Thesis Defense

Indirect Supervised Learning of Strategic Generation Logic

Pablo Ariel Duboue

Committe

Dr. Hirschberg (chair)
Dr. McKeown (advisor)
Dr. Jebara (internal)
Dr. Rambow (external)
Dr. Jurafsky (external)

Computer Science Department
Columbia University
in the city of New York

(A) Thesis-in-a-slide

- This is a thesis in Natural Language Generation (NLG).
  – NLG deals with the creation of text starting from knowledge.
- The knowledge needs to be:
  – filtered, **selected**;
  – ordered, **structured**.
- Selection and structuring are **domain dependent**.
  – Knowing how to structure medical reports does not help at all to structure biographies.
- This thesis:
  – uses machine learning to provide domain independent solutions to the Content Selection and Document Structuring problems.

- Choosing the right information to communicate.
  - Arguably the most critical part from the user's perspective.

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

(A) Content Selection

- Choosing the right information to communicate.
  - Arguably the most critical part from the user's perspective.

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

Always include $\langle$name $\rightarrow$ `first`$\rangle$.

(A) Content Selection

- Choosing the right information to communicate.
  - Arguably the most critical part from the user's perspective.

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

Never include $\langle$`weight`$\rangle$ or $\langle$`height`$\rangle$.

(A) Content Selection

- Choosing the right information to communicate.
  - Arguably the most critical part from the user's perspective.

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

Include only if $\langle\texttt{award} \rightarrow \texttt{name}\rangle \in \{\text{“Oscar”}\}$.

(A) Content Selection

- Choosing the right information to communicate.
  - Arguably the most critical part from the user's perspective.

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

Include only if $\langle\texttt{relative} \rightarrow \texttt{type}\rangle \in \{\text{c-son}\}$.
Include only if $\langle\texttt{relative} \rightarrow \texttt{name} \leftarrow \texttt{type}\rangle \in \{\text{c-son}\}$.

(A) Document Structuring

- Ordering and imposing a hierarchy to the information.
  - Conciseness and coherence goals.

Compare:

- *Diane Cilento is the mother of Jason. The movie ‘James Bond’ received an Oscar. Micheline Roquebrune is the wife of Sean Connery. Jason Connery is son of Sean Connery. Diane Cilento is an ex-wife of Sean Connery. The movie ‘James Bond’ is starred by Sean Connery.*

- *Sean Connery is an actor and producer. He married and later divorced the actress Diane Cilento and they have a child, Jason. He also married Micheline Roquebrune, a painter. Because he starred in the movie ‘James Bond’, he received an Oscar.*

(A) DS schemata

- A schema produces a sequence of **messages**.

$$
\begin{bmatrix}
\text{pred} & \texttt{education} \\
pred_0 & \texttt{person-32} \\
pred_1 & \texttt{"Columbia University"} \\
pred_2 & \texttt{"Computer Science"} \\
\text{mods} & \begin{bmatrix}
\text{time} & \begin{bmatrix} \text{start} & \texttt{"1999/8/27"} \\ \text{end} & \texttt{"2005/1/17"} \end{bmatrix} \\
\text{place} & \texttt{"New York, NY"}
\end{bmatrix}
\end{bmatrix}
$$

(A) DS schemata

- A schema produces a sequence of **messages**.
- Messages are instantiated a **predicates**.

```
predicate Education
  variables
     person             :  c-person
     education-event    :  c-education-event
  properties
    education-event ≡ person.education
  output
```

$$\left[\begin{array}{ll}
\text{pred} & \texttt{education} \\
pred_0 & \texttt{person} \\
pred_1 & \texttt{education-event}\rightarrow\texttt{teaching-agent} \\
pred_2 & \texttt{education-event}\rightarrow\texttt{subject-matter} \\
\text{mods} & \left[\begin{array}{ll}
\text{time} & \left[\begin{array}{ll} \text{start} & \texttt{education-event}\rightarrow\texttt{date-start} \\ \text{end} & \texttt{education-event}\rightarrow\texttt{date-end} \end{array}\right] \\
\text{place} & \texttt{education-event}\rightarrow\texttt{place} \\
\text{reason} & \texttt{education-event}\rightarrow\texttt{reason}
\end{array}\right]
\end{array}\right]$$

(A) DS schemata

- A schema produces a sequence of **messages**.
- Messages are instantiated a **predicates**.
- A **schema** is a finite state automaton over the language of predicates

```
intro-person(self),
education(self,education)* ,
(spouse(self,spouse), intro-person(spouse);
   { child(spouse,self,child),
      intro-person(child) } )*
(movie(self,movie), intro-movie(movie);
   { award(movie,self,award),
      intro-award(award,self) } )*
```

(A) DS schemata

- A schema produces a sequence of **messages**.
- Messages are instantiated a **predicates**.
- A **schema** is a finite state automaton over the language of predicates
- Example **document plan** (sequence of messages)

```
 intro-person(person-1), ex-spouse(person-1,person-2),
intro-person(person-2), spouse(person-1,person-3),
intro-person(person-3), child(person-1,person-4),
intro-person(person-4), movie(bond-1,person-1),
intro-award(oscar-1,person-1)
```

(A) Indirect Supervised Learning

- Use the Text-Knowledge corpus.
  - To obtain matched texts.
- Use the matched texts.
  - To obtain Content Selection labels.
  - To obtain semantic sequences.
- Use the Content Selection labels
  - To learn CS rules.
- Use the semantic sequences
  - To learn schemata.

(A) Matched Text Example

TEXT

Ryder, Winona (1971—) Actress. Born Winona Laura Horowitz on October 29, 1971 in Winona, Minnesota. Named after the city where she was born, she is the third of four siblings (including one half−brother and one half−sister from her mother's first marriage). Ryder's parents, Michael and Cindy (née Palmer) Horowitz, were hippie intellectuals, and family friends included the likes of beat poet Allen Ginsberg, and counterculture guru Timothy Leary who was Ryder's godfather. Ryder's family lived briefly in Colombia with Chilean revolutionaries before returning to northern California in 1974. Later, the family moved to a commune in Mendocino, where they lived for four years without television or electricity. They relocated to Petaluma, California in the early 1980s, where Ryder attended school and developed an interest in dramatic arts. At the age of 12, her parents encouraged her to enroll in the American Conservatory Theater (ACT) in San Francisco.

In 1985, Ryder was performing a monologue chosen from J.D. Salinger's "Franny & Zooey" at ACT when Deborah Lucchesi, a talent scout, ...

KNOWLEDGE

1	<birth date day>	29
2	<birth date month>	10
3	<birth date year>	1971
4	<birth father name first>	Michael
5	<birth father name last>	Horowitz
6	<birth name first>	Winona
7	<birth name givenname>	Laura
8	<birth name last>	Horowitz
9	<birth mother name first>	Cindy
10	<birth mother name last>	Horowitz
11	<birth place city>	Winona
12	<birth place province>	MN
13	<birth place country>	USA
14	<name last>	Ryder
15	<name first>	Winona
16	<occupation>	c-actress
17	<occupation>	c-model
18	<relative relative name first>	Michael,Cindy
19	<education place city>	San Francisco
20	<education teaching-agent>	American Conservatory Theater
21	<significant-other name first>	David

- Given training input I and output O pairs.
- To find the entity $e^* \in \mathrm{Schemas\ or\ Rules}$ such that

$$e^* = \underset{e}{argmax}\, P(e|I, O)$$

- Replace the probability with a likelihood $f(e, I, O)$
  - Define f by using e on I to obtain $O' = e(I)$.
  - $f(e, I, O) = ||O - e(I)|| = ||O - O'||$.
- The distances are fitness functions for a stochastic search process.

(A) CS Fitness Function Over Training Set

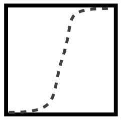
I use the weighted F-measure over the labels as fitness:

$$Fitness = F^*_\alpha + \textsc{mdl}$$

where

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

$$\textsc{mdl} = \text{a minimum description length term}$$



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

(A) DS Fitness: Three Tiers

1. Content Selection.
   - Same as before, but now measures Content Selection in-place.
2. Order Constraints.
   - Order Constraints mined in the training data, to qualify poor instances with crossing alignments.
3. Alignments.
   - Efficient, dynamic programming-based alignments (do not allow crossing alignments) with recurrences that compare sequences of atomic values to sequences of messages.

- Indirect Supervised Learning
  - Obtained hundreds of CS training instances, with an F^* as high as 0.7 and hundreds of DS training instances, with a Kendall's τ as high as 0.94.
- Content Selection
  - Three different learning methods, with different strengths and weaknesses. Results 8% below training material quality.
- Document Structuring
  - Mined order constraints in two domains.
  - Succeeded learning a simple schema in medical domain.
  - Promising results in biographies domain.

(B) Indirect Supervised Learning.

- Unsupervised construction of the matched texts.
- Biographies domain, 4 different styles.

(C) Content Selection Learning.

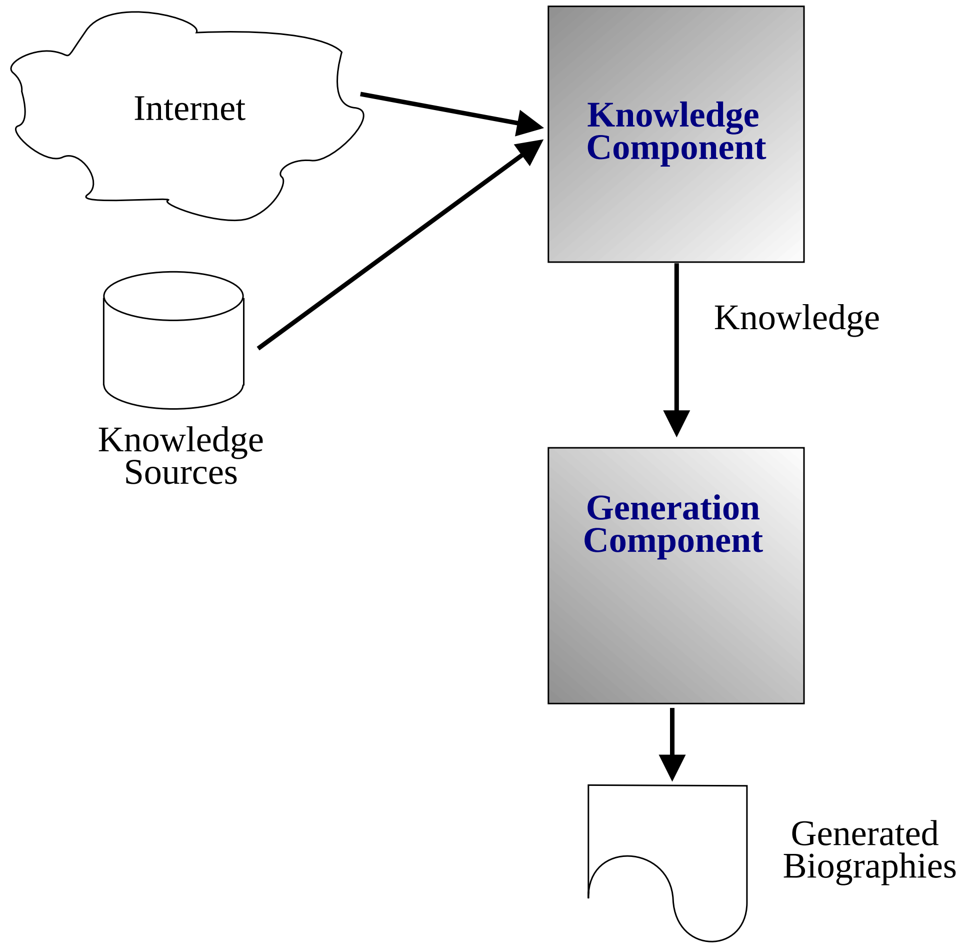
- Supervised learning of Content Selection rules.
- Biographies domain, 4 different styles.

(D) Document Structuring Learning.

- Unsupervised learning of order constraints.
- Supervised learning of schemata.
- Medical and biographies domain.

