Indirect Supervised Learning
Of Content Selection Rules

Pablo Duboue

Computer Science Department
Columbia University
in the city of New York
Talk Structure

- High Level Perspective
  - Biographical Descriptions
  - Content Planning
  - Content Selection

- The Problem
- My Solution
- Experiments
- Conclusions
PROGENIE: A Biographical Generator

- **PROGENIE: Automatic Biographical Descriptions**

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

- Columbia University—University of Colorado AQUAINT
  - Open Question Answering
  - Funded by ARDA
Content Planning

- **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**

  - (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

- **Example Verbalization**

  *John Doe is a writer, producer, ...*
Indirect Supervised Learning

Knowledge

Pre-selected Data

Sequence of Messages

Text

Content Selection

Text Structuring

Aggregation & Lexicalization
Indirect Supervised Learning

Knowledge

Pre-selected Data

Content Selection

Text
Indirect Supervised Learning

Knowledge

Pre-selected Data

Content Selection

Text
Example Learned Rules

- **name** $\rightarrow$ **first** and **name** $\rightarrow$ **last**
  Rule: $\text{TRUE}()$
  Always say first and last names.

- **education** $\rightarrow$ **place** $\rightarrow$ **country**
  Rule: $\text{IN}("Scotland","England")$
  As I used U.S. biographies, the country of education is only mentioned when it is abroad.

- **significant-other** $\rightarrow$ **#TYPE**
  Rule: $\text{IN}("c-husband","c-wife")$
  Mention husband and wives (but not necessarily boyfriends, girlfriends or lovers).
Talk Structure

- High Level Perspective
- **The Problem**
  - Learning Content Selection Rules
  - Text-Knowledge Corpus
- My Solution
- Experiments
- Conclusions
• **Input To My Learning System**
  – A set of text and associated knowledge base pairs

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

\[\rightarrow \ldots \rightarrow\]

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.
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

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, 1999

"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... an that list of great actors, Sean ranks way high."
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
  - Knowledge and biographies crawled from independent websites
Output: Content Selection Rules

All rules take a node in the knowledge representation and return true or false.

**TRUE ()** Always select.

**IN (1994,1995)** Select if the value is in the list.

**TRAVERSE (../../relative/#TYPE, IN (c-cousin))** Select if this is the name of a cousin.

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

- High Level Perspective
- The Problem
- **My Solution**
  - Indirect Supervised Learning
  - Technique Overview
  - Example
  - Details
- Experiments
- Conclusions
Indirect Supervised Learning: Overview

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

  \[
  \begin{array}{c|c|c}
  \text{name first} & \text{name last} & \text{example} \\
  \text{John} & \text{Doe} & \text{John Doe, American writer, born in Maryland in 1967, famous for his strong prose and ...} \\
  \text{150Kg} & \text{160cm} & \\
  \end{array}
  \]

  vs.

  \[
  \begin{array}{c|c|c}
  \text{name first} & \text{name last} & \text{example} \\
  \text{John} & \text{Doe} & \text{John Doe, American writer, born in Maryland in 1967, famous for his strong prose and ...} \\
  \text{150Kg} & \text{160cm} & \\
  \end{array}
  \]

- **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:**
  - $(K B_1, Bio_1), (K B_2, Bio_2), (K B_3, Bio_3), (K B_4, Bio_4)$

- **If:**
  - $\{K B_1, K B_2\}$ contain $(\langle birth \rightarrow place \rightarrow state\rangle, 'MD')$
  - $\{K B_3, K B_4\}$ contain $(\langle birth \rightarrow place \rightarrow state\rangle, 'NY')$

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

- $Bio_1 \Rightarrow "\ldots born in Maryland\ldots"$
- $Bio_2 \Rightarrow "\ldots from Maryland\ldots"$
- $Bio_3 \Rightarrow "\ldots native of New York\ldots"$
- $Bio_4 \Rightarrow "\ldots born in New York\ldots"$
Methods: Indirect Supervised Learning

