Statistical Acquisition of Content Selection Rules for Natural Language Generation

Pablo A. Duboue and Kathleen R. McKeown

ARDA

Computer Science Department
Columbia University
in the city of New York

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

$\langle$`name first`$\rangle$	John	$\langle$`name last`$\rangle$	Doe
$\langle$`weight`$\rangle$	150Kg	$\langle$`height`$\rangle$	160cm

$\longleftarrow \ldots \longrightarrow$

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

$\longleftarrow \dots \longrightarrow$

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

$\longleftrightarrow$

⟨**`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

$\leftarrow \ldots \rightarrow$

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

- 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

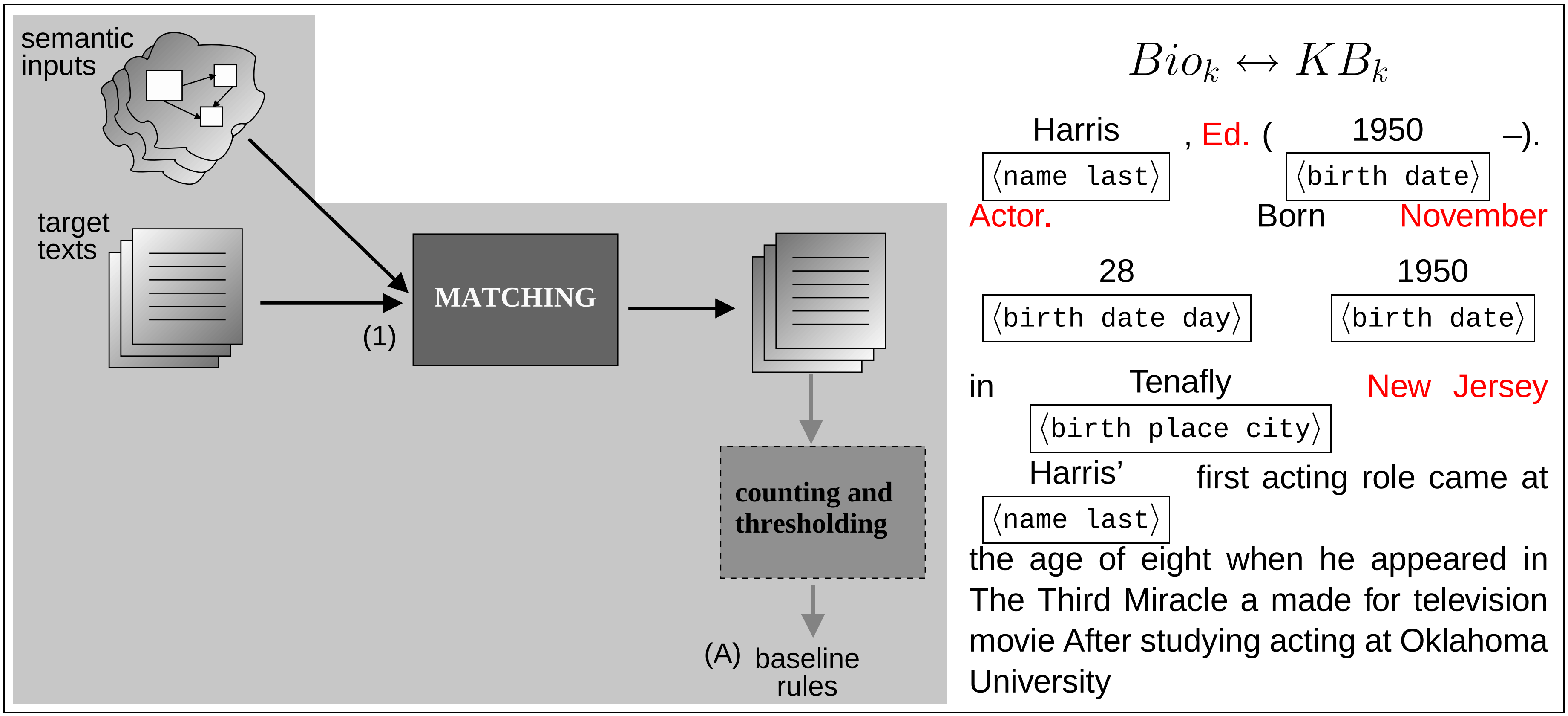
- PROGENIE has three major components
  1. Knowledge Component
  2. Generation Component
  3. Learning Component
- The Knowledge Component provides structured knowledge for the Generation Component
  - Noisy input
- The Learning Component trains off-line major parts of the Generation Component
  - Using 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`$\rangle$.