Indirect Supervised Learning

(B) Problem Revisited (Indirect Supervised Learning)



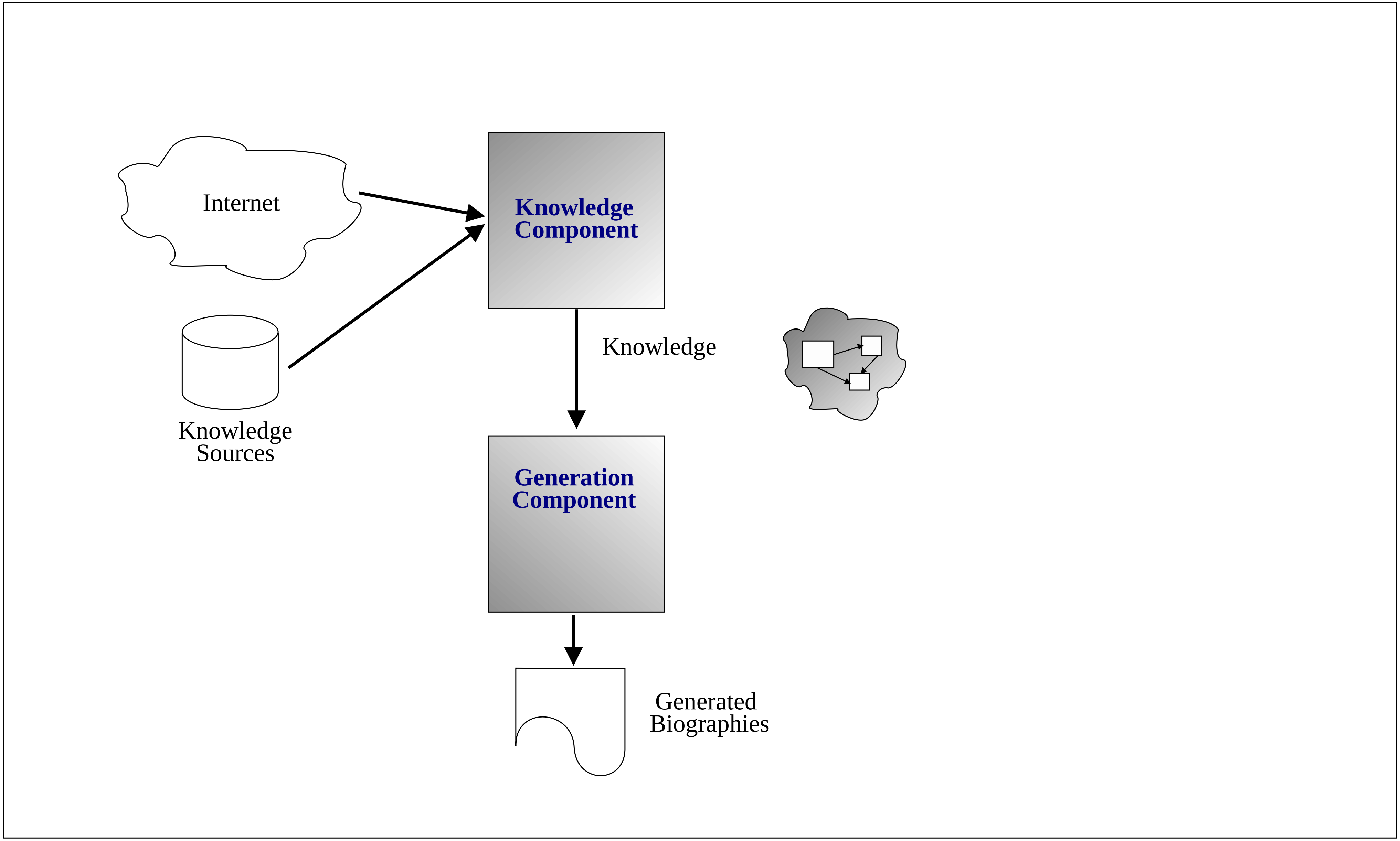
(B) Extraction Example

ALIASES	His aliases include Fadel Nazzal al- Khalayleh , Fadil al- Khalaylah , . . .
EDUCATION	He 's thought to be a high school dropout
TRASLATION	Zarqawi went to Afghanistan to fi ght the Soviets in the late 1980s
JOIN_ORG	In Afghanistan , Zarqawi plugged into the al Qaeda
PRISON	in the early 1990s , he was jailed and spent seven years in jail
BIRTH_DATE	al- Zarqawi (Arabic :) (possibly born on October 20 , 1966)
BIRTH_COUNTRY	al- Zarqawi (Arabic :) (possibly born on October 20 , 1966) is a shadowy . . .
BIRTH_NAME	Ahmad Fadeel al- Nazal al- Khalayleh (Arabic :) , is believed to be his real name
DESCRIBED	Zarqawi is usually described as somber and unintelligent
PRISON	in 2001 , al- Zarqawi was arrested again in Jordan
MASTERMIND	On July 11 , 2004 , Zarqawi claimed responsibility for a July 8 mortar attack in Samarra
MASTERMIND	Zarqawi has also claimed responsibility for the Canal Hotel bombing of the U . N . . . .
BIRTH_DATE	al- Zarqawi (possibly born on October 20 , 1966)
BIRTH_COUNTRY	al- Zarqawi (possibly born on October 20 , 1966) is a shadowy Jordanian national

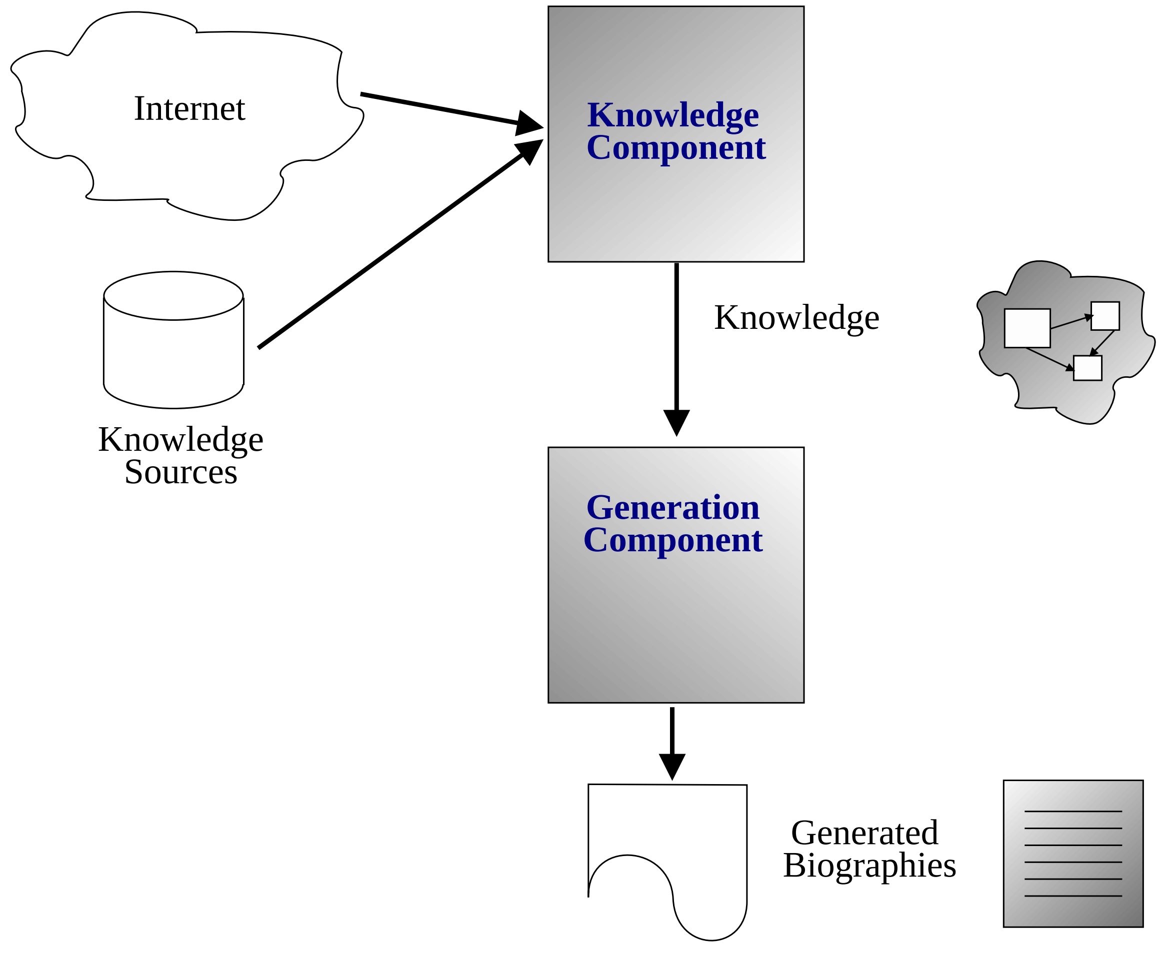
Extraction Example

PRISON	Zarqawi was jailed briefly in the 1980s for sexual assault
TRASLATION	In 1 989 , Zarqawi traveled to Afghanistan to fi ght against the Soviet invasion of . . .
TRASLATION	in the mid- 1990s , al- Zarqawi travelled to Europe
PRISON	he was arrested in Jordan in 1992
CREATE_ORG	In Afghanistan , al- Zarqawi established a terrorist training camp
PRISON	in 2001 , al- Zarqawi was arrested again in Jordan
MASTERMIND	On July 11 , 2004 , Zarqawi claimed responsibility for a July 8 mortar attack in Samarra
OCCUPATION	al- Zarqawi is a Palestinian jihadi leader
ALIAS	al- Zarqawi , A . K . A . Fedel Nazzel Khalayleh ,
JOIN_ORG	He is from the Beni Hassan tribe
MASTERMIND	Zarqawi has been implicated in terrorist activity worldwide
MASTERMIND	He has also been implicated in a foiled chemical weapons attack against Jordan ’s . . .
MASTERMIND	Zarqawi was behind the assassination of US diplomat Lawrence Foley in Amman , . . .
OCCUPATION	Al- Qaeda Zarqawi has been named as the leader of Jund al- Shams

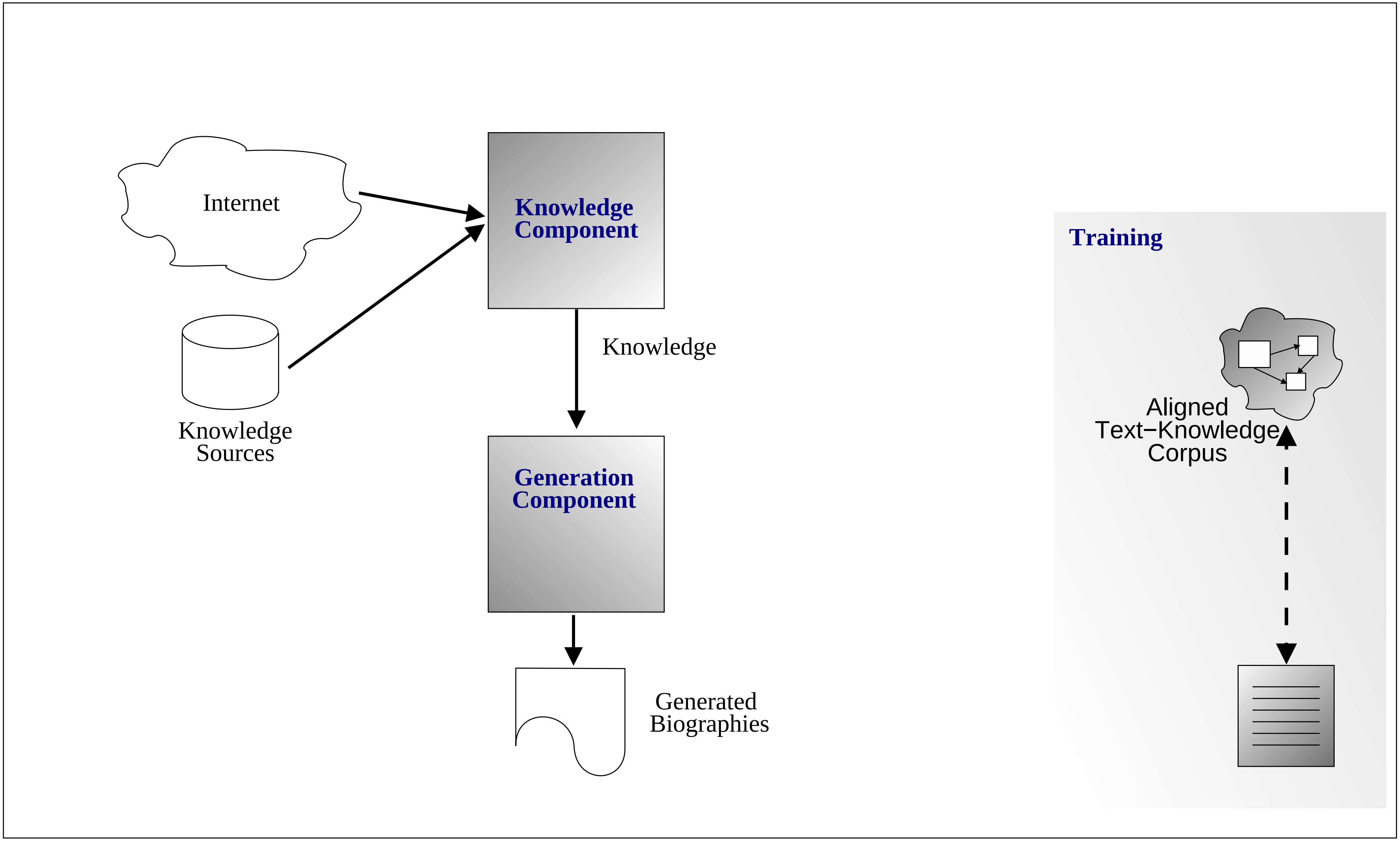
(B) Problem Revisited (Indirect Supervised Learning)



