Indirect Supervised Learning

Of Content Selection Rules

Pablo Duboue

Computer Science Department
Columbia University
in the city of New York

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.

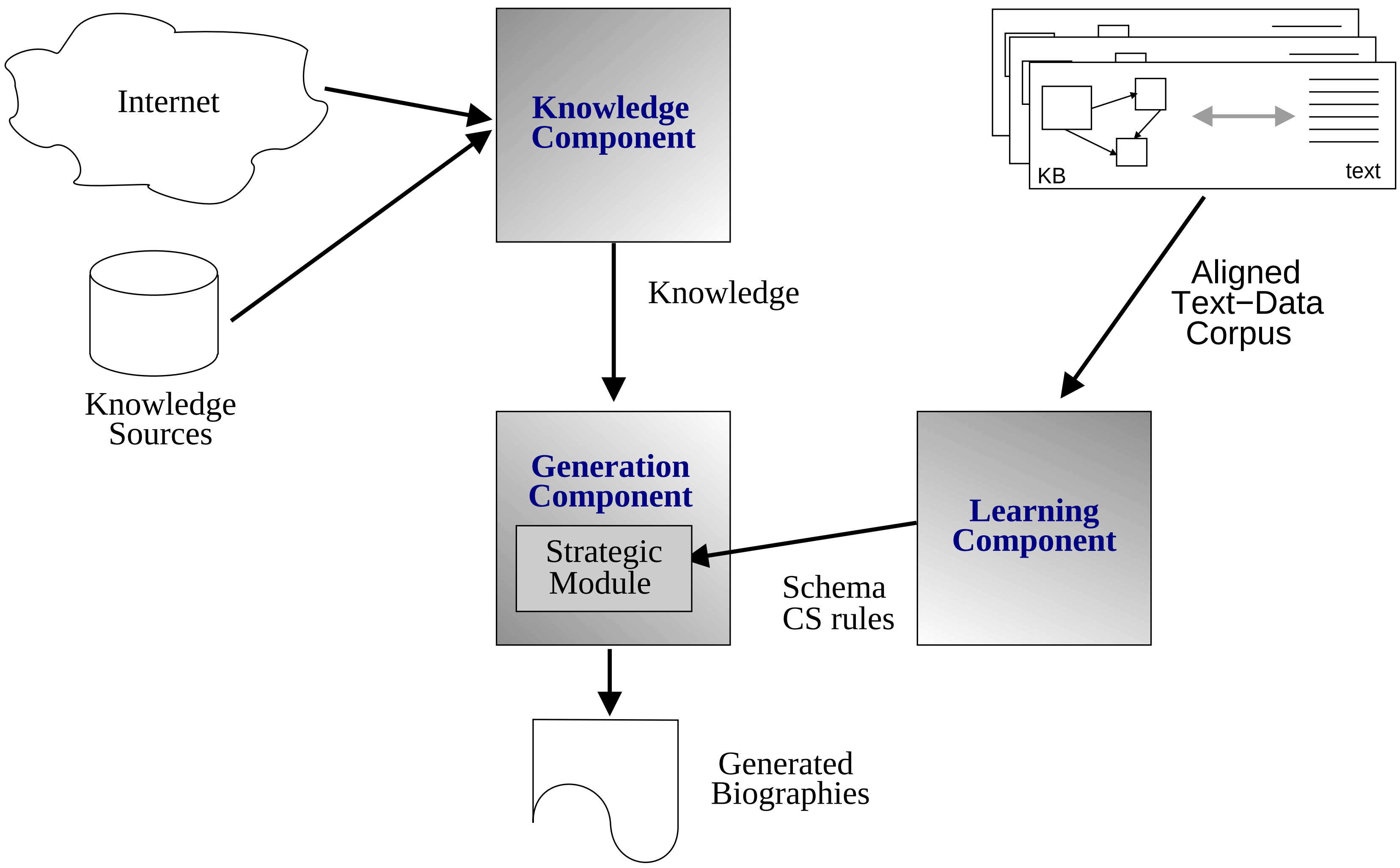
Talk Structure

- **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

PROGENIE


Internet
Knowledge Sources
Knowledge Component
Knowledge
KB
text
Aligned Text−Data Corpus
Generation Component
Strategic Module
Schema CS rules
Learning Component
Generated Biographies

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.
6. **Surface Realizer** Recursive-descent realizer.

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

⟨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, . . .

Learning Problem

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

⟨name first⟩	John	⟨name last⟩	Doe
⟨weight⟩	150Kg	⟨height⟩	160cm

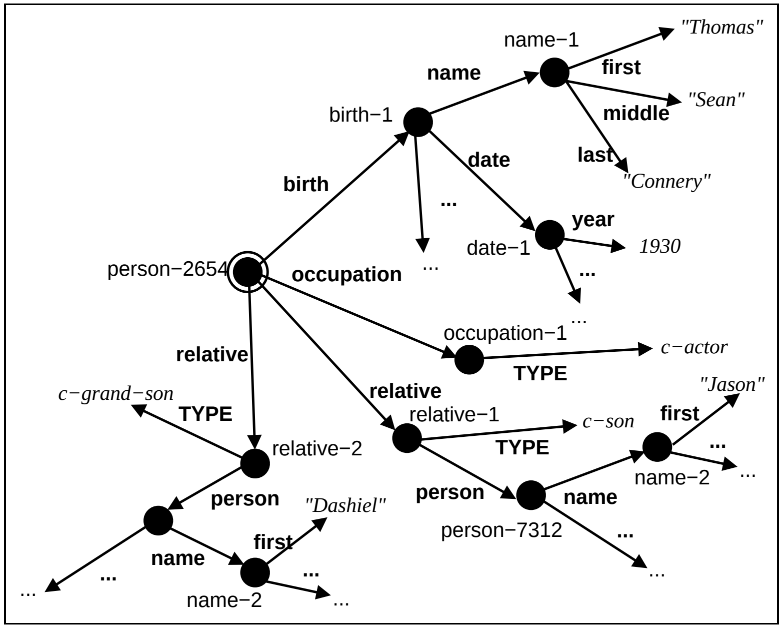
$\leftarrow \ldots \rightarrow$

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.
  - Rich in Anchors.

