Indirect Supervised Learning
Of Content Selection Rules

Pablo Duboue
My Background

- **High-school**
  - Latin, Greek, French, Italian, Art History, etc.
  - Math Olympiads.

- **Undergrad**
  - Computer Science at Math School.
  - Thesis: LFG Parser in Haskell.

- **Grad-school**
  - Joined Columbia in 1999.
  - First year: WSD for bioinformatics.
Talk Structure

- **The Problem**
  - High Level Perspective
  - Learning Content Selection Rules
  - Text-Knowledge Corpus

- **My Solution**

- **Experiments**

- **Conclusions**
PROGENIE: A Biographical Generator

- **PROGENIE: Automatic Biographical Descriptions**

- **Generate immediate up-to-date biographical profiles**
  - Different, Learned Content Plans
    - Different, Possible Users

- **Columbia University—University of Colorado AQUAINT**
  - Open Question Answering
  - Funded by ARDA
Knowledge Component

Generation Component
- Strategic Module

Learning Component
- Schema CS rules

Generated Biographies

Knowledge Sources

Internet

Aligned Text–Data Corpus
- KB
- text

Knowledge Component

Knowledge
PROGENIE: Knowledge Component

- **Knowledge Bases for Training**
  - Knowledge as clean as possible.

- **Knowledge Bases for Execution**
  - GATE, an Information Extraction pipeline
  - University of Colorado Semantic Parser
  - Publicly Available Knowledge as a Test Bed

- Both Knowledge Bases share the same **Ontology**
PROGENIE: Generation Component

1. **Strategic Module** Content Selection rules and Document Structuring schemas.

2. **Text Planner** Splits a rhetorical tree into paragraphs.

3. **Referring Expression Generator** Pronominalization.

4. **Aggregation** Mixes together clauses with similar structure.

5. **Lexical Chooser** Selects words for concepts.

How Complex Is Strategic Generation?
Strategic Generation

• **Content Selection**
  – Choosing the right information to communicate.
  – Arguably the most critical part from the user’s perspective.

• **Document Structuring**
  – Conciseness and coherence goals.
  – Information in context.

• **Domain Dependent Complex Tasks**
Content Selection Example

- **Input: Set of Attribute Value Pairs**
  
<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name first</td>
<td>John</td>
</tr>
<tr>
<td>weight</td>
<td>150Kg</td>
</tr>
<tr>
<td>occupation</td>
<td>c-writer</td>
</tr>
<tr>
<td>award title</td>
<td>BAFTA</td>
</tr>
<tr>
<td>relative type</td>
<td>c-grandson</td>
</tr>
<tr>
<td>rel. lastN</td>
<td>Doe</td>
</tr>
<tr>
<td>rel. firstN</td>
<td>Dashiel</td>
</tr>
<tr>
<td>rel. birthD</td>
<td>1990</td>
</tr>
</tbody>
</table>

- **Output: Selected Attribute-Value Pairs**
  
<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name first</td>
<td>John</td>
</tr>
<tr>
<td>name last</td>
<td>Doe</td>
</tr>
<tr>
<td>occupation</td>
<td>c-writer</td>
</tr>
<tr>
<td>occupation</td>
<td>c-producer</td>
</tr>
</tbody>
</table>

- **Example Verbalization**

  *John Doe is a writer, producer, …*
Learning Problem

- **Input To My Learning System**
  - A set of text and associated knowledge base pairs

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

← ... → 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

- **Domain Limitations**
  - Descriptive Text.
  - Rich in Anchors.
Actor, born Thomas Connery on August 25, 1930, in Fountainbridge, Edinburgh, Scotland, the son of a truck driver and charwoman. He has a brother, Neil, born in 1938. Connery dropped out of school at age fifteen to join the British Navy. Connery is best known for his portrayal of the suave, sophisticated British spy, James Bond, in the 1960s.
THE FACTS
Sean Connery

**get the goods**
- search for Sean Connery
- products:
  - movies
  - collectibles

**Birth Name:** Thomas Sean Connery
**Birthdate:** August 25, 1930
**Birthplace:** Edinburgh, Scotland
**Occupations:** Actor, Director, Model, Producer
**Quote:** "I would drink Sean Connery’s bath water." —Whoopi Goldberg, Cable Magazine, 1989

“He’s...one of the best actors there is, simple as that... With Sean, in addition to brilliant talent, there is a persona that every great star has. When Sean’s...on the screen, it’s hard to look at anything else. To be a great star, you have to be a first-rate actor, too, or that list of great actors... Sean makes you wish..."
Input Availability

- **Biology**
  - Biological KB and Species Descriptions.