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

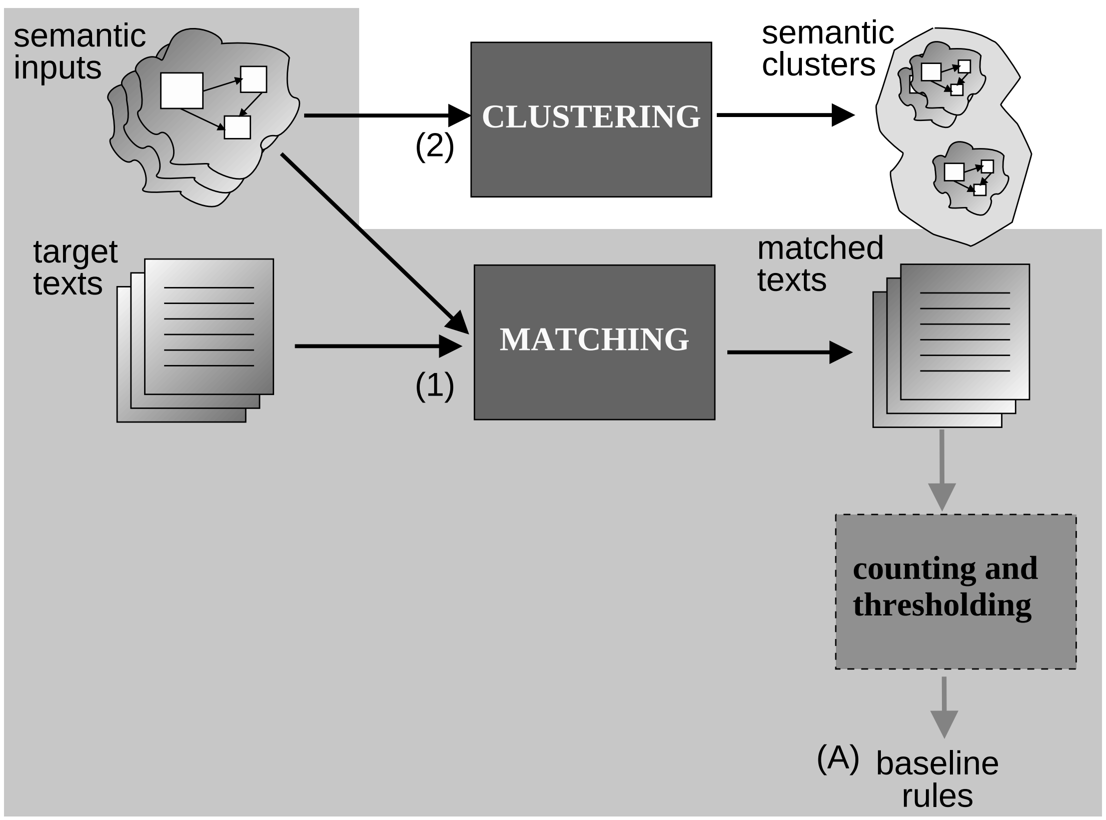
System


semantic inputs
target texts
(1)
MATCHING
counting and thresholding
(A) baseline rules
$Bio_k \leftrightarrow KB_k$
Harris
⟨name last⟩
, Ed. (
1950
⟨birth date⟩
–).
Actor.
Born
November
28
⟨birth date day⟩
1950
⟨birth date⟩
in
Tenafly
⟨birth place city⟩
New Jersey
Harris'
⟨name last⟩
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 ⟨`name last`⟩
  - Never select ⟨`height`⟩