Input Example

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. . . .



Factsheets


online
It's a Must Have This Season
style.
Get Our Free Newsletter
search
go
HOME
NEWS
FEATURES
GOSSIP
REVIEWS+
CELEBS
FUN&GAMES
MULTIMEDIA
E!TV
TODAY S NEWS
BE ON TV
FRESH FEATURES
tonight on
THE FACTS
Sean Connery
the facts
credits
stories
multimedia
fanclubs
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
It's a Must Have This Season
style.

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 Web-sites

Output: Content Selection Rules

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

`TRUE()` Always select.
Example: for node $\in$ **name**$\rightarrow$**last**, **select node**.

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

Example: for node $\in$ **education**$\rightarrow$**place**$\rightarrow$**country**, if node_value $\in$ {*"Scotland"*, *"England"*}, then **select node**.

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

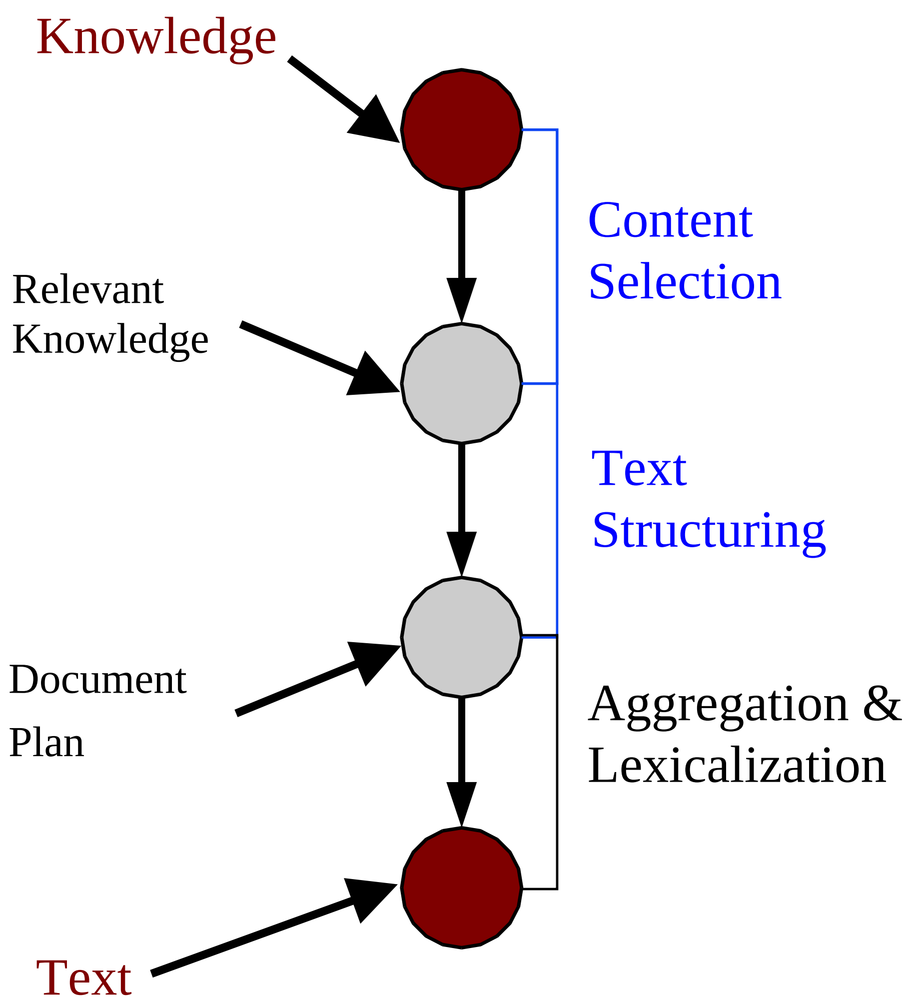
Example: for node $\in$ **relative**$\rightarrow$**relative**$\rightarrow$**name**$\rightarrow$**first**, if node$\leftarrow$name$\leftarrow$relative$\rightarrow$type $\in$ $\{\text{son}, \text{daughter}\}$, then **select node**.

`AND,OR` Plus logic combinators.