- Input: Texts
- Semantic Inputs
- Target Texts
- Content Selection
- Content Selection Rules
- Dataset Construction
- Supervised Learning
Methods: Dataset Construction

**Semantic Inputs**: Make selection from input texts

**Target Texts**: Selection of content dataset

**Dataset Construction**: Combination of "Exact" and "Statistical" datasets

**Dataset Combination**: Final selection of content dataset
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 . . .

\(\text{name last}\) “Harris”
\(\text{name first}\) “Edward”
\(\text{birth date year}\) 1950
\(\text{occupation}\) c-actor
\(\text{birth date month}\) 11
\(\text{birth date day}\) 28
\(\text{birth place city}\) “Tenafly”
\(\text{birth place province}\) “NJ” . . .
Dataset Construction: Statistical Pipeline

"STATISTICAL" PIPELINE

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

\[
\downarrow
\]

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

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

Genetic Search

- content selection
- rules

instance pool
- ruleset
- ruleset
- ruleset

fitness fn

content selection dataset

mutations crossover
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.**
I use the weighted F-measure from IR 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

• High Level Perspective
• 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><strong>0.71</strong></td>
</tr>
</tbody>
</table>

- **Testing Overall Indirect Supervised Algorithm:**
  - Obtain rules over $Train$
  - Hand tag $Test$
  - 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>EMNLP’03</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>imdb.com</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></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>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
<td>1.00</td>
<td>0.41</td>
</tr>
<tr>
<td>EMNLP’03</td>
<td>550</td>
<td>0.41</td>
<td>0.94</td>
<td>0.58</td>
</tr>
<tr>
<td>only exact match</td>
<td>359</td>
<td>0.64</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>combined</td>
<td>292</td>
<td>0.57</td>
<td>0.81</td>
<td>0.67</td>
</tr>
<tr>
<td>test set</td>
<td>293</td>
<td>-</td>
<td>-</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>EMNLP’03</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>
<tr>
<td>Experiment</td>
<td>biography.com</td>
<td></td>
<td></td>
<td></td>
<td>imdb.com</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------</td>
<td>------------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>------------</td>
<td>---------</td>
<td>----------</td>
</tr>
<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><strong>0.48</strong></td>
<td>0.36</td>
<td>369</td>
<td>0.25</td>
<td><strong>0.50</strong></td>
<td>0.33</td>
</tr>
<tr>
<td>select-all</td>
<td>1129</td>
<td>0.26</td>
<td><strong>1.00</strong></td>
<td>0.41</td>
<td>1584</td>
<td>0.23</td>
<td><strong>1.00</strong></td>
<td>0.37</td>
</tr>
<tr>
<td>EMNLP’03</td>
<td>550</td>
<td>0.41</td>
<td><strong>0.94</strong></td>
<td>0.58</td>
<td>891</td>
<td>0.36</td>
<td><strong>0.88</strong></td>
<td>0.51</td>
</tr>
<tr>
<td>only exact match</td>
<td>359</td>
<td>0.64</td>
<td><strong>0.61</strong></td>
<td>0.62</td>
<td>432</td>
<td>0.48</td>
<td><strong>0.65</strong></td>
<td>0.55</td>
</tr>
<tr>
<td>combined</td>
<td>292</td>
<td>0.57</td>
<td><strong>0.81</strong></td>
<td>0.67</td>
<td>432</td>
<td>0.49</td>
<td><strong>0.68</strong></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>
### 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>EMNLP’03</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>
Talk Structure

- High Level Perspective
- 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**
  - Use Machine Learning to Achieve Domain Independence.

- **Indirect Supervised Learning**
  - Machine Learning Without Hand-tagging
  - Applicable In A Number Of Domains
  - May Be Applicable In Other Areas Of NLG
    * Sentence Planning.
    * Surface Realization.