System

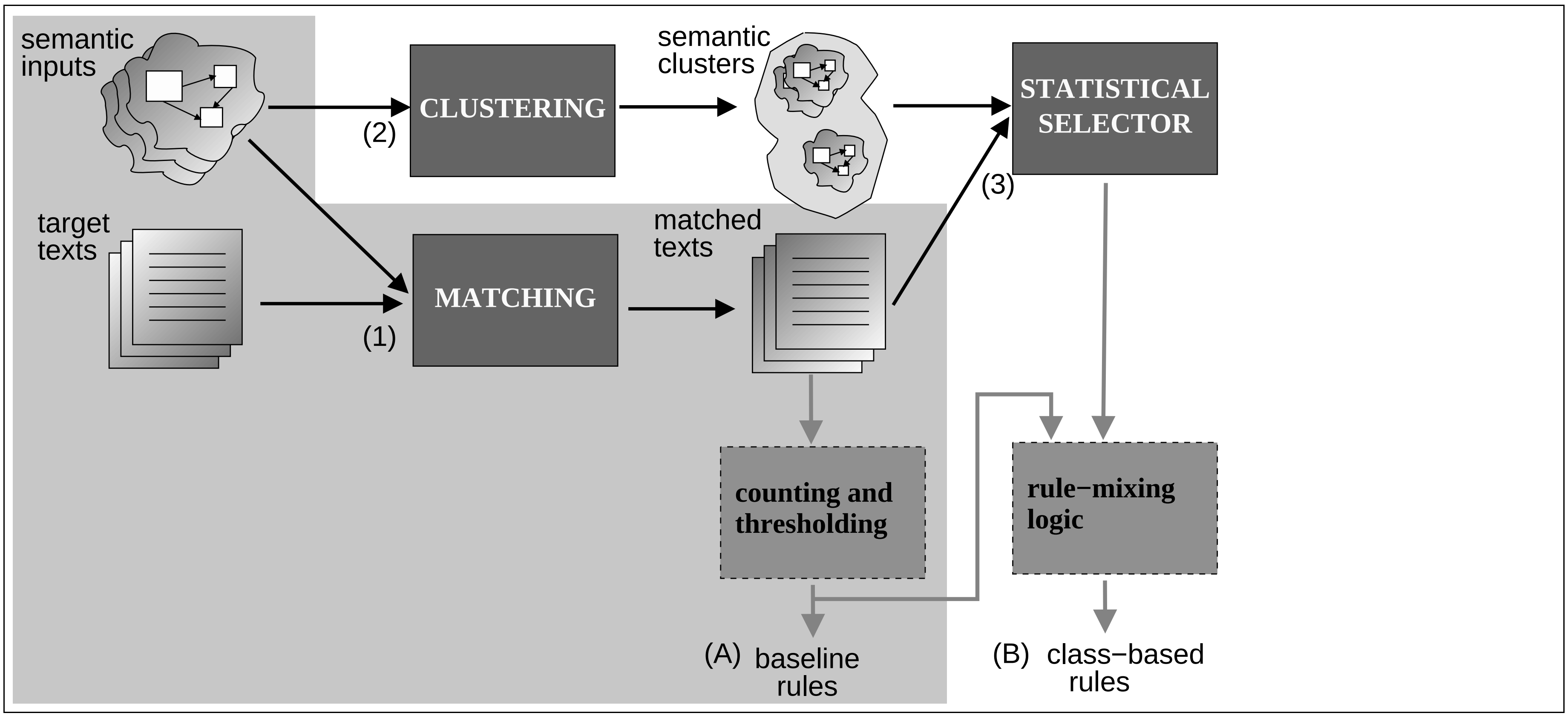


$$\{KB_1, KB_2, KB_3, KB_4\}$$
$$\downarrow$$

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

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

System

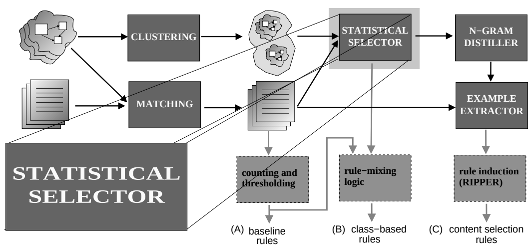

semantic inputs
CLUSTERING
semantic clusters
STATISTICAL SELECTOR
(2)
target texts
MATCHING
matched texts
(3)
(1)
counting and thresholding
rule−mixing logic
(A) baseline rules
(B) class−based rules