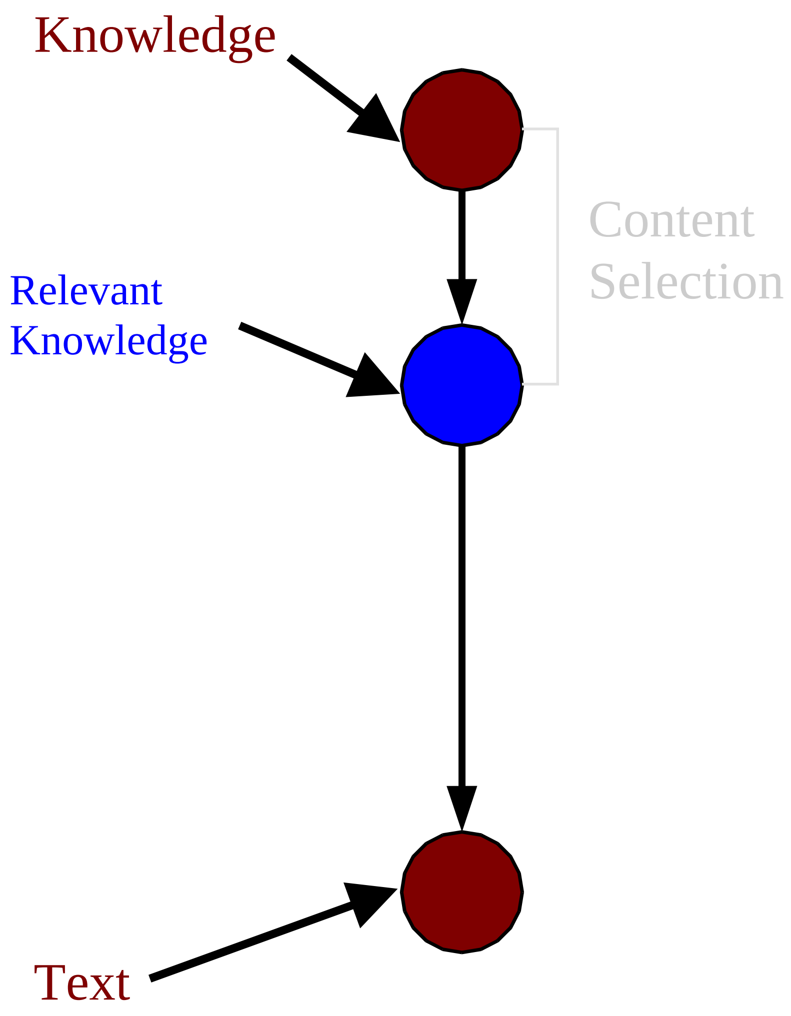
Talk Structure

- The Problem
- **My Solution**
  - Indirect Supervised Learning
  - Technique Overview
  - Example
  - Details
- Experiments
- Conclusions