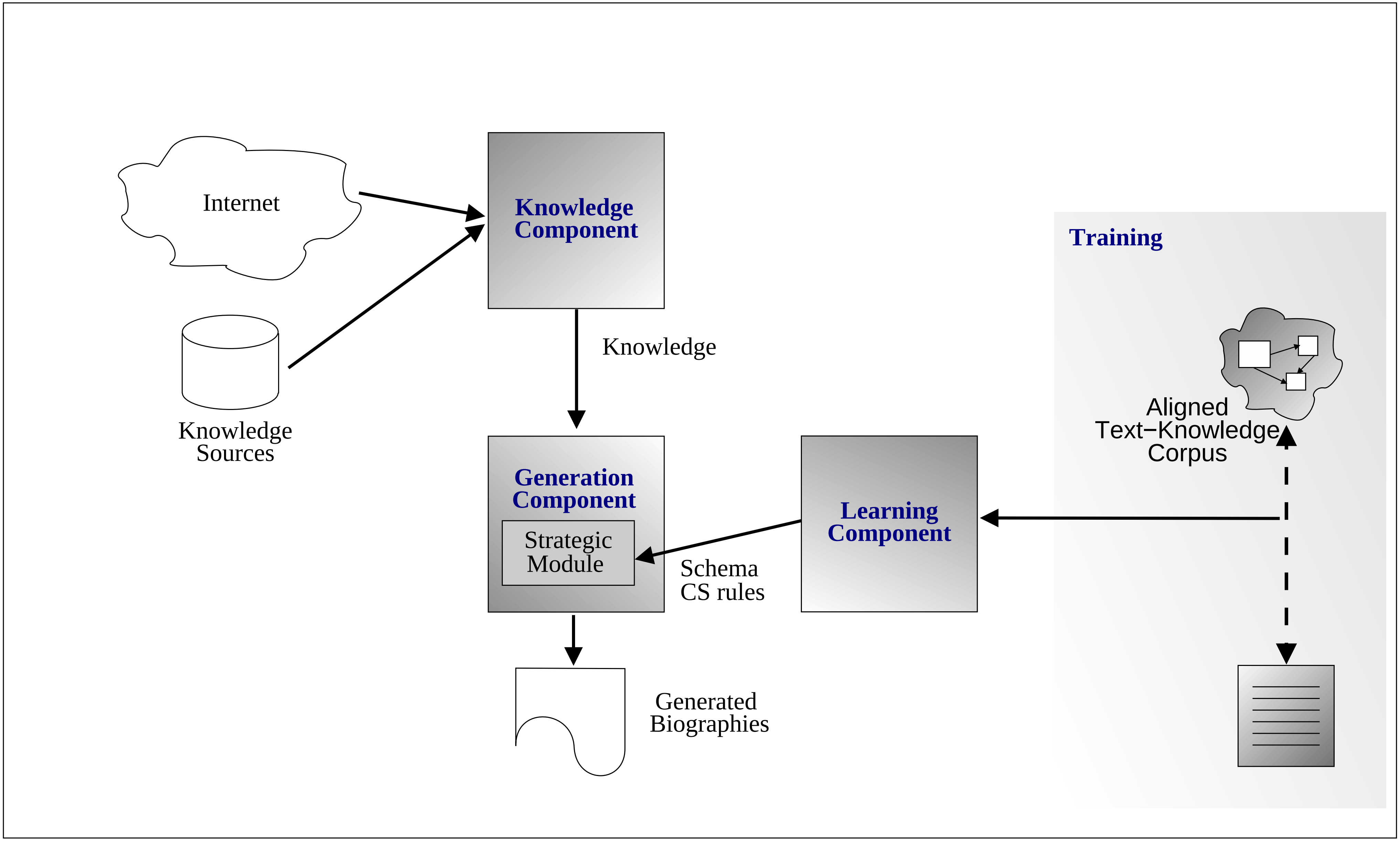
(B) Problem Revisited (Indirect Supervised Learning)



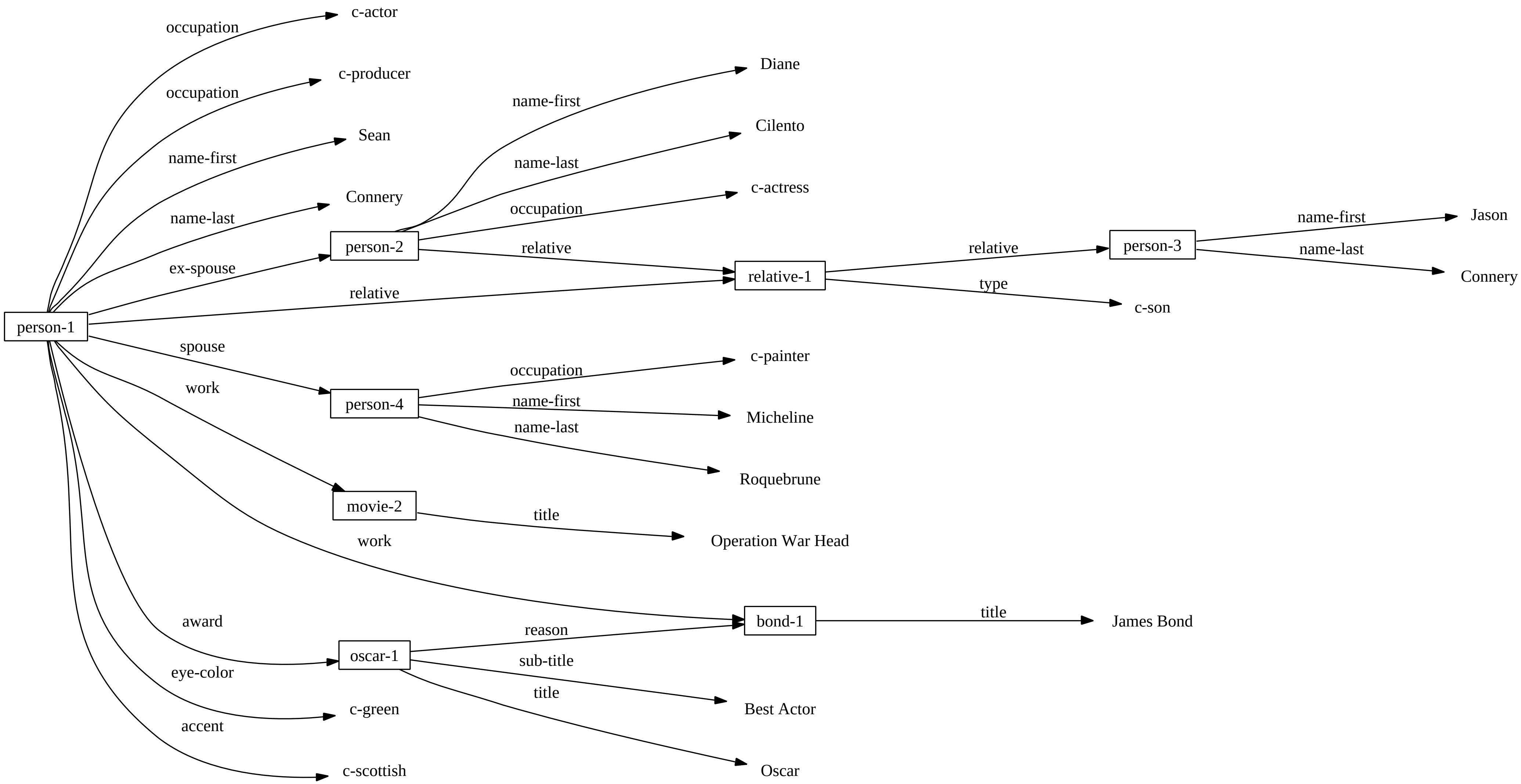
(B) Problem Revisited (Indirect Supervised Learning)



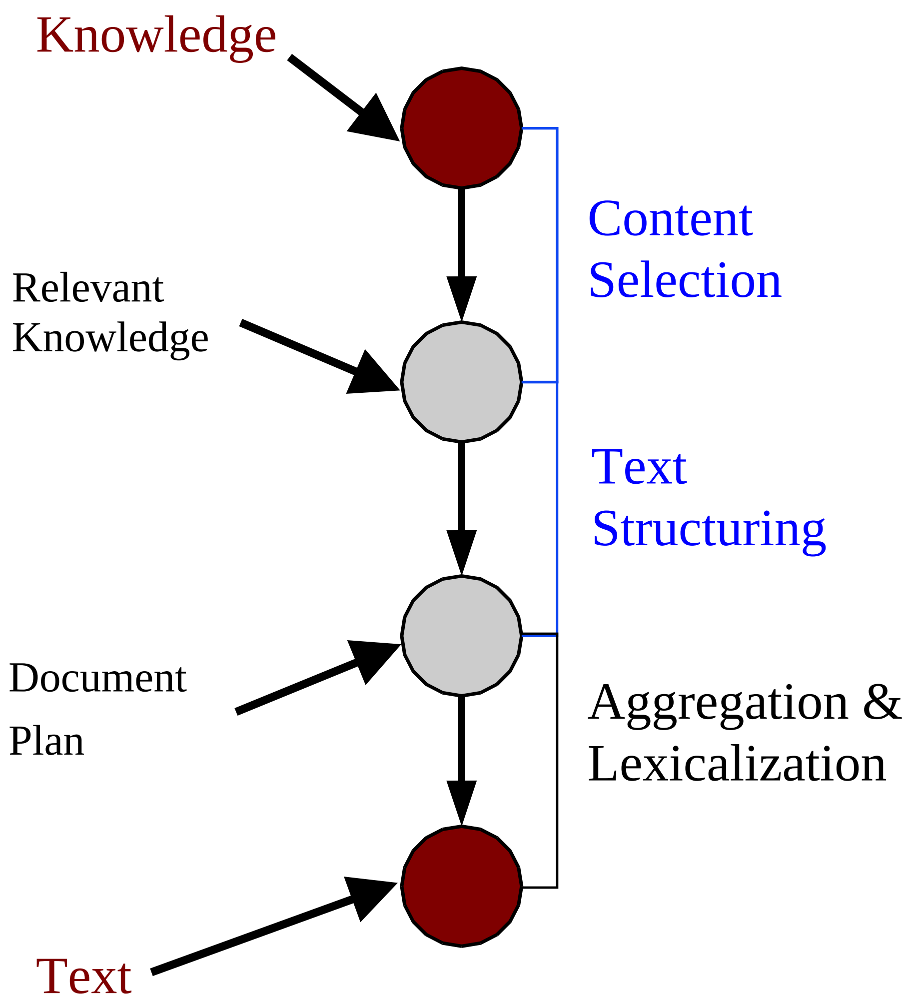
(B) Problem Revisited (Indirect Supervised Learning)



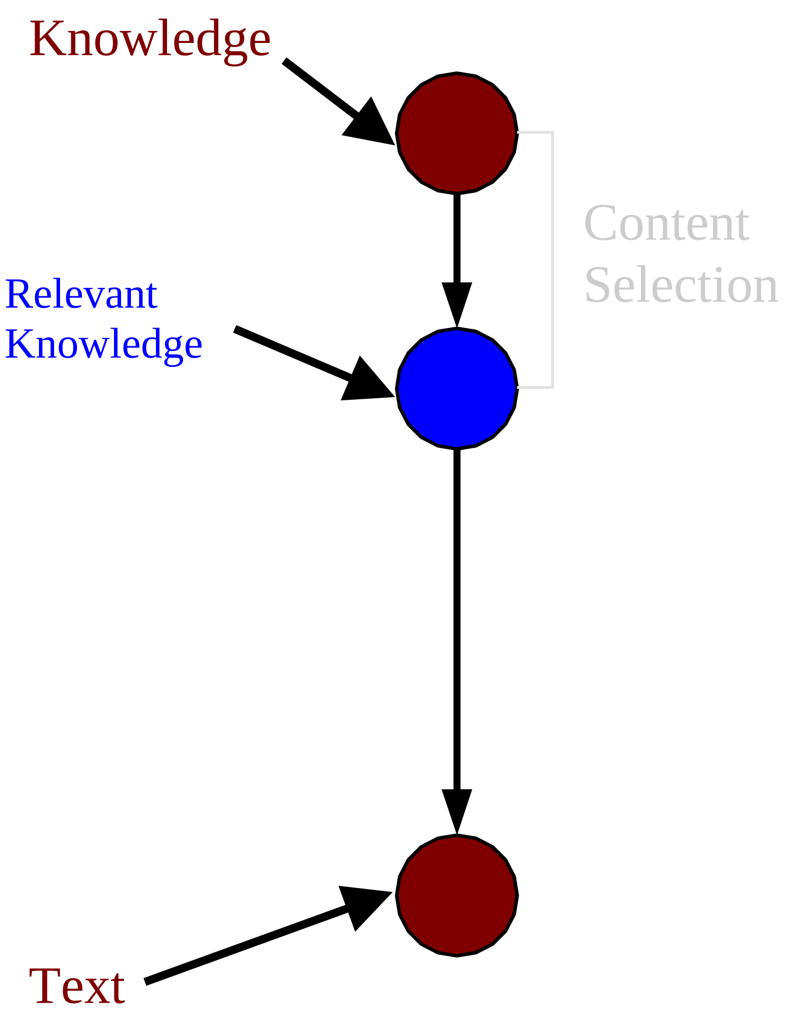
(B) Knowledge Representation



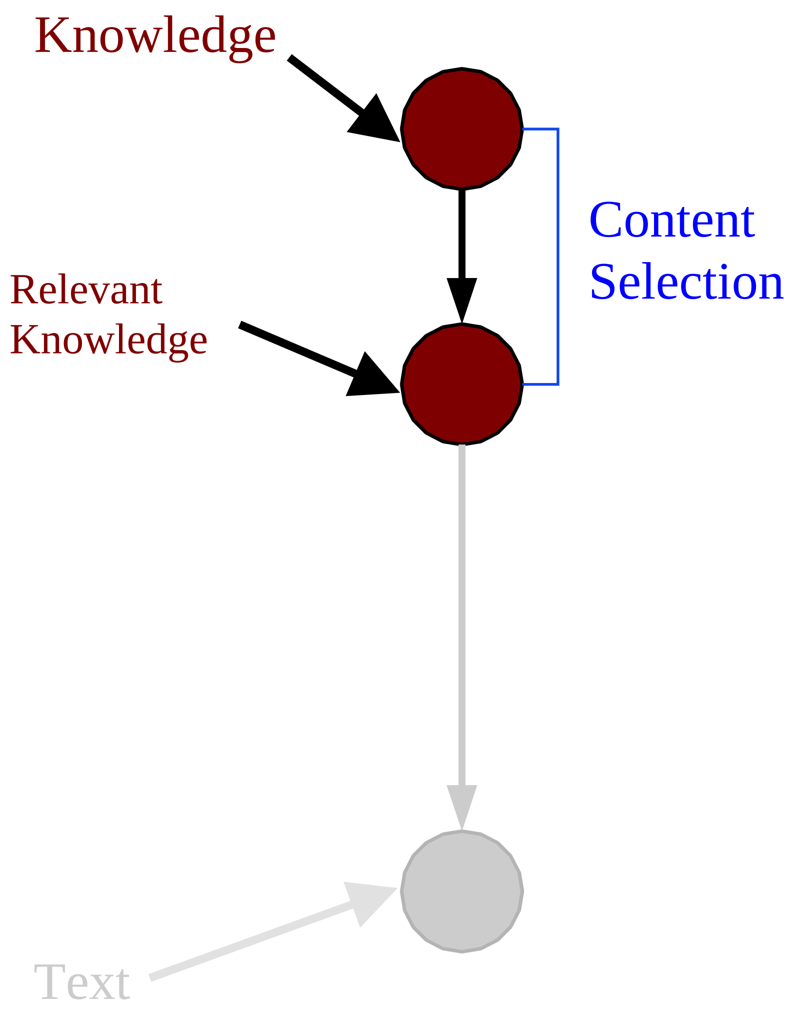
(B) Graphical Model



(B) Graphical Model



(B) Graphical Model



- Very simple model
  - Similar to IBM Model-1 for MT
  - $\mathcal{C}$: set of concepts
  - $\mathcal{P}$: set of phrases
  - $\mathcal{V}(c)$: set of phrases for concept c