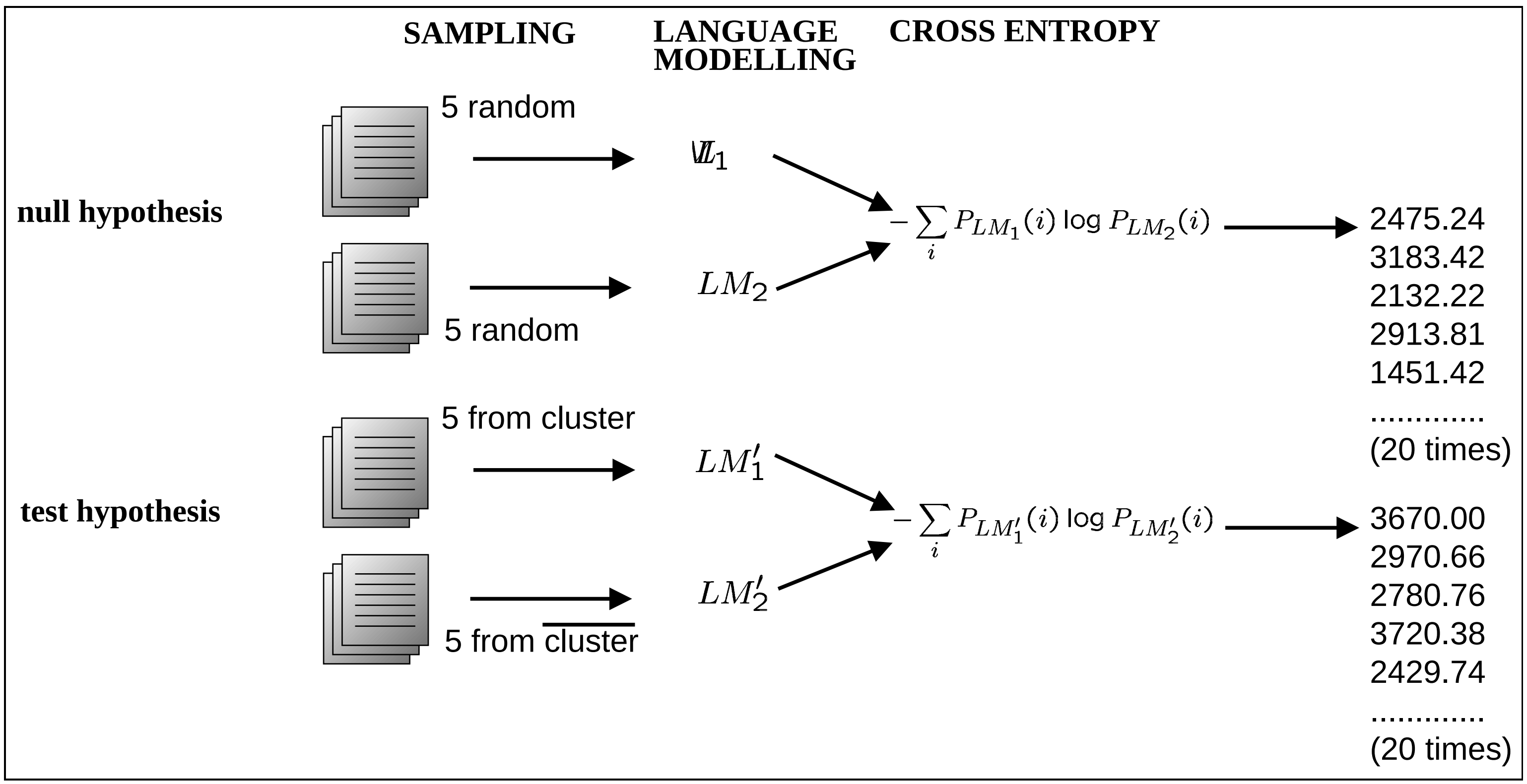
(B) Class-based Content Selection Rules

- Augment the baseline rules
- Select or unselect each and every instance of a given data path
- Example
  - Will add to the baseline rules like $\langle$`birth place state`$\rangle$
- Impact
  - Include datapaths where no exact match between data and text can be found (e.g., *“MD”* $\longrightarrow$ *“Maryland”*).