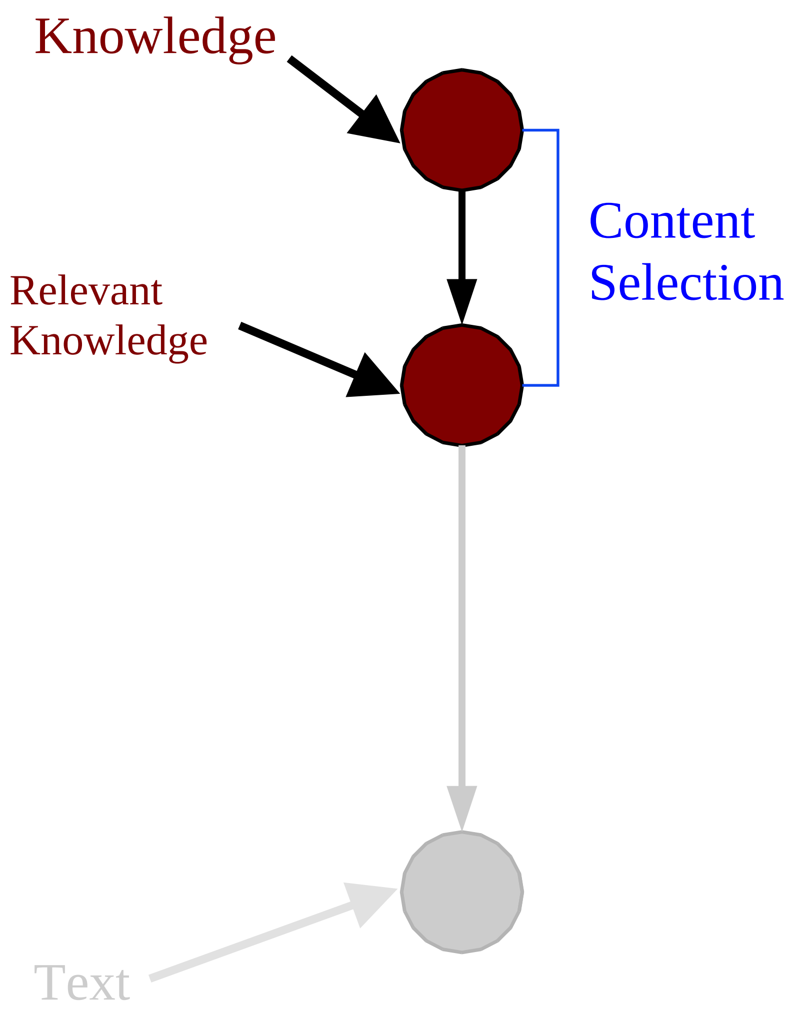
Indirect Supervised Learning: Overview



Indirect Supervised Learning: Overview



Indirect Supervised Learning: Overview



- **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

↔

⟨**name first**⟩	John	⟨**name last**⟩	Doe
⟨weight⟩	150Kg	⟨height⟩	160cm

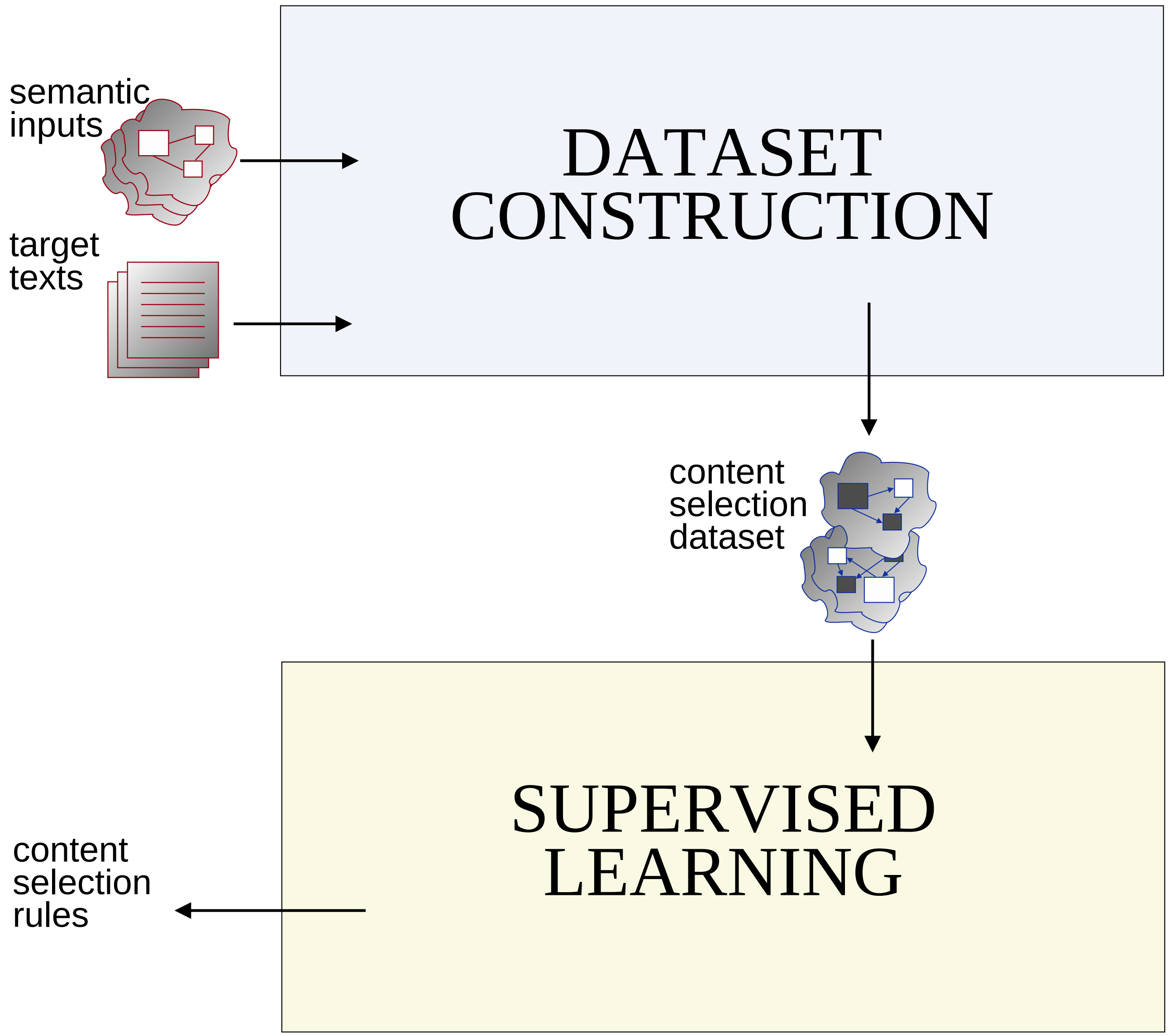
- **Learning As Automated Knowledge Acquisition**
  - Learning Structures That Humans Can Understand.
  - Mixing Machine Learning And Knowledge-based Approaches.
  - Domain-independence Through Learning.

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

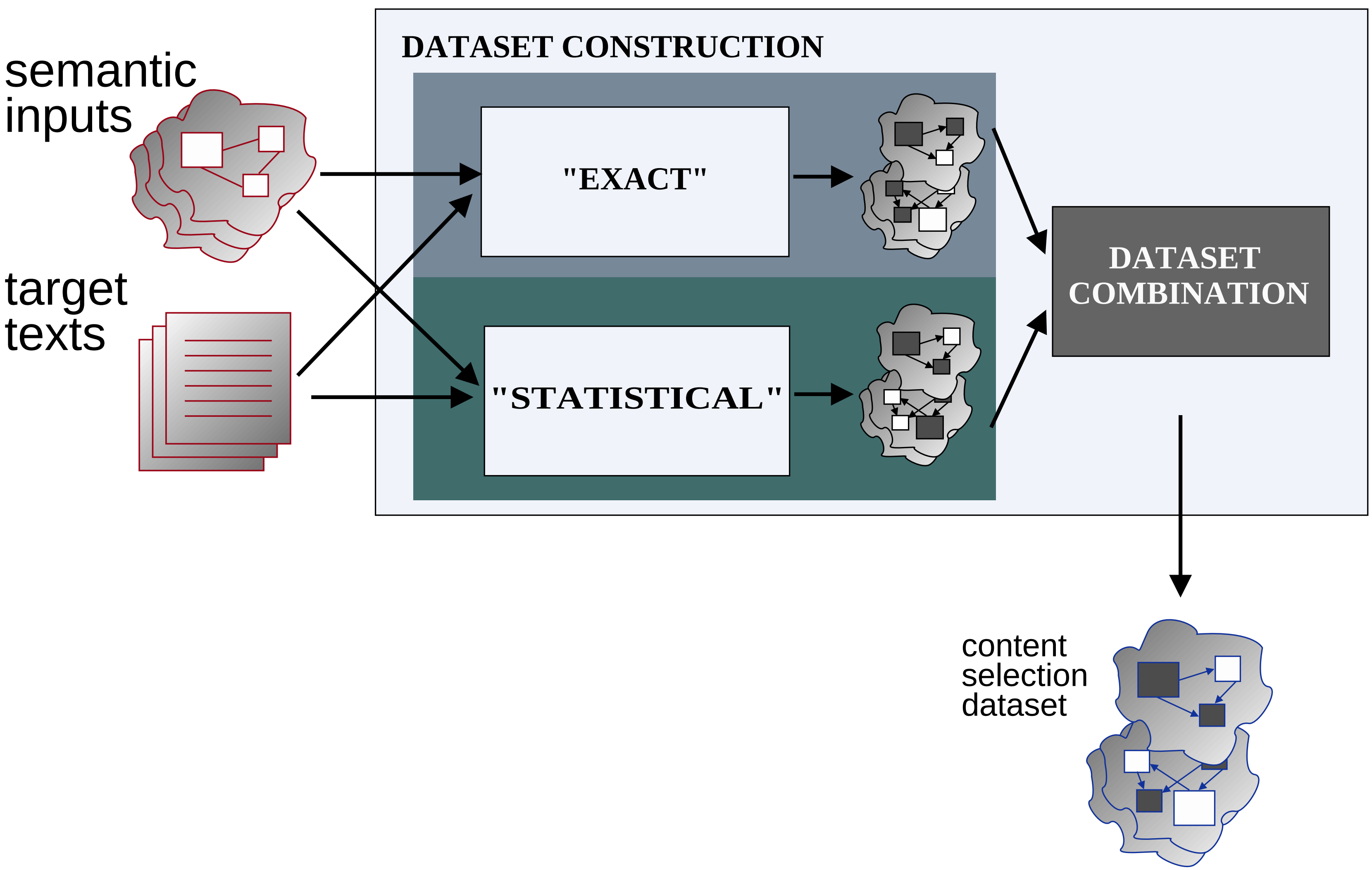
- Given:
  - $(KB_1, Bio_1), (KB_2, Bio_2), (KB_3, Bio_3), (KB_4, Bio_4)$
- Cluster Knowledge Bases By Value:
  - $\{KB_1, KB_2\}$ contain $(\langle \texttt{birth} \rightarrow \texttt{place} \rightarrow \texttt{state} \rangle, `MD')$
  - $\{KB_3, KB_4\}$ contain $(\langle \texttt{birth} \rightarrow \texttt{place} \rightarrow \texttt{state} \rangle, `NY')$
- Compare Language Models Of Clusters:
  - Compare the models of $\{Bio_1, Bio_2\}$ against $\{Bio_3, Bio_4\}$.
  - If the models differ, select $\langle \texttt{birth} \rightarrow \texttt{place} \rightarrow \texttt{state} \rangle$.