- Test

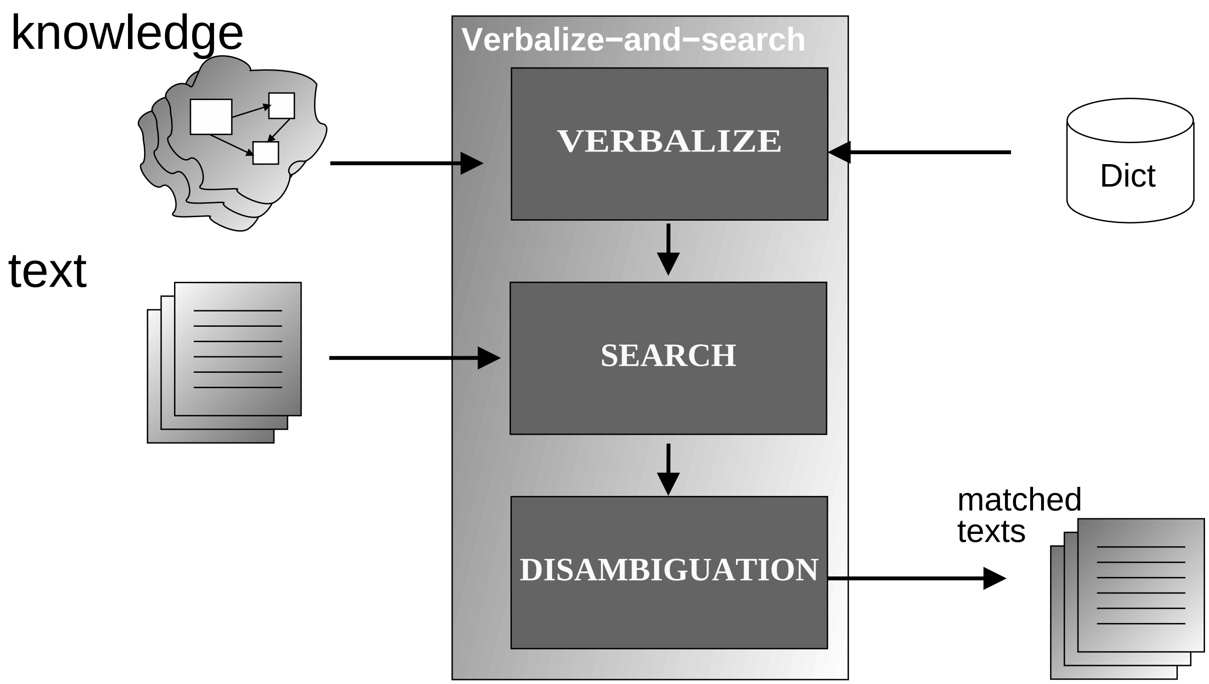
$$
\begin{aligned}
H_0: &\quad P\left(p \in \mathcal{P} | c \in \mathcal{C}\right) = p_0 = P\left(p \in \mathcal{P}\right) \quad \text{if } p \notin \mathcal{V}(c) \\
H_1: &\quad P\left(p \in \mathcal{P} | c \in \mathcal{C}\right) = p_1 \gg p_2 = P\left(p \in \mathcal{P}\right) \quad \text{if } p \in \mathcal{V}(c)
\end{aligned}
$$

(B) Example

- 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..."*

(B) Verbalize-and-search

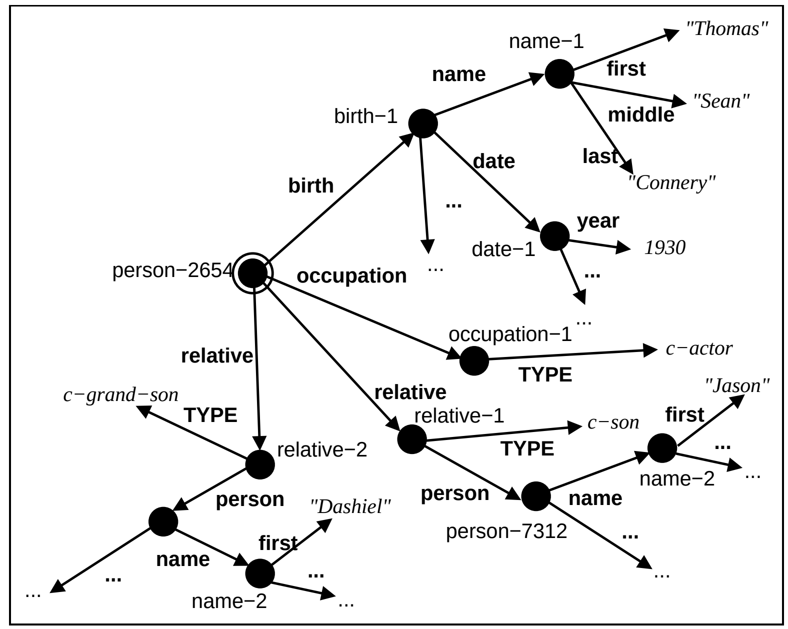


(B) Evaluation Methodology

- Split all training material into $Train$ and $Test$ sets.
  - Hand-tag $Test$ (for Content Selection, Ordering or both).
- Testing the Unsupervised algorithm:
  - Obtain labels over $Test$ and compare them to the hand-annotated ones.
  - (Actually, obtain labels over $Train+Test$ to have more training material and also have more insights of how well the system runs over $Train$.)
- Testing the overall Indirect Supervised algorithm:
  - Obtain in an unsupervised manner tags over $Train$.
  - Learn rules or schemata over the tags obtained over $Train$.
  - Execute the rules or schemata over $Test$ and compare to the hand-annotated tags.

(B) Corpora

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. . . .



(B) Corpora

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.

(B) Corpora

biography.com	Total	Average	Train	Test
# pairs	102	-	91	11
# triples	10,628	104.20	9,500	1,128
# words	54,001	529.42 ± 301.15	49,220	4,781
s9.com				
# pairs	578	-	558	20
# triples	95,032	164.42	92,969	2,063
# words	21,037	36.40 ± 34.04	20,192	845
imdb.com				
# pairs	199	-	185	14
# triples	31,676	159.18	29,323	2,353
# words	64,196	322.59 ± 285.63	60,086	4,110
wikipedia.org				
# pairs	361	-	341	20
# triples	108,009	299.19	102,297	5,712
# words	68,953	191.01 ± 55.17	64,784	4,169

(B) Baseline Variant

Corpus	Prec.	Rec.	F^*	selected
`biography.com`	0.74	0.64	0.69	297
`s9.com`	0.51	0.53	0.52	184
`imdb.com`	0.71	0.53	0.61	295
`wikipedia.org`	0.70	0.47	0.56	420

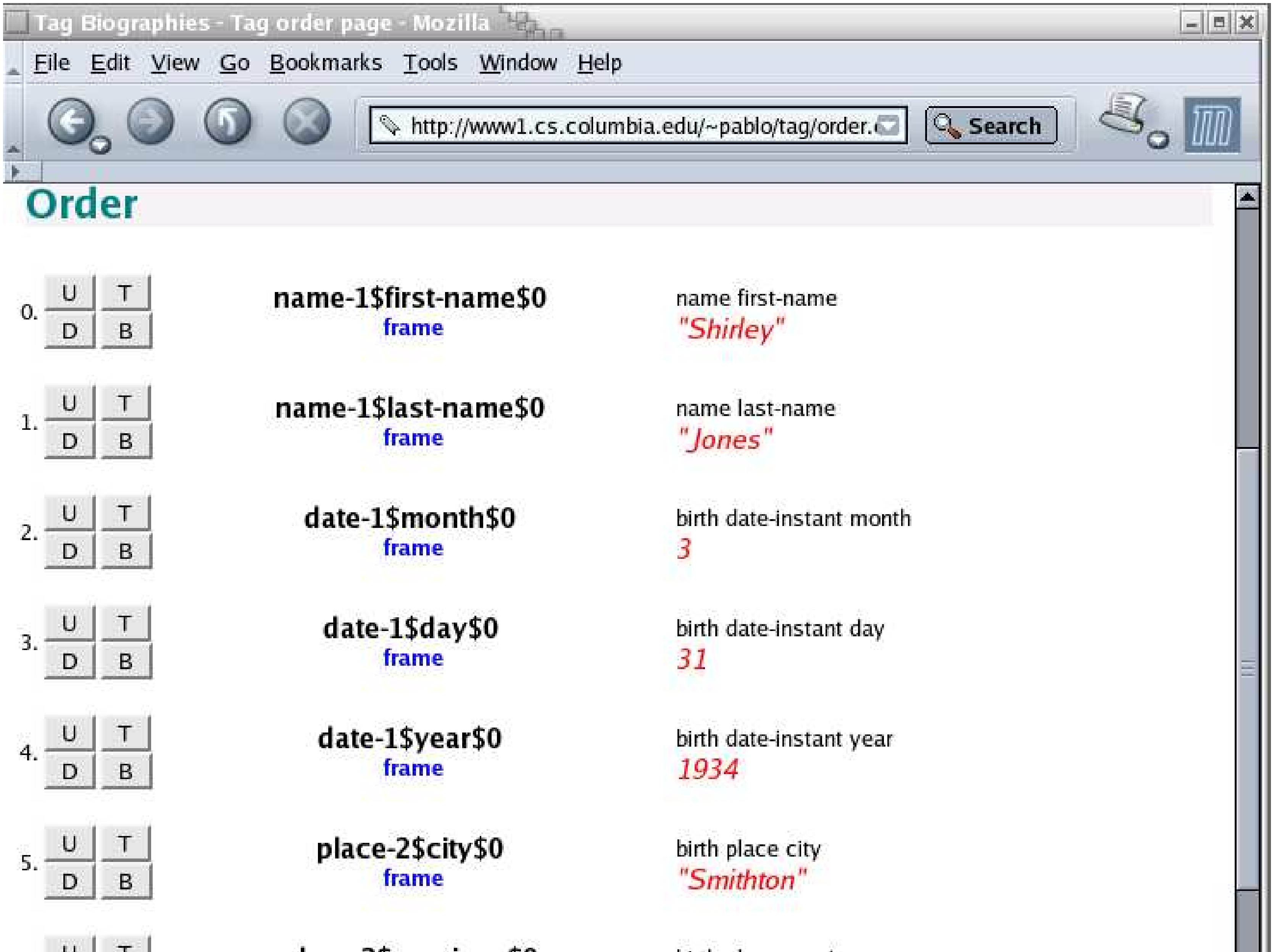
- Error Analysis
  - Had to select 334, system selected 297 with 111 misses.
  - `claimtofame canned-text`, 11 misses out of 11.
  - `education #TYPE`, 6 misses out of 6.
  - `occupation #TYPE`, 16 misses out of 16.
  - `significant-other #TYPE`, 15 misses out of 15.
  - `relative #TYPE`, 17 misses out of 17.
  - `relative relative name last`, 9 misses out of 11.
  - (covers 66% of all errors)

(B) Variant 4

Corpus	Prec.	Rec.	F*	selected
`biography.com`	0.75	0.67	0.71	300
`s9.com`	0.52	0.55	0.54	181
`imdb.com`	0.68	0.59	0.59	284
`wikipedia.org`	0.65	0.52	0.57	481

- Error Analysis, compared to Baseline Variant
  - `occupation #TYPE`, now missed 13 (instead of 16).
    - Dictionary Induction
  - `relative relative name last`, now missed 5 (instead of 9).
    - Disambiguation

(B) Document Structuring results



(B) Document Structuring results

- Kendall's τ

$$\tau = 1 - \frac{2(\text{number of inversions})}{N(N-1)/2}$$

- Results
  - Wikipedia corpus, average sequence length in test set is 29.80 ± 10.86.

	Recall	τ
Baseline	0.47	0.94 ± 0.10
Variant 4	0.52 ± 11.43	0.89 ± 0.12

(B) Learned Words Examples

$(\langle$**`birth date month`**$\rangle, 3)$ *March*.

