<td>(weight)</td>
<td>150Kg</td>
<td>(height)</td>
<td>160cm</td>
</tr>
</tbody>
</table>

$\leftrightarrow \ldots \leftrightarrow$

John Doe, American writer, born in Maryland in 1967, famous for his strong prose and ...

$\downarrow$

$\leftrightarrow$

<table>
<thead>
<tr>
<th>(name first)</th>
<th>John</th>
<th>(name last)</th>
<th>Doe</th>
</tr>
</thead>
<tbody>
<tr>
<td>(weight)</td>
<td>150Kg</td>
<td>(height)</td>
<td>160cm</td>
</tr>
</tbody>
</table>
Experimental Setting

*Two phases of training and testing*

- **Knowledge bases from E! on-line (celebrities)**

  **Development**
  - 102 biographies
  - *From* biography.com
  - Split into development training (91) and test (11)
  - Hand-tagged the test set
  - Average length: 450 words

  **Test**
  - 205 new biographies
  - *From* imdb.com
  - Split into training (191) and test (14)
  - Hand-tagged the test set
  - Average length: 250 words

- **Content selection rules to be learned were different**
## Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th><strong>development</strong></th>
<th></th>
<th></th>
<th><strong>imdb.com</strong></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
<td>1.00</td>
<td>0.41</td>
<td>1584</td>
<td>0.23</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>baseline</td>
<td>530</td>
<td>0.40</td>
<td>0.72</td>
<td>0.51</td>
<td>727</td>
<td>0.35</td>
<td>0.68</td>
<td>0.46</td>
</tr>
<tr>
<td>class-based</td>
<td>550</td>
<td>0.41</td>
<td>0.94</td>
<td>0.58</td>
<td>891</td>
<td>0.36</td>
<td>0.88</td>
<td>0.51</td>
</tr>
<tr>
<td>content-selection</td>
<td>336</td>
<td>0.46</td>
<td>0.53</td>
<td>0.49</td>
<td>375</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>369</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>development</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
<td>1.00</td>
<td>0.41</td>
<td>1584</td>
<td>0.23</td>
<td>1.00</td>
<td>0.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>530</td>
<td>0.40</td>
<td>0.72</td>
<td>0.51</td>
<td>727</td>
<td>0.35</td>
<td>0.68</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-based</td>
<td>550</td>
<td>0.41</td>
<td>0.94</td>
<td>0.58</td>
<td>891</td>
<td>0.36</td>
<td>0.88</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>content-selection</td>
<td>336</td>
<td>0.46</td>
<td>0.53</td>
<td>0.49</td>
<td>375</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>369</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>development</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
<td></td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
<td>1.00</td>
<td>0.41</td>
<td></td>
<td>1584</td>
<td>0.23</td>
<td>1.00</td>
</tr>
<tr>
<td>baseline</td>
<td>530</td>
<td>0.40</td>
<td>0.72</td>
<td>0.51</td>
<td></td>
<td>727</td>
<td>0.35</td>
<td>0.68</td>
</tr>
<tr>
<td>class-based</td>
<td>550</td>
<td>0.41</td>
<td>0.94</td>
<td>0.58</td>
<td></td>
<td>891</td>
<td>0.36</td>
<td>0.88</td>
</tr>
<tr>
<td>content-selection</td>
<td>336</td>
<td>0.46</td>
<td>0.53</td>
<td>0.49</td>
<td></td>
<td>375</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td>369</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

| Experiment      | imdb.com    |             |         |             |       |         |         |       |
|                 | Selected    | Prec.       | Rec.    | F*          |       | Selected| Prec.   | Rec.  | F* |
|                 |             |             |         |             |       |         |         |       |

The table above shows the results for different experiments on the development and imdb.com datasets.
## Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>development</th>
<th></th>
<th></th>
<th></th>
<th>imdb.com</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
<td>Selected</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F*</td>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
<td>1.00</td>
<td>0.41</td>
<td>1584</td>
<td>0.23</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>baseline</td>
<td>530</td>
<td>0.40</td>
<td>0.72</td>
<td>0.51</td>
<td>727</td>
<td>0.35</td>
<td>0.68</td>
<td>0.46</td>
</tr>
<tr>
<td>class-based</td>
<td>550</td>
<td>0.41</td>
<td>0.94</td>
<td>0.58</td>
<td>891</td>
<td>0.36</td>
<td>0.88</td>
<td>0.51</td>
</tr>
<tr>
<td>content-selection</td>
<td>336</td>
<td>0.46</td>
<td>0.53</td>
<td>0.49</td>
<td>375</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>369</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Conclusions

- **We filter out half the input data**
  - Keeping a 90%+ recall

- **Class-based model is best**
  - Aid in the Content Selection Knowledge Engineering task.
  - Ripper approach requires a better instance representation

- **Novel method for learning Content Selection rules**
  - Content Selection is a difficult, domain dependent, task

- **Further work**
  - Incorporate knowledge (improve clustering and matching)
  - Improve \( n \)-gram distillation and rule-induction instance representation