- $Bio_1 \Rightarrow$ *"... born in Maryland..."*
- $Bio_2 \Rightarrow$ *"... from Maryland..."*
- $Bio_3 \Rightarrow$ *"... native of New York..."*
- $Bio_4 \Rightarrow$ *"... born in New York..."*

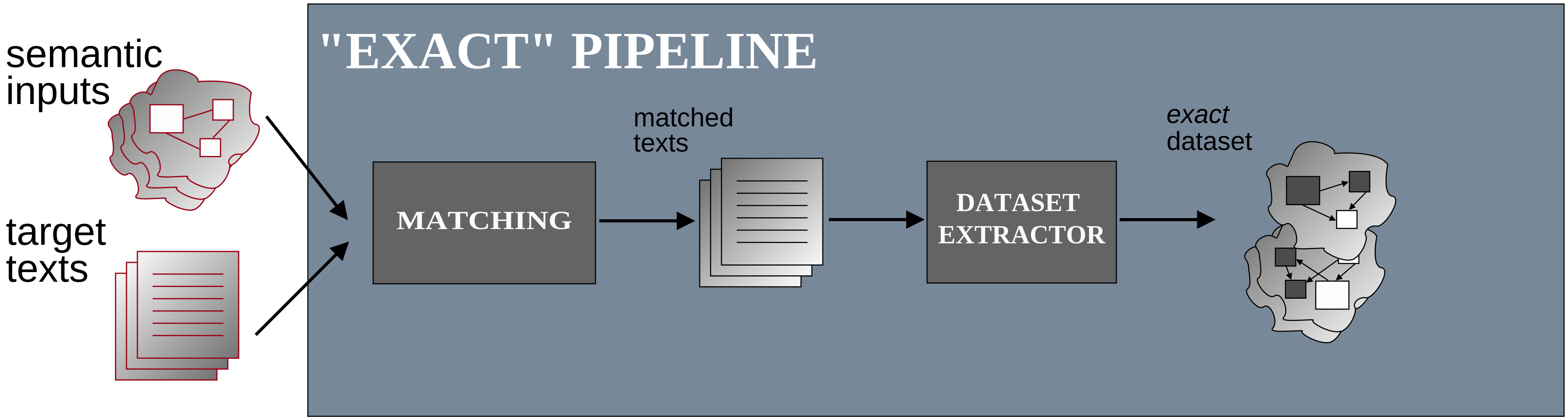
Methods: Indirect Supervised Learning



Methods: Dataset Construction



Dataset Construction: Exact Match Pipeline



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"

$\neg sel$ ⟨`name first`⟩ "Edward"