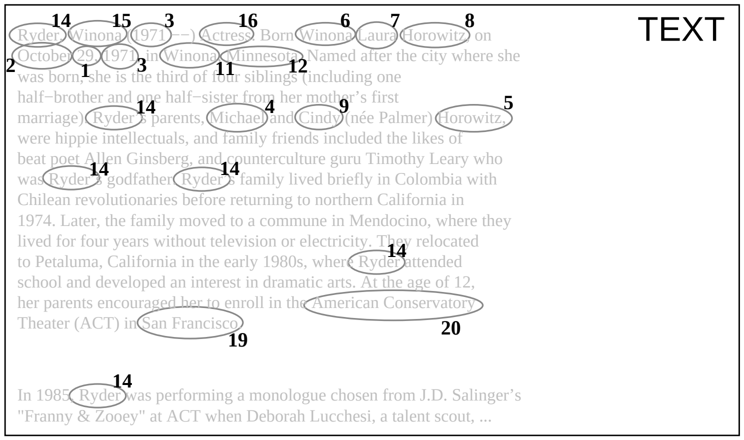
$(\langle$**`birth date day`**$\rangle, 17)$ *17*.

$(\langle$**`birth place country`**$\rangle, \mathit{England})$ *England, Britain, UK, British*.

$(\langle$**`significant-other #TYPE`**$\rangle, \mathit{c\text{-}fiancee})$ *dated, engaged*.

$(\langle$**`occupation #TYPE`**$\rangle, \mathit{c\text{-}job\text{-}comedian})$ *comic, stand, Comedian, Comedy, comedian, comedy, comedic, Comedians*.

(B) Semantic Sequence Example



KNOWLEDGE

1	<birth date day>	29
2	<birth date month>	10
3	<birth date year>	1971
4	<birth father name first>	Michael
5	<birth father name last>	Horowitz
6	<birth name first>	Winona
7	<birth name givenname>	Laura
8	<birth name last>	Horowitz
9	<birth mother name first>	Cindy
10	<birth mother name last>	Horowitz
11	<birth place city>	Winona
12	<birth place province>	MN
13	<birth place country>	USA
14	<name last>	Ryder
15	<name first>	Winona
16	<occupation>	c–actress
17	<occupation>	c–model
18	<relative relative name first>	Michael,Cindy
19	<education place city>	San Francisco
20	<education teaching–agent>	American Conservatory Theater
21	<significant–other name first>	David

Semantic Sequence:

⟨name last⟩	name-1$last
⟨name first⟩	name-1$first
⟨birth date year⟩	date-22$year
⟨occupation⟩	occupation-2$TYPE
⟨birth name first⟩	name-2$first
⟨birth name givenname⟩	name-2$givenname$0
⟨birth name last⟩	name-2$last
⟨birth date month⟩	date-22$month
⟨birth date day⟩	date-22$day
⟨birth place city⟩	place-1$city
⟨birth place province⟩	place-1$province
⟨birth father name first⟩	name-15$first
⟨birth mother name first⟩	name-17$first
⟨birth father name last⟩	name-15$last
⟨education teaching-agent⟩	name-20$name
⟨education place city⟩	place-7$city

...

(B) Indirect Supervised Learning (ISL) Conclusions

- ISL is a feasible way to perform supervised learning without hand-tagging.
- Its unsupervised nature makes for quite some level of noise.
- More research can focus on improving the matching model.
- The text part of the Text-Knowledge corpus is normally very small, but step-wise construction of the matched text helps to remedy the lack of data.

Content Selection

- CS is labelling atomic pieces of knowledge
  - Labelling with two labels, select (sel) or omit ($\neg sel$).

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

Always include $\langle$`name` $\rightarrow$ `first`$\rangle$.

(C) Problem Revisited (Content Selection)

- CS is labelling atomic pieces of knowledge
  - Labelling with two labels, select (sel) or omit ($\neg sel$).

$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

Include only if $\langle$`award` $\rightarrow$ `name`$\rangle \in \{\text{"Oscar"}\}$.

(C) Problem Revisited (Content Selection)

- CS is labelling atomic pieces of knowledge
  - Labelling with two labels, select (sel) or omit ($\neg sel$).

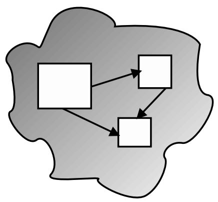
$\langle$`name` $\rightarrow$ `first`$\rangle$	"John"
$\langle$`weight`$\rangle$	150Kg
$\langle$`award` $\rightarrow$ `name`$\rangle$	"Oscar"
$\langle$`award` $\rightarrow$ `name`$\rangle$	"MTV"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-son
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Steve"
$\langle$`relative` $\rightarrow$ `type`$\rangle$	c-step-cousin
$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$	"Martin"

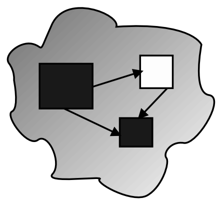
Include only if $\langle$`relative` $\rightarrow$ `type`$\rangle \in \{\text{c-son}\}$.

Include only if $\langle$`relative` $\rightarrow$ `name` $\leftarrow$ `type`$\rangle \in \{\text{c-son}\}$.

(C) Approach: Content Selection Rules

- Learning from
  - Input
  - Output

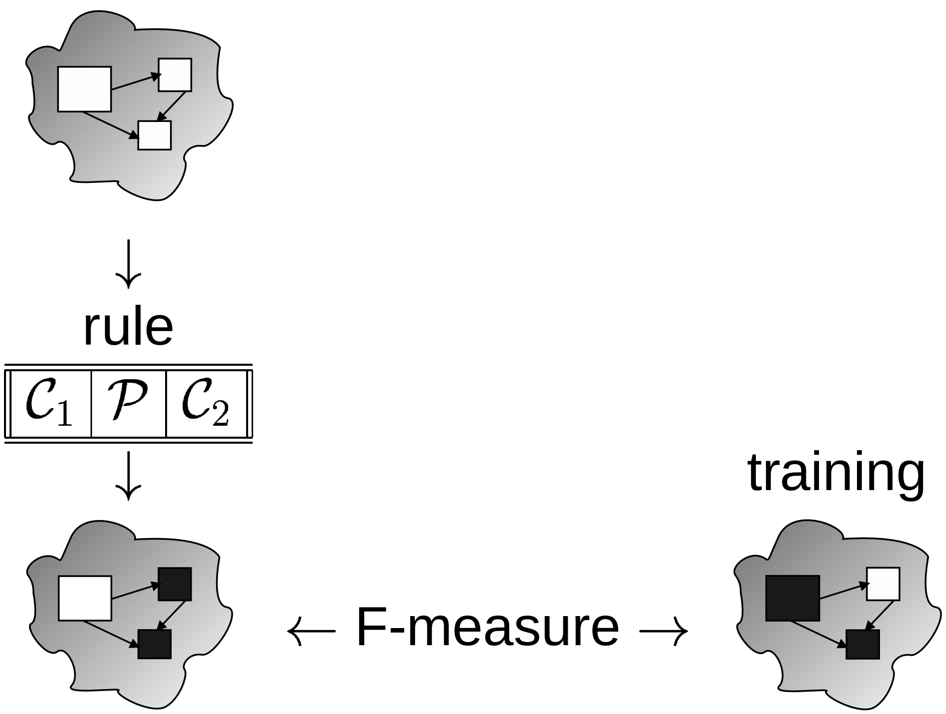




- Representation (rules)

$\mathcal{C}_1$: constraints in node	$\mathcal{P}$: path to other node	$\mathcal{C}_2$: constraints in other node

(C) Approach: Content Selection Rules



- Each rule is executed and its output compared to the automatically obtained reference

(C) Approach: Content Selection Rules

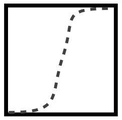
I use the weighted F-measure over the labels as fitness:

$$Fitness = F^*_\alpha + \textsc{mdl}$$

where

$$F^*_\alpha \;=\; \frac{\left(\alpha^2+1\right) Prec\; Rec}{\alpha^2 Prec + Rec}$$

$$\textsc{mdl} \;=\; \text{a minimum description length term}$$



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

(C) Details CS Rules

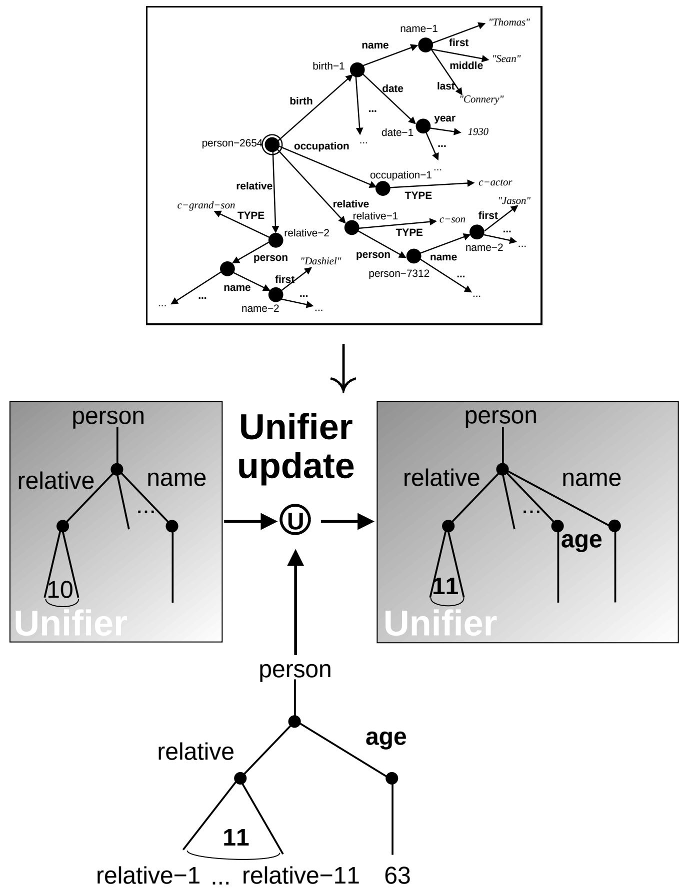
- Combining two rules

$$
\begin{array}{c}
\begin{array}{||c|c|c||}
\hline
\mathcal{C}_1 & \mathcal{P} & \mathcal{C}_2 \\
\hline
\end{array} \\
\updownarrow \qquad\qquad\qquad \updownarrow \\
\begin{array}{||c|c|c||}
\hline
\mathcal{C}_1' & \mathcal{P} & \mathcal{C}_2' \\
\hline
\end{array} \\
\hline
\downarrow \qquad\qquad\qquad \downarrow \\
\begin{array}{||c|c|c||}
\hline
c\left(\mathcal{C}_1, \mathcal{C}_1'\right) & \mathcal{P} & c\left(\mathcal{C}_2, \mathcal{C}_2'\right) \\
\hline
\end{array}
\end{array}
$$

- The new rule share some of the constraints of its parents

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

(C) Approach: Machine Learning



$\neg sel$ ⟨ August, .,
., c-wife, .,
., 1, ., ., ., ., ., ., ., .,
., 1, ., ., ., ., ., ., ., ., ., ., ., ., ., ., ., ., ., 1, ., ., ., ., ., 1, ., ., ., ., ., ., ., .,
., 1, ., ., ., ., ., ., 1, ., ., ., ., ., ., ., ., ., ., ., ., ., ., .,
., ., ., ., Colleen, ., ., ., ., ., ., ., ., ., ., ., ., ., ., ., ., Dewhurst, .,
., ., ., ., ., ., ., . ⟩

(C) SELECT-ALL/SELECT-NONE rules

- Simpler rules
  - Compute the F^* of selecting all elements in a data-path.
  - If the F^* is greater than 0.5, SELECT-ALL.
  - Otherwise, SELECT-NONE.
- Advantages
  - Trivially fast to learn.
  - Generalize well.
  - Very robust to noise.
- Disadvantages
  - Low accuracy.

(C) Experiment: CS Rules Overall Evaluation

- SELECT-ALL/SELECT-NONE Rules

Corpus	Variant 0 P	Variant 0 R	Variant 0 F^*	Variant 0 sel	Variant 4 P	Variant 4 R	Variant 4 F^*	Variant 4 sel
`biography.com`	0.60	0.61	0.61	36	0.58	0.66	0.62	55
`s9.com`	0.35	0.46	0.40	11	0.50	0.48	0.49	18
`imdb.com`	0.58	0.32	0.41	22	0.53	0.37	0.44	39
`wikipedia.org`	0.85	0.18	0.30	10	0.59	0.29	0.39	33

- Tri-partite (CS) Rules.