- **Geography**
  - CIA Factbook and Country Descriptions.

- **Financial Market**
  - Stock Data and Market Reports.

- **Entertainment**
  - Role Playing Character Sheets and Character Descriptions.
Input: Aligned Text-Knowledge Corpus

- **Celebrities**
  - Easily available
  - Representative of the learning issues
  - Possibility of corpus re-distribution

- **Size**
  - Knowledge frames for 1,100 different celebrities
  - Assorted biographies, ranging from 110 to 500 bios
  - Knowledge and biographies crawled from independent Websites
Output: Content Selection Rules

All rules take a node in the knowledge representation and return T or F.

**TRUE ()** Always select.

Example: for node ∈ name→last, select node.

**IN (list of values)** Select if the value is in the list.

Example: for node ∈ education→place→country, if node_value ∈ {“Scotland”, “England”}, then select node.

**TRAVERSE (path,recursive-rule)** Select if the node at the end of the path matches the recursive-rule.

Example: for node ∈ relative→relative→name→first, if node←name←relative→type ∈ {son, daughter}, then select node.

**AND, OR** Plus logic combinators.
Talk Structure

- The Problem

- My Solution
  - Indirect Supervised Learning
  - Technique Overview
  - Example
  - Details

- Experiments

- Conclusions
Indirect Supervised Learning: Overview
Indirect Supervised Learning: Overview

Knowledge

Relevant Knowledge

Content Selection

Text
Indirect Supervised Learning: Overview

Knowledge

Relevant Knowledge

Content Selection

Text
Indirect Supervised Learning

- **Learning Without Hand-labelling**
  - Employing evidence used by humans to learn

| 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 |

- **Learning As Automated Knowledge Acquisition**
  - Learning Structures That Humans Can Understand.
  - Mixing Machine Learning And Knowledge-based Approaches.
  - Domain-independence Through Learning.

- **My focus**
  - Descriptive Texts (Single, Informative, Communicative Goal).
  - High-level Content Selection Rules, To Filter Out The Input.
Example of the Approach

• Given:
  – \((KB_1, Bio_1), (KB_2, Bio_2), (KB_3, Bio_3), (KB_4, Bio_4)\)

• Cluster Knowledge Bases By Value:
  – \(\{KB_1, KB_2\}\) contain \(\langle birth \rightarrow place \rightarrow state, 'MD' \rangle\)
  – \(\{KB_3, KB_4\}\) contain \(\langle birth \rightarrow place \rightarrow state, 'NY' \rangle\)

• Compare Language Models Of Clusters:
  – Compare the models of \(\{Bio_1, Bio_2\}\) against \(\{Bio_3, Bio_4\}\).
  – If the models differ, select \(\langle birth \rightarrow place \rightarrow state \rangle\).

• \(Bio_1 \Rightarrow \ldots \text{born in Maryland}\ldots\) 
• \(Bio_2 \Rightarrow \ldots \text{from Maryland}\ldots\) 
• \(Bio_3 \Rightarrow \ldots \text{native of New York}\ldots\) 
• \(Bio_4 \Rightarrow \ldots \text{born in New York}\ldots\)
Methods: Indirect Supervised Learning

- Semantic inputs
- Target texts
- Content selection rules

Dataset Construction

Supervised Learning
Methods: Dataset Construction

DATASET CONSTRUCTION

"EXACT"

"STATISTICAL"

content selection dataset

semantic inputs

target texts
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 . . .

sel ⟨name last⟩ “Harris”
¬sel ⟨name first⟩ “Edward”
sel ⟨occupation⟩ c-actor
¬sel ⟨birth date year⟩ 1950
¬sel ⟨birth date month⟩ 11
sel ⟨birth date day⟩ 28
sel ⟨birth place city⟩ “Tenafly”
¬sel ⟨birth place province⟩ “NJ” . . .
Dataset Construction: Statistical Pipeline