sel ⟨`birth date year`⟩ 1950

$\neg sel$ ⟨`occupation`⟩ `c-actor`

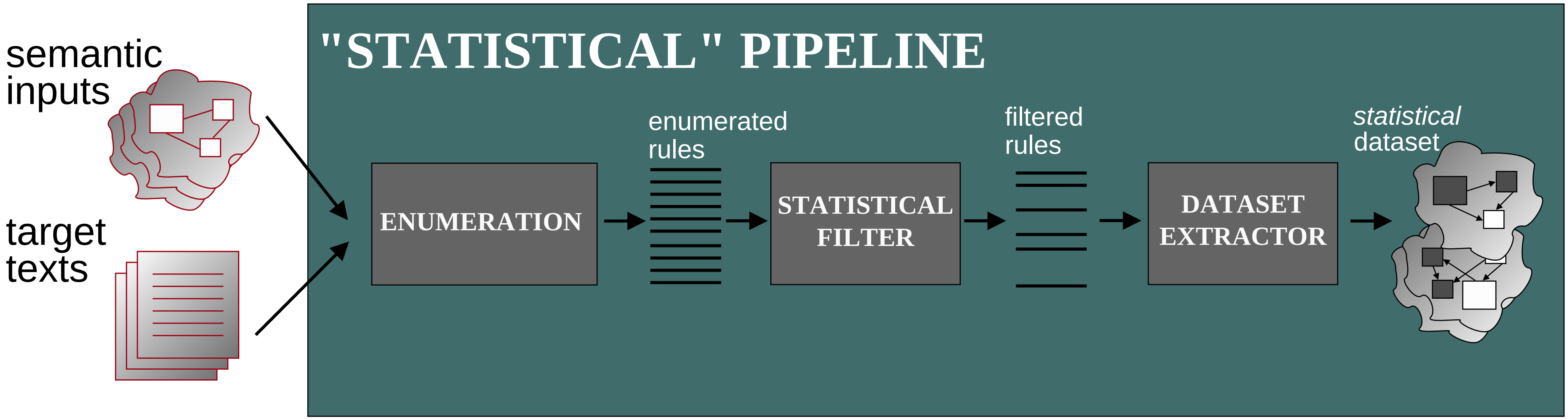
$\neg sel$ ⟨`birth date month`⟩ 11

sel ⟨`birth date day`⟩ 28

sel ⟨`birth place city`⟩ "Tenafly"

$\neg sel$ ⟨`birth place province`⟩ "NJ" . . .

Dataset Construction: Statistical Pipeline



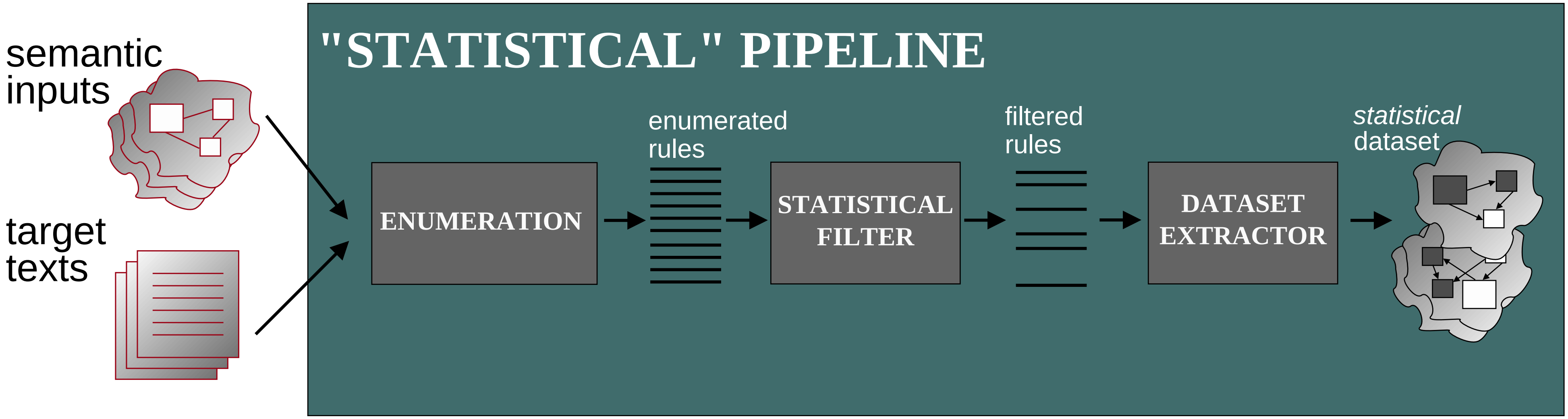
$$\{KB_1, KB_2, KB_3, KB_4\}$$

$$\downarrow$$

$$(\langle \texttt{birth place state} \rangle, \text{`}MD\text{'}) \Rightarrow \{KB_1, KB_2\} \Rightarrow \{Bio_1, Bio_2\}$$

$$(\langle \texttt{birth place state} \rangle, \text{`}NY\text{'}) \Rightarrow \{KB_3, KB_4\} \Rightarrow \{Bio_3, Bio_4\}$$

Dataset Construction: Statistical Pipeline



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

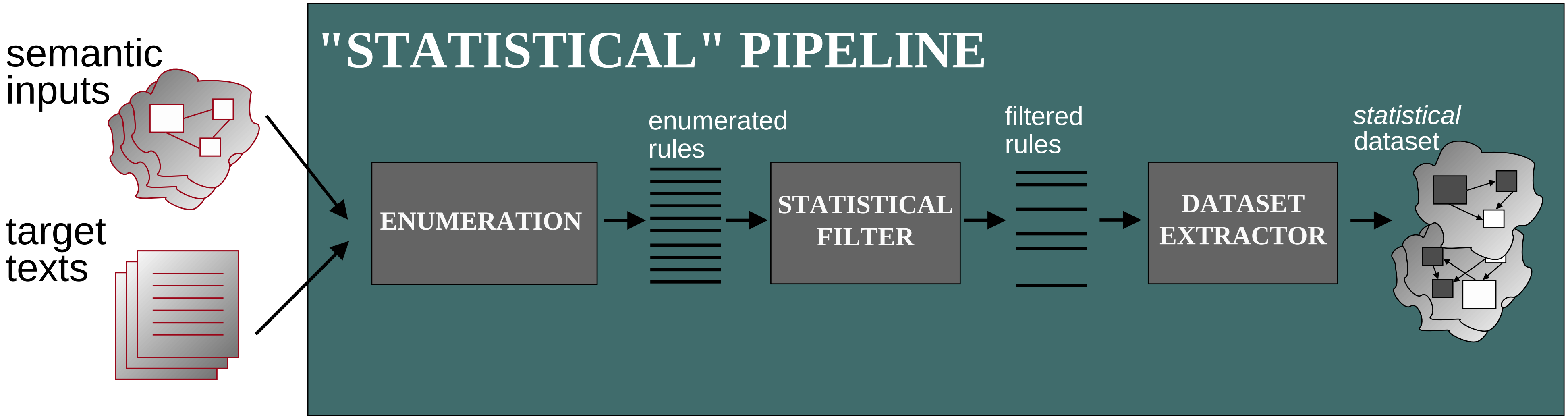
Cluster

Word	Count
New	6
York	5
The	10
other	7

Outside Cluster

Word	Count
New	1
York	0
The	11
other	6

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.