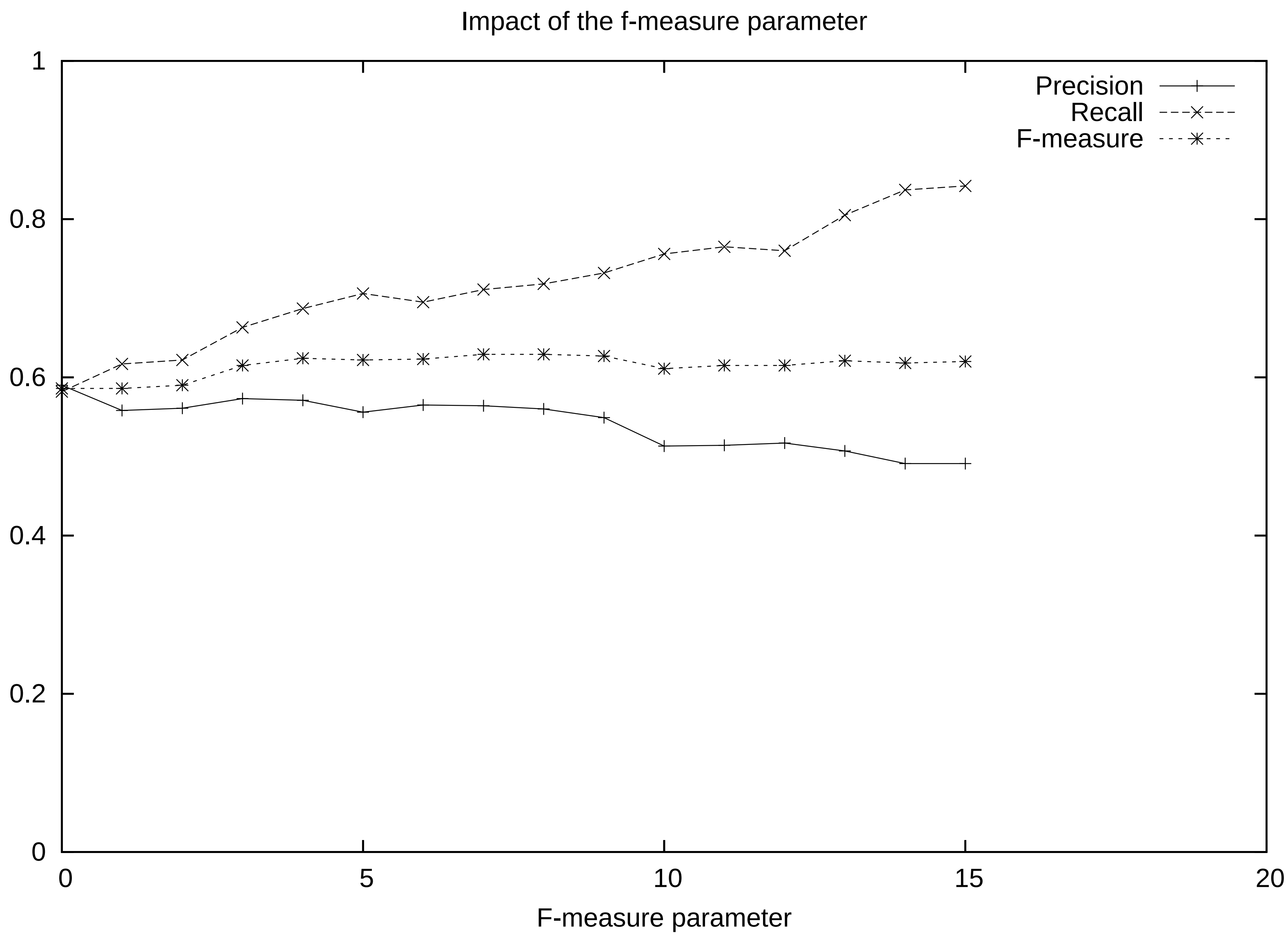
Corpus	P	R	F^*	sel
`biography.com`	0.58	0.72	0.64	410
`s9.com`	0.34	0.49	0.40	248
`imdb.com`	0.50	0.46	0.48	338
`wikipedia.org`	0.52	0.37	0.43	433

(C) Experiment: Machine Learning

- Only performed in `biography.com`

Metric	Prec.	Rec.	F^*
j48 (C4.5)	0.68	0.49	0.57
Naïve Bayes	0.62	0.49	0.55
SMO (SVM)	0.61	0.50	0.55
Logistic	0.62	0.57	0.59
Baseline	0.60	0.61	0.61
SELECT-ALL/SELECT-NONE	0.58	0.66	0.62
CS Rules	0.58	0.72	0.64

(C) Experiment: Prec. and Rec. at different values of α



(C) Experiment: Cross-Corpora application of the rules

Tested on	Trained on `biography.com`	Trained on `s9.com`	Trained on `imdb.com`
`biography.com`	$\frac{P{:}0.58}{R{:}0.72}$ F^*: **0.64**	$\frac{P{:}0.17}{R{:}0.79}$ F^*: 0.28	$\frac{P{:}0.40}{R{:}0.67}$ F^*: 0.50
`s9.com`	$\frac{P{:}0.66}{R{:}0.35}$ F^*: **0.46**	$\frac{P{:}0.34}{R{:}0.49}$ F^*: 0.40	$\frac{P{:}0.46}{R{:}0.25}$ F^*: 0.32
`imdb.com`	$\frac{P{:}0.56}{R{:}0.37}$ F^*: 0.44	$\frac{P{:}0.23}{R{:}0.59}$ F^*: 0.33	$\frac{P{:}0.50}{R{:}0.46}$ F^*: **0.48**

(C) Example Rules

$\langle$`person` $\rightarrow$ `name` $\rightarrow$ `first`$\rangle$:

(-, -, -). ;TRUE

Always say the first name of the person being described.

$\langle$`education` $\rightarrow$ `place` $\rightarrow$ `country` $\rightarrow$ `name` $\rightarrow$ `last`$\rangle$:

($\mathrm{value} \in \{$*"Scotland"*, *"England"*$\}$,-,-).

As I used U.S. biographies, the country of education is only mentioned when it is abroad.

$\langle$`significant-other` $\rightarrow$ `#TYPE`$\rangle$:

($\mathrm{value} \in \{$`c-husband`, `c-wife`$\}$,-,-).

Mention husband and wives (but not necessarily boyfriends, girlfriends or lovers).

$\langle$`relative` $\rightarrow$ `name` $\rightarrow$ `last`$\rangle$:

(-,$\langle$`-last -name -relative #TYPE`$\rangle$, $\mathrm{value} \in \{$`c-father`$\}$).

Only mention the last name of the father of the person.

(C) Content Selection Conclusions

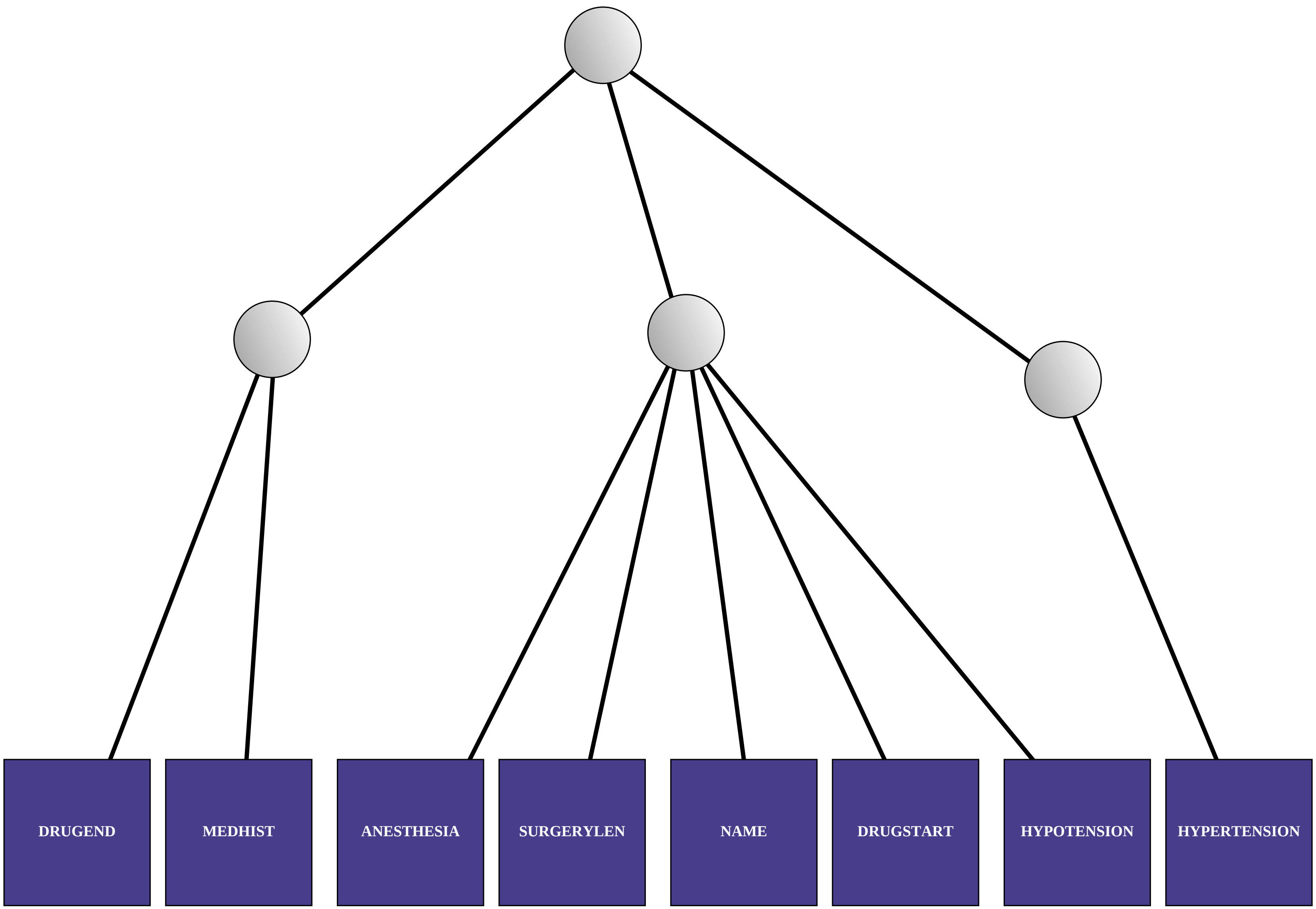
- Proposed, implemented and evaluated three learning methodologies
  - SELECT-ALL/SELECT-NONE rules.
  - Tri-partite rules.
  - Classification systems (traditional Machine Learning).
- Each methodology has its own strengths and weaknesses.
  - SELECT-ALL/SELECT-NONE rules: more robust.
  - Tri-partite rules: best compromise.
  - Classification systems: more precise.

Document Structuring

(D) Problem Revisited (Document Structuring)

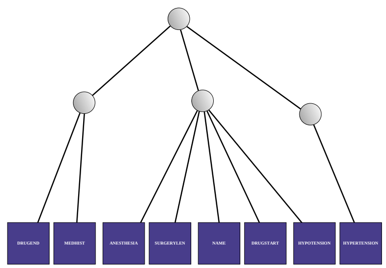
- Document Structuring.
  - Input: knowledge to be structured
  - It uses communicative predicates to produce messages.
  - Output: document plan (sequence of messages).
- Learning Document Structuring schema.
  - From sequences of atomic values.
  - Problem: sequence of atomic values is not a sequence of messages.
- Two Domains.
  - Medical: simpler, fewer data, simpler schema.
  - Biographical: more complex, more data, full-fledged schema.

(D) MAGIC representation



(D) MAGIC representation

planner

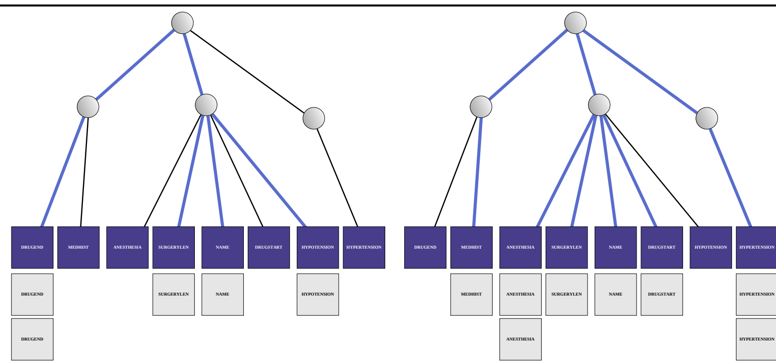


Semantic input sets **(unordered)**

patient A drugend-1,drugend-2,hypotension-1,name-1,surgerylen-1

patient B anesthesia-1,anesthesia-2,drugstart-1,hypertension-1,hypertension-2,medhist-1,name-1surgerylen-1

planning



Plans **(ordered)**

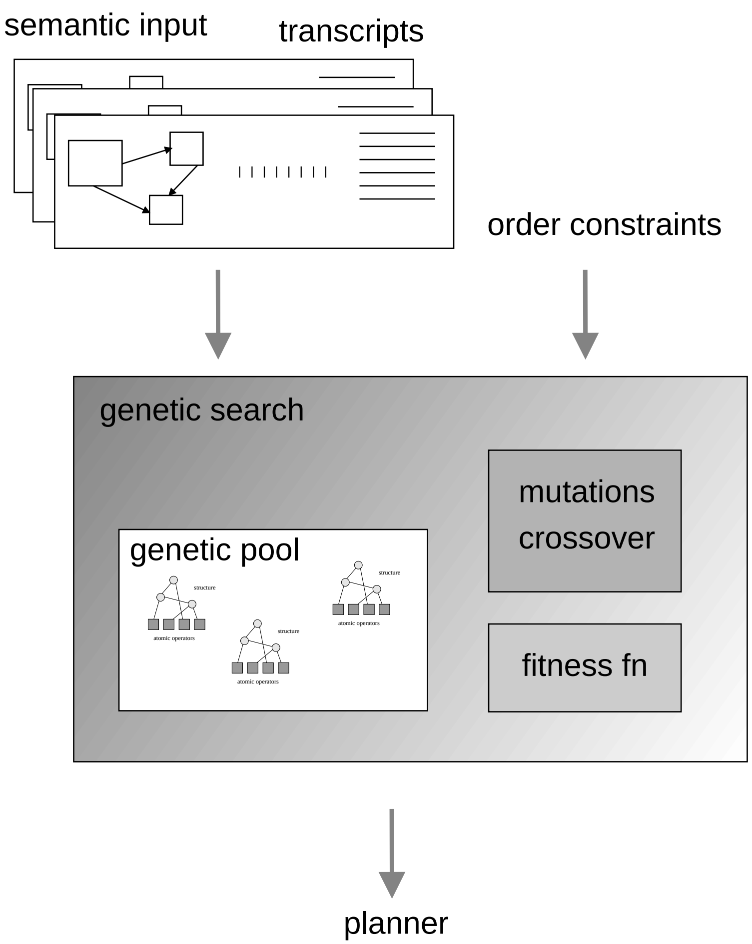
plan for A drugend-1,drugend-2,surgerylen-1,name-1,hypotension-1

plan for B medhist-1,anesthesia-1,anesthesia-2,surgerylen-1,name-1,drugstart-1,hypertension-1,hypertension-2