"STATISTICAL" PIPELINE

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

\[
\downarrow
\]

\[
((\text{birth place state }), 'MD') \Rightarrow \{KB_1, KB_2\} \Rightarrow \{Bio_1, Bio_2\}
\]

\[
((\text{birth place state }), 'NY') \Rightarrow \{KB_3, KB_4\} \Rightarrow \{Bio_3, Bio_4\}
\]
Dataset Construction: Statistical Pipeline

"STATISTICAL" PIPELINE

- From the cluster.
- From outside the cluster.

**Sample word counts**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>York</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>The</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outside Cluster</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>York</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>The</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Dataset Construction: Statistical Pipeline

"STATISTICAL" PIPELINE

- **Sample word counts**
  - From the cluster.
  - From outside the cluster.

- **Use Student’s t-test**
  - Look for words counts that show a statistically significant difference on the counts.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Word</th>
<th>Count</th>
<th>Outside Cluster</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>6</td>
<td>New</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>York</td>
<td>5</td>
<td>York</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Dataset Construction: Statistical Pipeline

**"STATISTICAL" PIPELINE**

- **Sample word counts**
  - From the cluster.
  - From outside the cluster.

- **Use Student's t-test**
  - Look for words counts that show a statistically significant difference on the counts.

- **Words found?**
  - The information is included in the text.
  - The words are signals of that inclusion.
Methods: Supervised Learning

\[
\begin{align*}
    sel & \langle \text{name last} \rangle \text{“Harris”} \\
    \neg sel & \langle \text{name first} \rangle \text{“Edward”} \\
    sel & \langle \text{birth date year} \rangle 1950 \\
    sel & \langle \text{occupation} \rangle \text{c-actor} \\
    sel & \langle \text{birth date month} \rangle 11 \\
    sel & \langle \text{birth date day} \rangle 28 \\
    \neg sel & \langle \text{birth place province} \rangle \text{“NJ”} \ldots
\end{align*}
\]
Supervised Learning: Genetic Algorithms

- **Genetic Algorithms (GAs)**
  - An Empirical Risk Minimization Method
  - A good optimization technique
    - To explore huge search spaces with highly interrelated features.
  - Biological Metaphor
  - I use them as Symbolic Learners.

- **GAs are driven by a **Fitness Function** that tells good solutions from bad.**
Genetic Algorithms: Description

- **How GAs Work**
  - In a genetic search, at all times a population of possible solutions, called chromosomes is kept.
  - Each chromosome has an associated fitness value, indicating its apparent goodness.
  - In each step of the search, or generation, a percentage of the worst-fitted chromosomes is discarded.
  - The empty slots are filled by applying operators, that create new chromosomes by mixing two existing ones (combination) or by making changes in a existing one (mutation).
I use the weighted F-measure over the labels as fitness:

\[
Fitness = F^*_\alpha + \text{MDL}
\]

where

\[
F^*_\alpha = \frac{(\alpha^2 + 1) \text{PrecRec}}{\alpha^2 \text{Prec} + \text{Rec}}
\]

\[
\text{MDL} = \text{a minimum description length term}
\]

This function captures the problem well and allows selecting solutions that prefer precision or recall through the \(\alpha\) parameter.
Talk Structure

- The Problem
- My Solution
- **Experiments**
  - Data
  - Dataset evaluation
  - Rules evaluation
- Conclusions
Experimental Setting

Two phases of training and testing

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

  Corpus 1
  - 102 biographies
  - From biography.com
  - Split into development training (91) and test (11)
  - Hand-tagged the test set
  - Average length: 450 words

  Corpus 2
  - 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
Evaluation Of Extracted Dataset

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Exact Match</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec.</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>Rec.</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>$F^*$</td>
<td>0.69</td>
<td>0.71</td>
</tr>
</tbody>
</table>

- **Testing Overall Indirect Supervised Algorithm:**
  - Obtain rules over $Train$.  
  - Test rules over $Test$.  