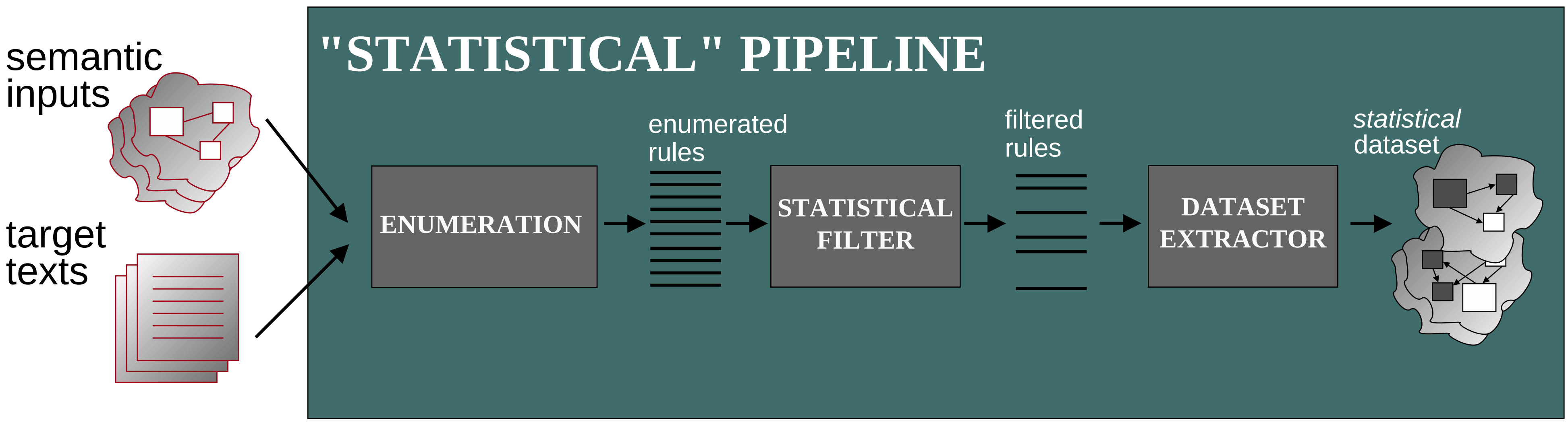
Cluster

Word	Count
New	6
York	5

Outside Cluster

Word	Count
New	1
York	0

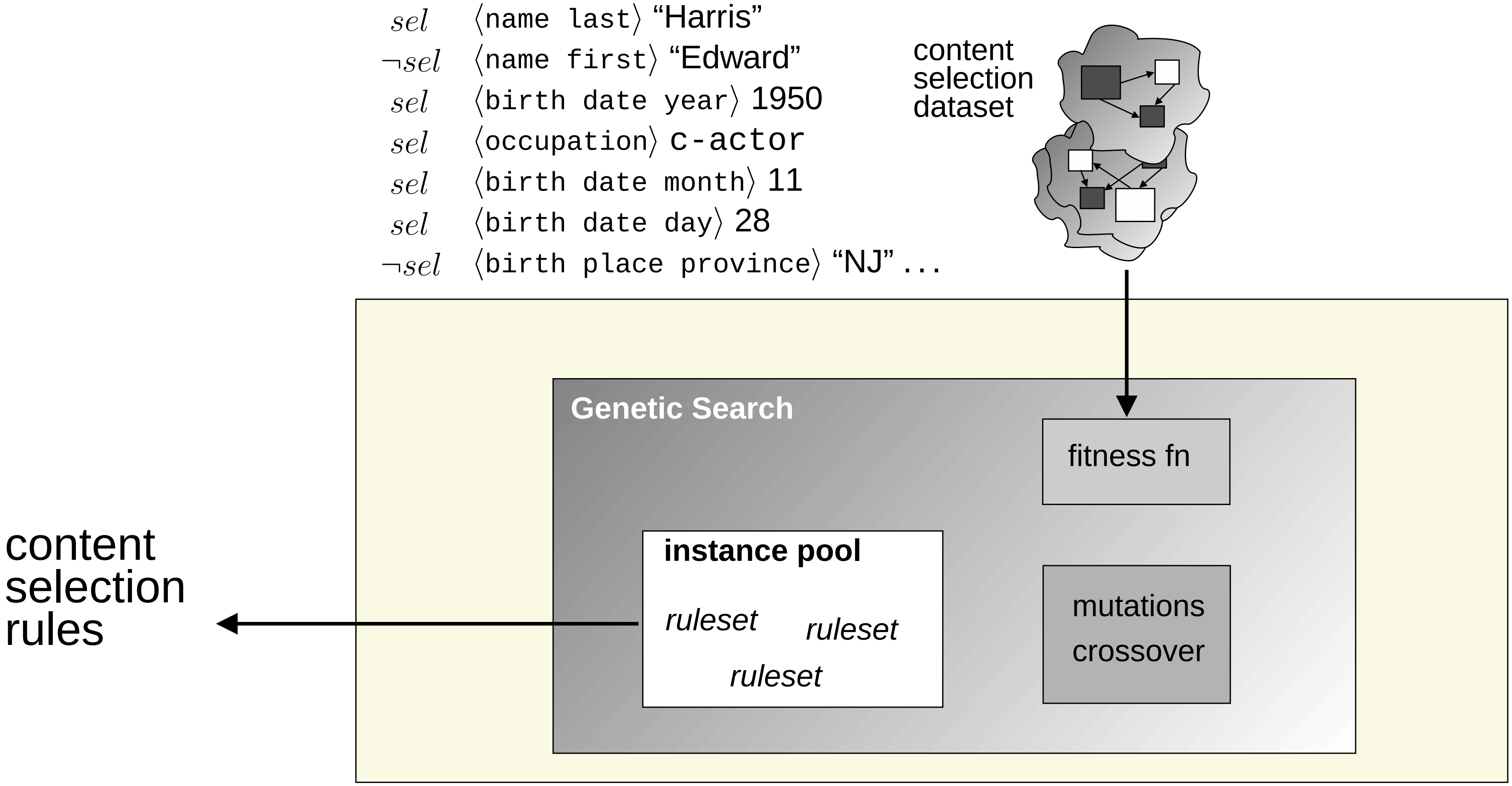
Dataset Construction: Statistical Pipeline