(D) MAGIC data

- From a past evaluation (McKeown et al., 2000)
  - Annotated Transcriptions of Physicians Briefings
- Semantic Annotation
  - Assisted by a Domain Expert
  - Semantically Tagged Non-overlapping Chunks (Clause Level)
  - Tag-set
    - Over 200 tags
    - 29 categories
- Expensive Task
  - Intensive Care Unit, a Busy Environment
  - Total Number: 24 Transcripts
  - Average Length: 33 tags ($\min = 13$, $\max = 66$, $\sigma = 11.6$)

(D) MAGIC approach



- Order Constraints

Sequences

$\Downarrow$

Pattern Detection

$\Downarrow$

Patterns

$\Downarrow$

Clustering

$\Downarrow$

Cluster of Patterns

$\Downarrow$

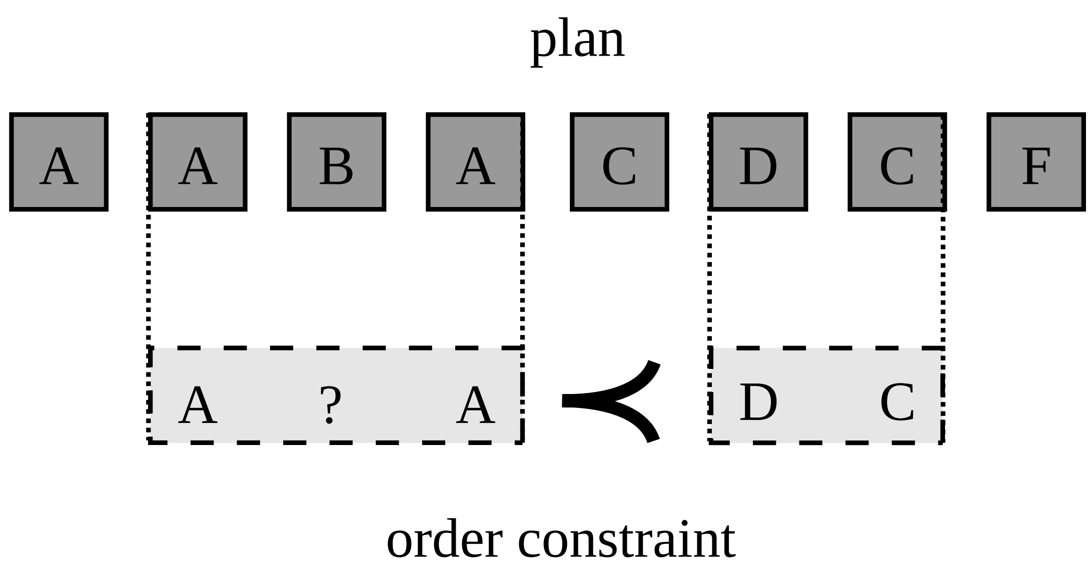
Constraints Inference

$\Downarrow$

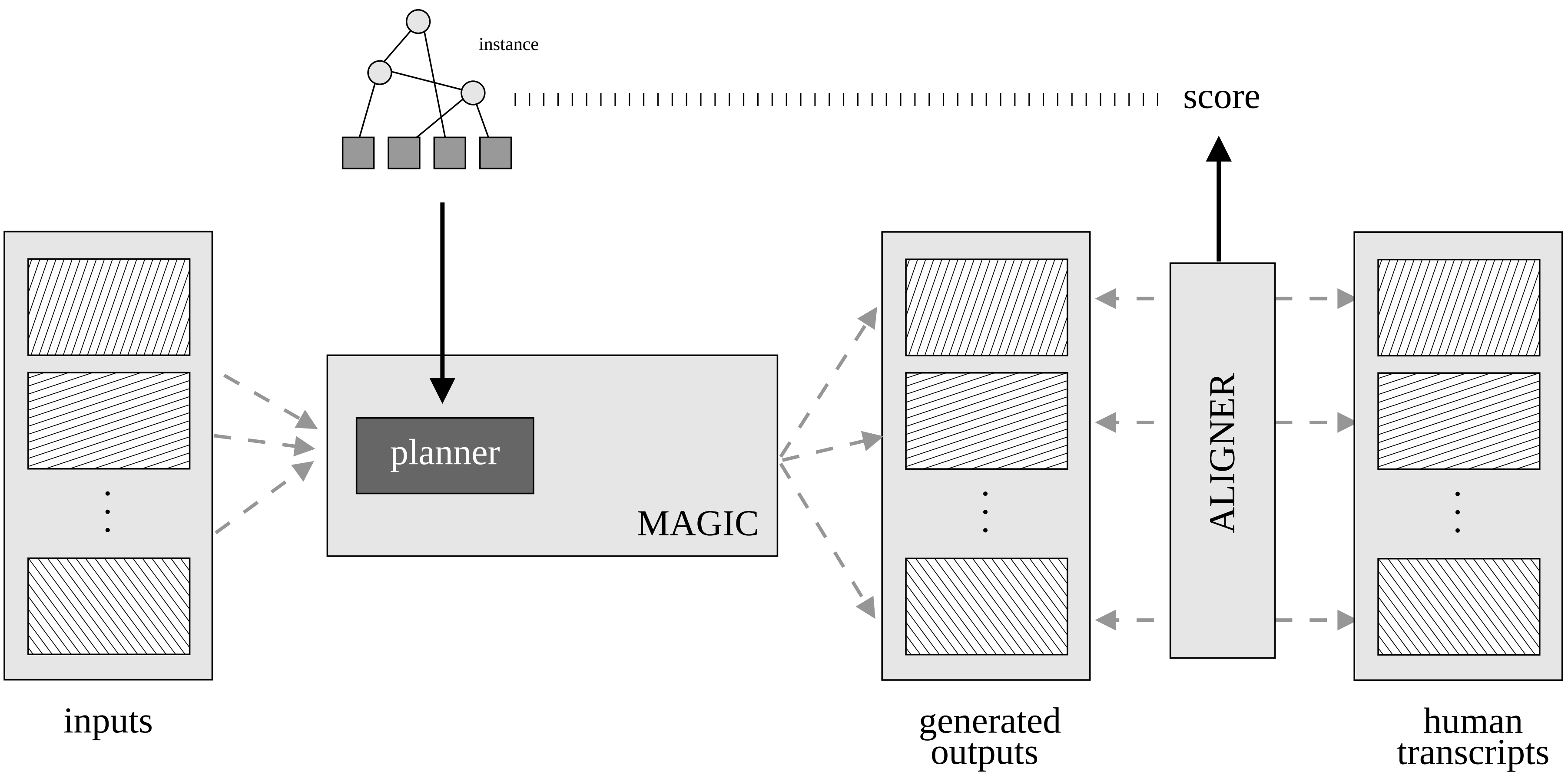
Order Constraints over Clusters

(D) MAGIC Approach

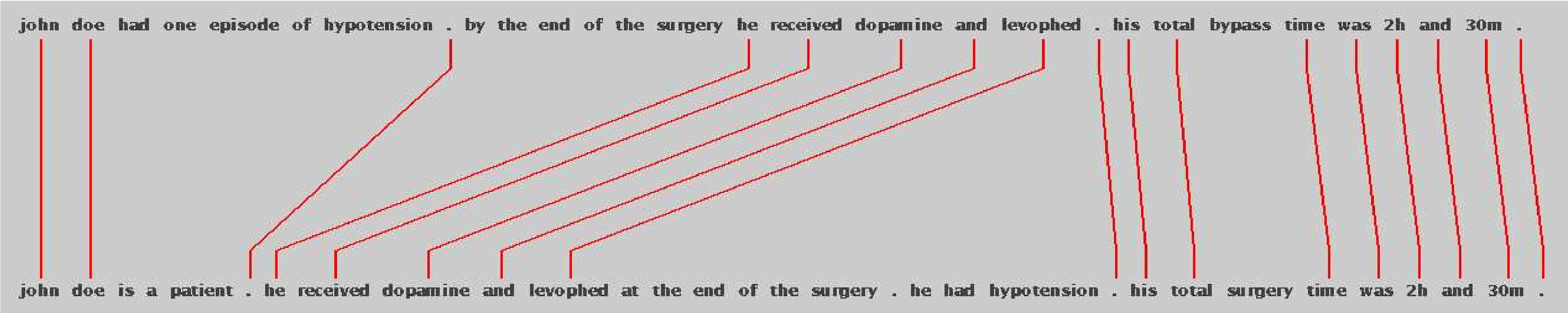
- F_O function works as follows:
  - given a set of semantic inputs;
  - the chromosome is used to generate corresponding plans;
  - then order constraints are checked for validity.



(D) MAGIC Approach



- Alignment



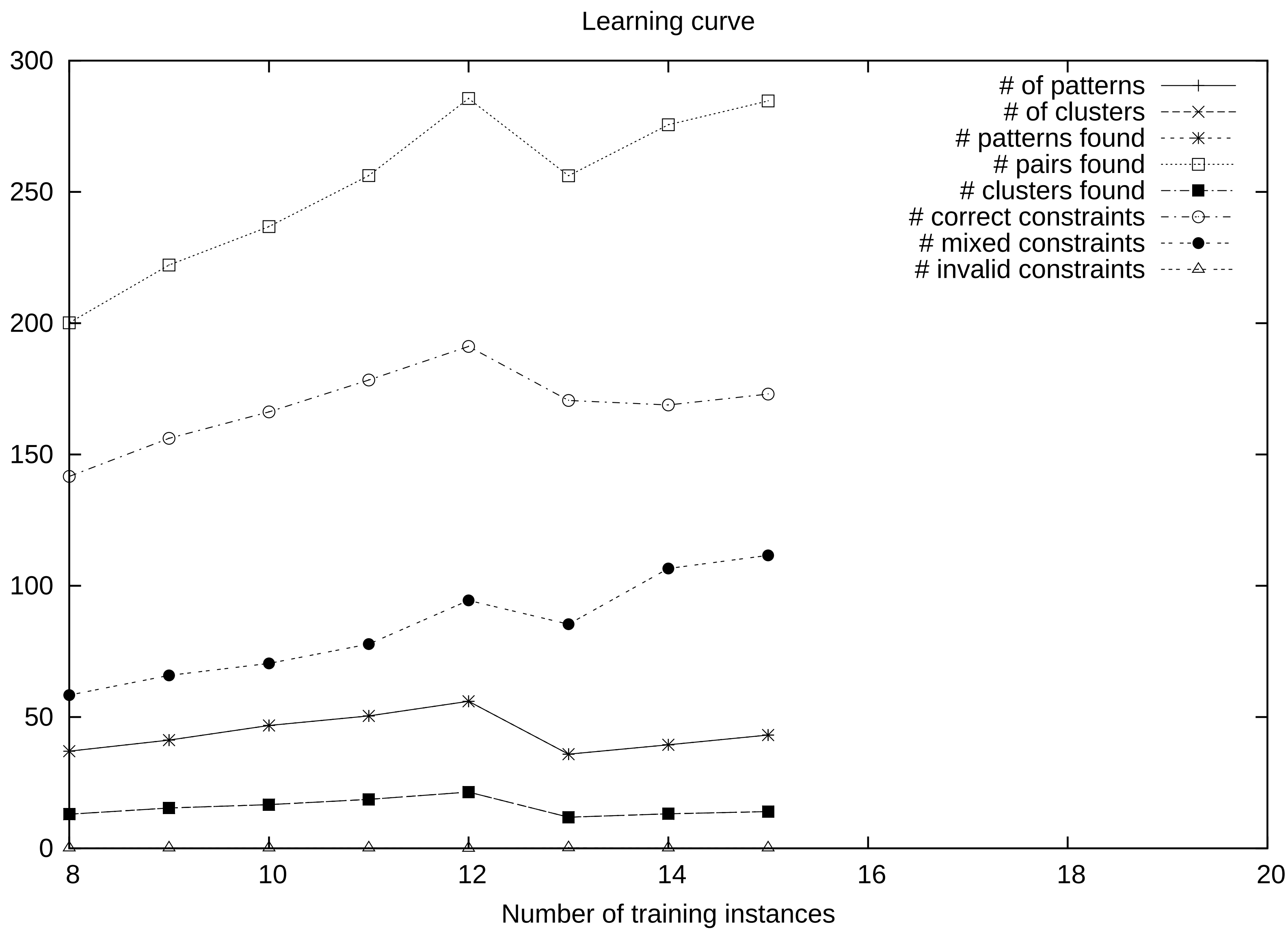
- This alignment produce an score that is then averaged over the different patients.