- **Testing The Unsupervised Part:**
  - Obtain labels over $Train + Test$.  
  - Compare with the Test labels over $Test$ with the ones obtained by hand.
## Evaluation Of Content Selection Rules

<table>
<thead>
<tr>
<th>Experiment</th>
<th>biography.com</th>
<th>imdb.com</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>Prec.</td>
</tr>
<tr>
<td>random</td>
<td>162</td>
<td>0.29</td>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
</tr>
<tr>
<td>true/false rules</td>
<td>550</td>
<td>0.41</td>
</tr>
<tr>
<td>only exact match</td>
<td>359</td>
<td>0.64</td>
</tr>
<tr>
<td>combined</td>
<td>292</td>
<td>0.57</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>-</td>
</tr>
</tbody>
</table>
Evaluation Of Content Selection Rules

<table>
<thead>
<tr>
<th>Experiment</th>
<th>biography.com</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>random</td>
<td>162</td>
<td>0.29</td>
<td>0.48</td>
<td>0.36</td>
<td>369</td>
<td>0.25</td>
<td>0.50</td>
<td>0.33</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>true/false rules</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>only exact match</td>
<td>359</td>
<td>0.64</td>
<td>0.61</td>
<td>0.62</td>
<td>432</td>
<td>0.48</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>combined</td>
<td>292</td>
<td>0.57</td>
<td>0.81</td>
<td>0.67</td>
<td>432</td>
<td>0.49</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>369</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

My new rules select comparable amount of data.
## Evaluation Of Content Selection Rules

<table>
<thead>
<tr>
<th>Experiment</th>
<th>biography.com</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>random</td>
<td>162</td>
<td>0.29</td>
<td>0.48</td>
<td>0.36</td>
<td>369</td>
<td>0.25</td>
<td>0.50</td>
<td>0.33</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>true/false rules</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>only exact match</td>
<td>359</td>
<td>0.64</td>
<td>0.61</td>
<td>0.62</td>
<td>432</td>
<td>0.48</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>combined</td>
<td>292</td>
<td>0.57</td>
<td>0.81</td>
<td>0.67</td>
<td>432</td>
<td>0.49</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>369</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

They increase precision, although the statistical pipeline has its toll.
The statistical pipeline increases recall.
# Evaluation Of Content Selection Rules

<table>
<thead>
<tr>
<th>Experiment</th>
<th>biography.com</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>random</td>
<td>162</td>
<td>0.29</td>
<td>0.48</td>
<td>0.36</td>
<td>369</td>
<td>0.25</td>
<td>0.50</td>
<td>0.33</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>true/false rules</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>only exact match</td>
<td>359</td>
<td>0.64</td>
<td>0.61</td>
<td>0.62</td>
<td>432</td>
<td>0.48</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>combined</td>
<td>292</td>
<td>0.57</td>
<td>0.81</td>
<td>0.67</td>
<td>432</td>
<td>0.49</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td></td>
<td></td>
<td></td>
<td>369</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The combined rules are the best overall.
Talk Structure

- The Problem
- My Solution
- Experiments
- **Conclusions**
  - Current Work
  - Conclusions
Current Work

- **Join The Two Pipelines**
  - The Statistical Pipeline now provides new verbalizations for the Search-in-Text approach.
  - Execute the Statistical Pipeline when no new verbalizations are found in the text.

- **Disambiguation**
  - Use the context of a found match to decide whether is a real or a spurious match.
  - Naïve Bayes.
Conclusions

- **Content Selection**
  - Complex Task Common to NLG and Template-based Systems.
  - Requires Customization When Moving to New Domains.

- **My Solution**
  - Improved Rule Language.
  - Improved Supervised Learning Step, with novel Fitness Function based on Training Material.

- **Indirect Supervised Learning**
  - Paired Unsupervised With Supervised Learning To Achieve Supervised Learning Without Hand-tagging.
  - May Be Applicable In Other Areas Of NLP
MDL term

$$\beta = 1.5 \log \left( \frac{s}{1-s} \right)$$

$$MDL = -\frac{1}{1 + e^{-\beta l}}$$

$t$: total number of items to be selected in the current data class.
$s$: user-provided saturation parameter (0.99).
$l$: length of the rule being evaluated measure in predicates.