(3) Statistical Selector Module

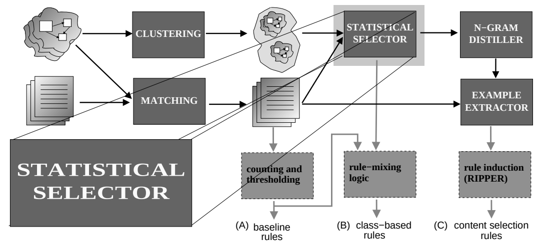
Find a change in word choice correlated with a change in data





(3) Statistical Selector Module

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



null hypothesis

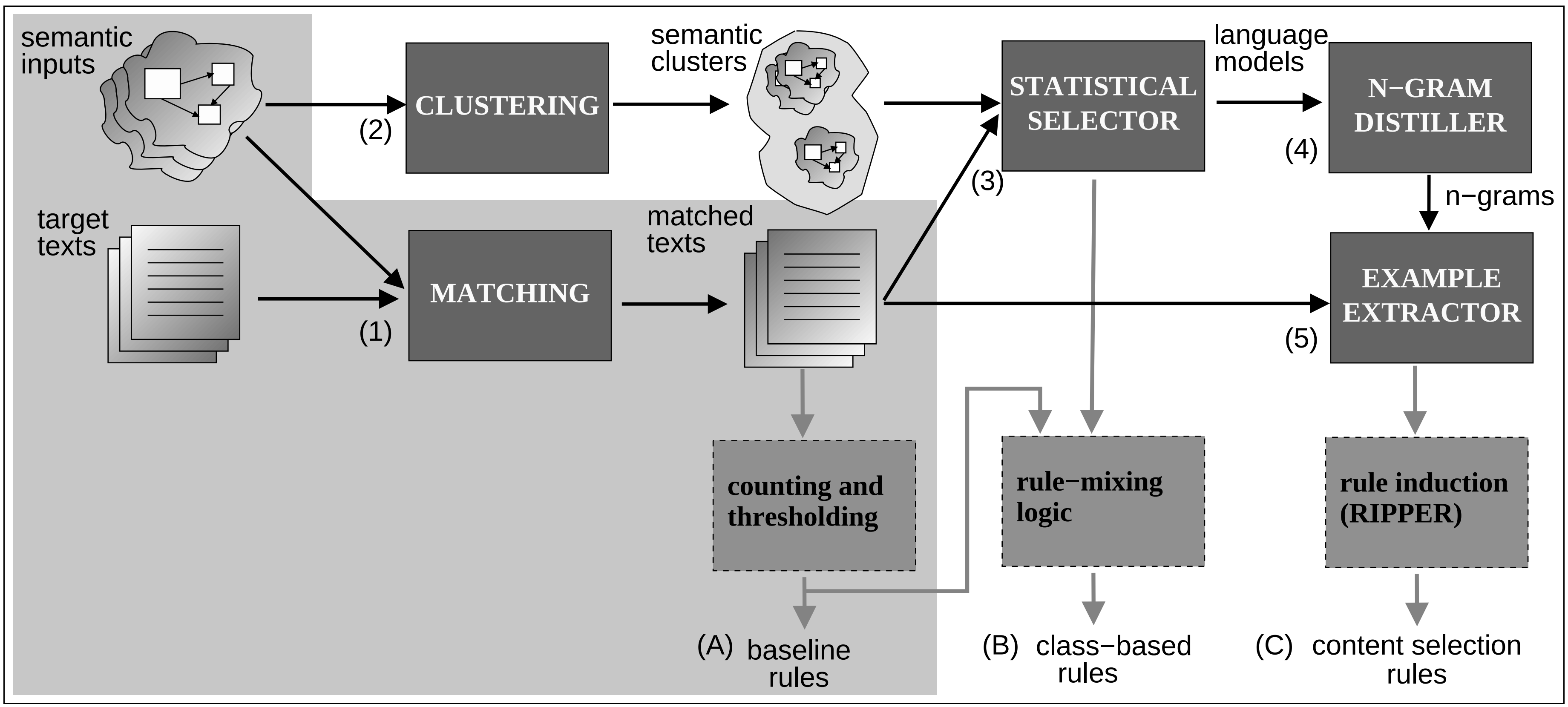
2475.24
3183.42
2132.22
2913.81
1451.42
....
(20 times)