- 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

sel ⟨name last⟩ “Harris”
$\neg sel$ ⟨name first⟩ “Edward”
sel ⟨birth date year⟩ 1950
sel ⟨occupation⟩ c-actor
sel ⟨birth date month⟩ 11
sel ⟨birth date day⟩ 28
$\neg sel$ ⟨birth place province⟩ “NJ” . . .



- 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.

- 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).

Novel Fitness Function Over Training Set

I use the weighted F-measure over the labels as fitness:

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

where

$$F^*_\alpha = \frac{(\alpha^2 + 1)\, PrecRec}{\alpha^2 Prec + Rec}$$

$$\text{MDL} = \text{a minimum description length term}$$

This function captures the problem well and allows selecting solutions that prefer precision or recall through the α 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

Exp.	Exact Match	Combined
Prec.	**0.75**	0.73
Rec.	0.64	0.69
F^*	0.69	**0.71**

- 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

Experiment	biography.com				imdb.com			
	Selected	Prec.	Rec.	F*	Selected	Prec.	Rec.	F*
random	162	0.29	0.48	0.36	369	0.25	0.50	0.33
select-all	1129	0.26	1.00	0.41	1584	0.23	1.00	0.37
true/false rules	550	0.41	0.94	0.58	891	0.36	0.88	0.51
only exact match	359	0.64	0.61	0.62	432	0.48	0.65	0.55
combined	292	0.57	0.81	0.67	432	0.49	0.68	0.57
test set	293	-	-	-	369	-	-	-

Evaluation Of Content Selection Rules

Experiment	biography.com				imdb.com			
	Selected	Prec.	Rec.	F*	Selected	Prec.	Rec.	F*
random	162	0.29	0.48	0.36	369	0.25	0.50	0.33
select-all	1129	0.26	1.00	0.41	1584	0.23	1.00	0.37
true/false rules	550	0.41	0.94	0.58	891	0.36	0.88	0.51
only exact match	359	0.64	0.61	0.62	432	0.48	0.65	0.55
combined	292	0.57	0.81	0.67	432	0.49	0.68	0.57
test set	293	-	-	-	369	-	-	-

My new rules select comparable amount of data.

Evaluation Of Content Selection Rules

Experiment	biography.com				imdb.com			
	Selected	**Prec.**	Rec.	F*	Selected	**Prec.**	Rec.	F*
random	162	**0.29**	0.48	0.36	369	**0.25**	0.50	0.33
select-all	1129	**0.26**	1.00	0.41	1584	**0.23**	1.00	0.37
true/false rules	550	**0.41**	0.94	0.58	891	**0.36**	0.88	0.51
only exact match	359	**0.64**	0.61	0.62	432	**0.48**	0.65	0.55
combined	292	**0.57**	0.81	0.67	432	**0.49**	0.68	0.57
test set	293	-	-	-	369	-	-	-

They increase precision, although the statistical pipeline has its toll.

Evaluation Of Content Selection Rules

Experiment	biography.com				imdb.com			
	Selected	Prec.	**Rec.**	F*	Selected	Prec.	**Rec.**	F*
random	162	0.29	**0.48**	0.36	369	0.25	**0.50**	0.33
select-all	1129	0.26	**1.00**	0.41	1584	0.23	**1.00**	0.37
true/false rules	550	0.41	**0.94**	0.58	891	0.36	**0.88**	0.51
only exact match	359	0.64	**0.61**	0.62	432	0.48	**0.65**	0.55
combined	292	0.57	**0.81**	0.67	432	0.49	**0.68**	0.57
test set	293	-	-	-	369	-	-	-

The statistical pipeline increases recall.

Evaluation Of Content Selection Rules

Experiment	biography.com				imdb.com			
	Selected	Prec.	Rec.	F*	Selected	Prec.	Rec.	F*
random	162	0.29	0.48	**0.36**	369	0.25	0.50	**0.33**
select-all	1129	0.26	1.00	**0.41**	1584	0.23	1.00	**0.37**
true/false rules	550	0.41	0.94	**0.58**	891	0.36	0.88	**0.51**
only exact match	359	0.64	0.61	**0.62**	432	0.48	0.65	**0.55**
combined	292	0.57	0.81	**0.67**	432	0.49	0.68	**0.57**
test set	293	-	-	-	369	-	-	-

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

$$\beta = 1.5\frac{log(\frac{s}{1-s})}{t}$$
$$MDL = -\frac{1}{1+e^{-\beta l}}$$

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.