larger? Mann−Whitney U test

test hypothesis

3670.00
2970.66
2780.76
3720.38
2429.74
....
(20 times)

System

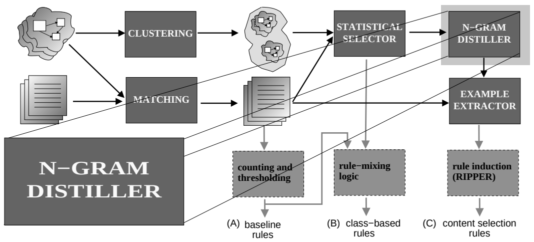


semantic inputs
CLUSTERING
(2)
semantic clusters
STATISTICAL SELECTOR
language models
N−GRAM DISTILLER
(4)
n−grams
(3)
target texts
MATCHING
(1)
matched texts
EXAMPLE EXTRACTOR
(5)
counting and thresholding
rule−mixing logic
rule induction (RIPPER)
(A) baseline rules
(B) class−based rules
(C) content selection rules

(C) Content Selection Rules

- Rules so far
  - Always include $\langle$`birth date day`$\rangle$ (baseline)
  - Always include $\langle$`birth place state`$\rangle$ (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* $\langle$*`award name`*$\rangle$

(4) n-gram Distiller Module



Obtaining finer grained information

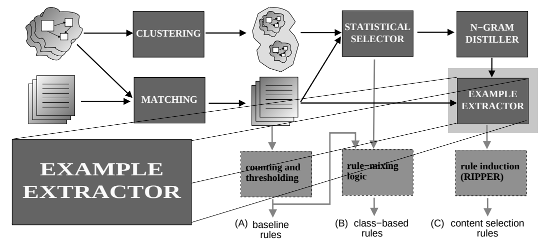
- The most significant n-grams were picked by looking at their overall contribution to the CE term

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

- 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

⟨`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 . . .

$\Downarrow$

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

$\leftrightarrow$

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

Experimental Setting

Two phases of training and testing

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

Development	Test
– 102 biographies	– 205 new biographies
– From `biography.com`	– From `imdb.com`
– Split into development training (91) and test (11)	– Split into training (191) and test (14)
– Hand-tagged the test set	– Hand-tagged the test set
– Average length: 450 words	– Average length: 250 words

- Content selection rules to be learned were different

Results

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

Results

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

Results

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

Results

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

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