(D) MAGIC Order Constraints Results

- Obtained an average of 58.54 (± 8.46) patterns, clustered into 19.71 (± 3.02) clusters.

- An average of 401.94 (± 51.23) constraints are found from which
  - 205.21 ± 45.95 (a 51.90%) are always correct,
  - 196.61 ± 68.13, (a 48.07%) sometimes contain errors,
  - and 0.14 ± 0.35, (a 0.04%) contains a large number of errors.

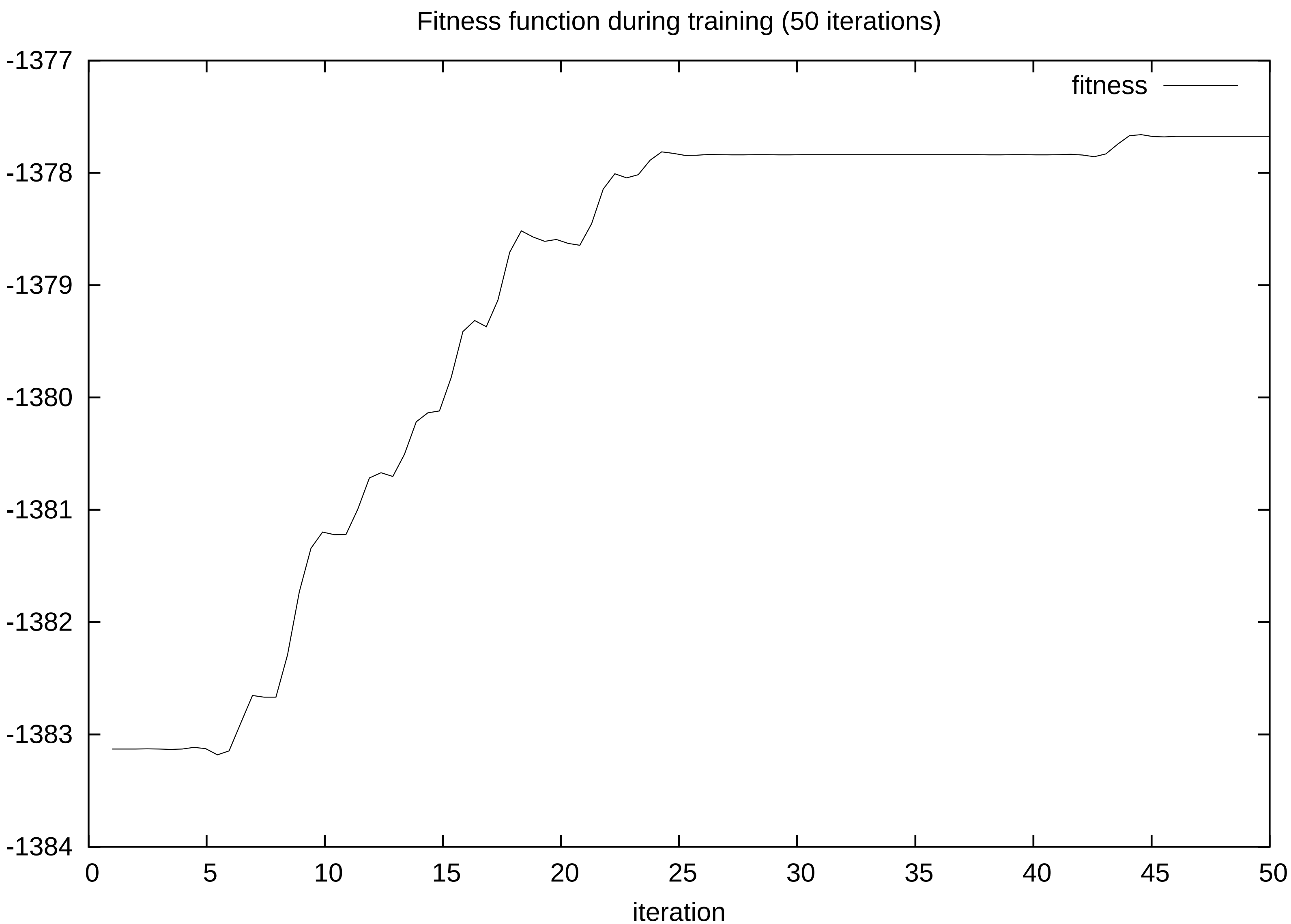
(D) MAGIC Order Constraints Results



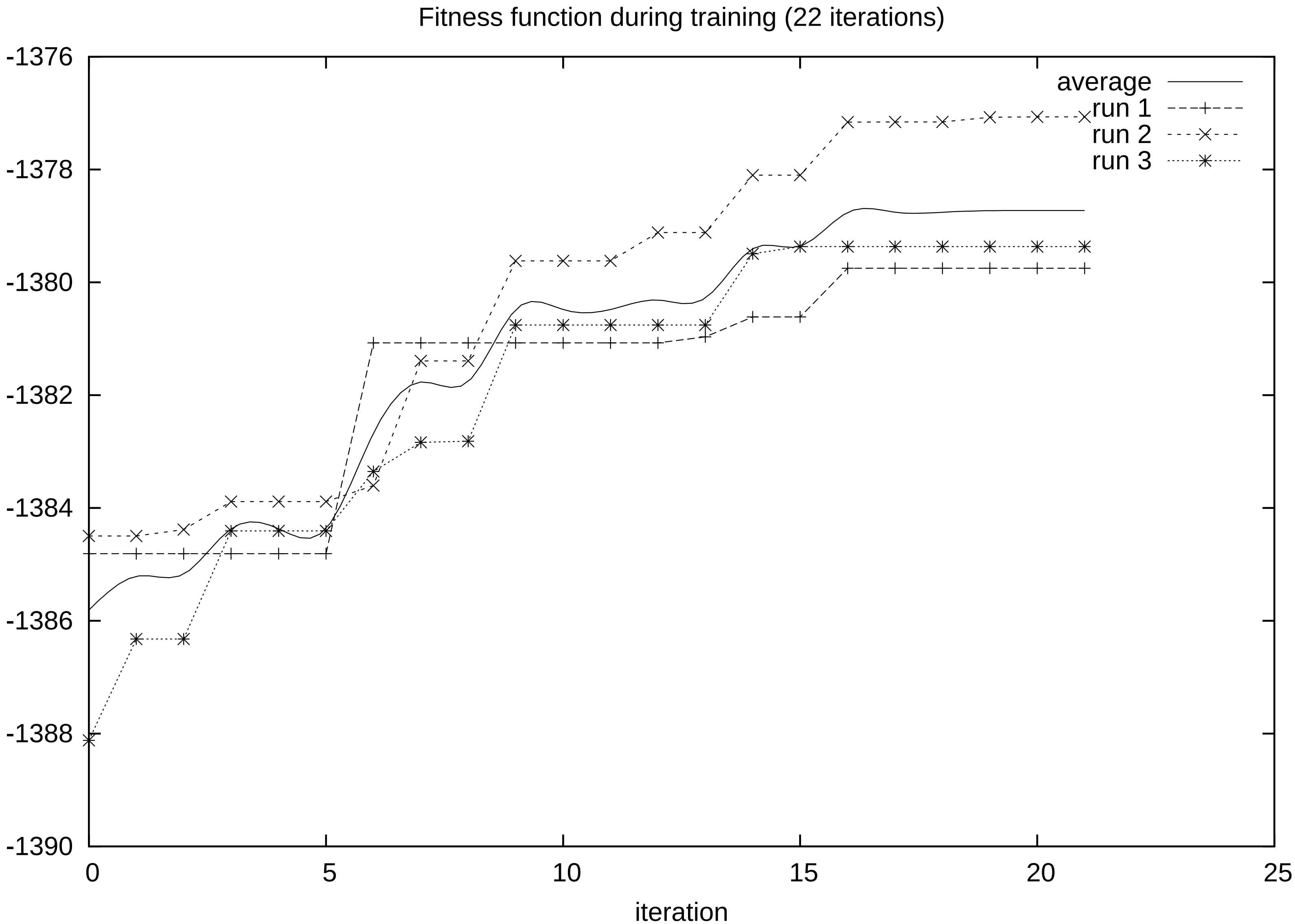
(D) MAGIC Document Structuring Results

- Baseline
  - The first generation of three runs (a total of 6,000 random instances).
  - Scored using Kendall's τ against the MAGIC planner, they had an average τ of 0.0952 ± 0.1144.
- Learned Planners
  - The best instance for each run at each iteration step is is scored against the sequence obtained from the MAGIC planner.
  - The average over the three runs gave τ of 0.2288 ± 0.0342.

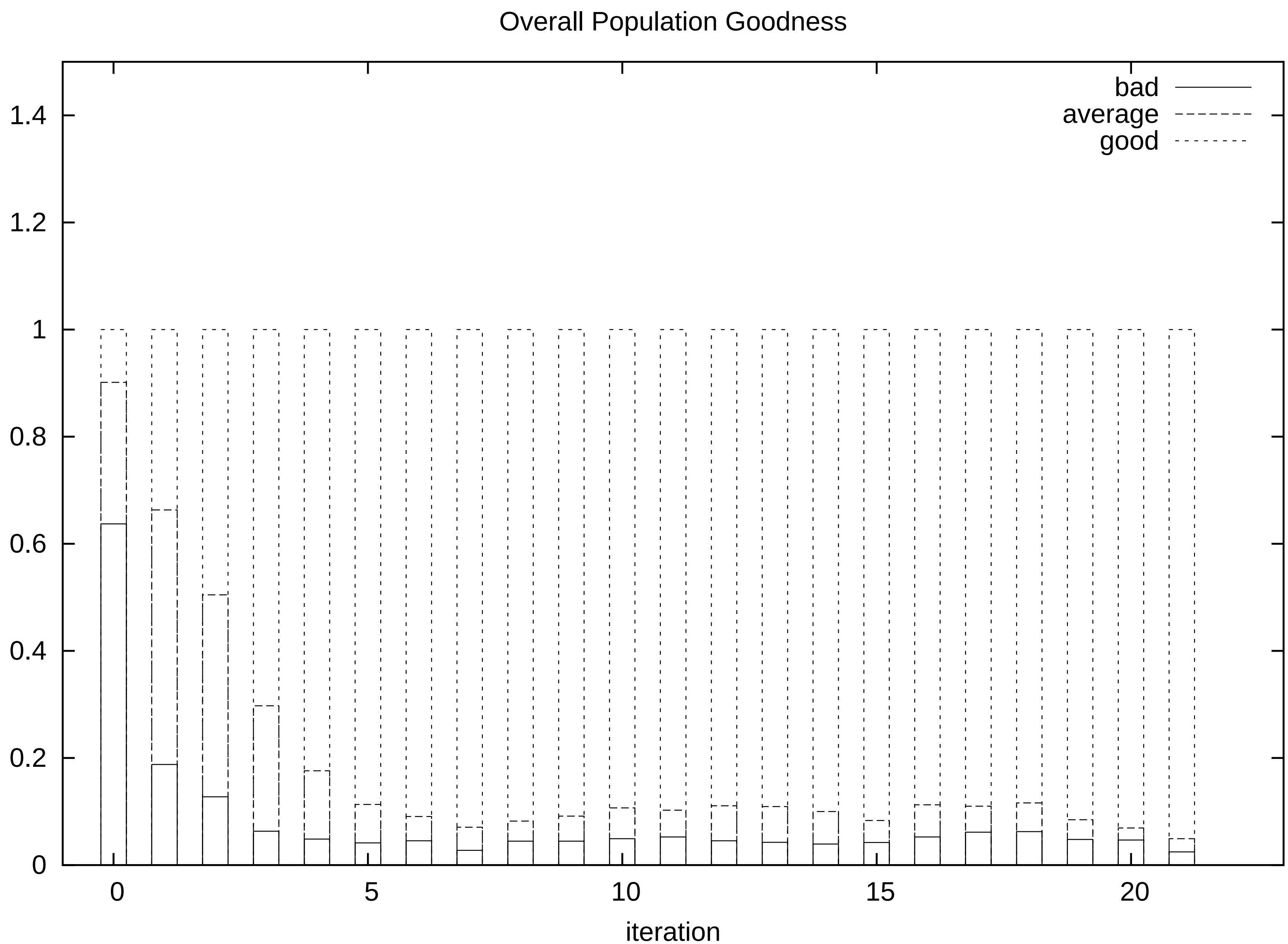
(D) MAGIC Document Structuring Results



(D) MAGIC Document Structuring Results



(D) MAGIC Document Structuring Results



(D) ProGenIE Data

- `wikipedia.org`

	Total	Average	Train	Test
# pairs	361	-	341	20
# frames	58,387	161.737	55,326	3,061
# triples	108,009	299.194	102,297	5,712
# words	68,953	191.006 $\pm$ 55.17	64,784	4,169
# chars	418,035	1,157.992 $\pm$ 334.01	392,925	25,110

- Orderings Quality

avg. length	τ
26.35 $\pm$ 11.4260	0.8909 $\pm$ 0.1154

(D) ProGenIE Approach

- Three Tiers:
  1. Content Selection
  2. Order Constraints
  3. Alignments

- Alignments
  - Comparing sequences of atomic values to sequences of messages (sequences of sets of atomic values).

$$T(i,j) = max \left\{ \begin{array}{ll} T(i-1,j) \text{ if } T(i-1,j) \text{was a mismatch} & \textit{(skip)} \\ T(i-1,j-1) & \textit{(match)} \\ T(i,j-1) & \textit{(stay)} \end{array} \right\} + c(i,j)$$

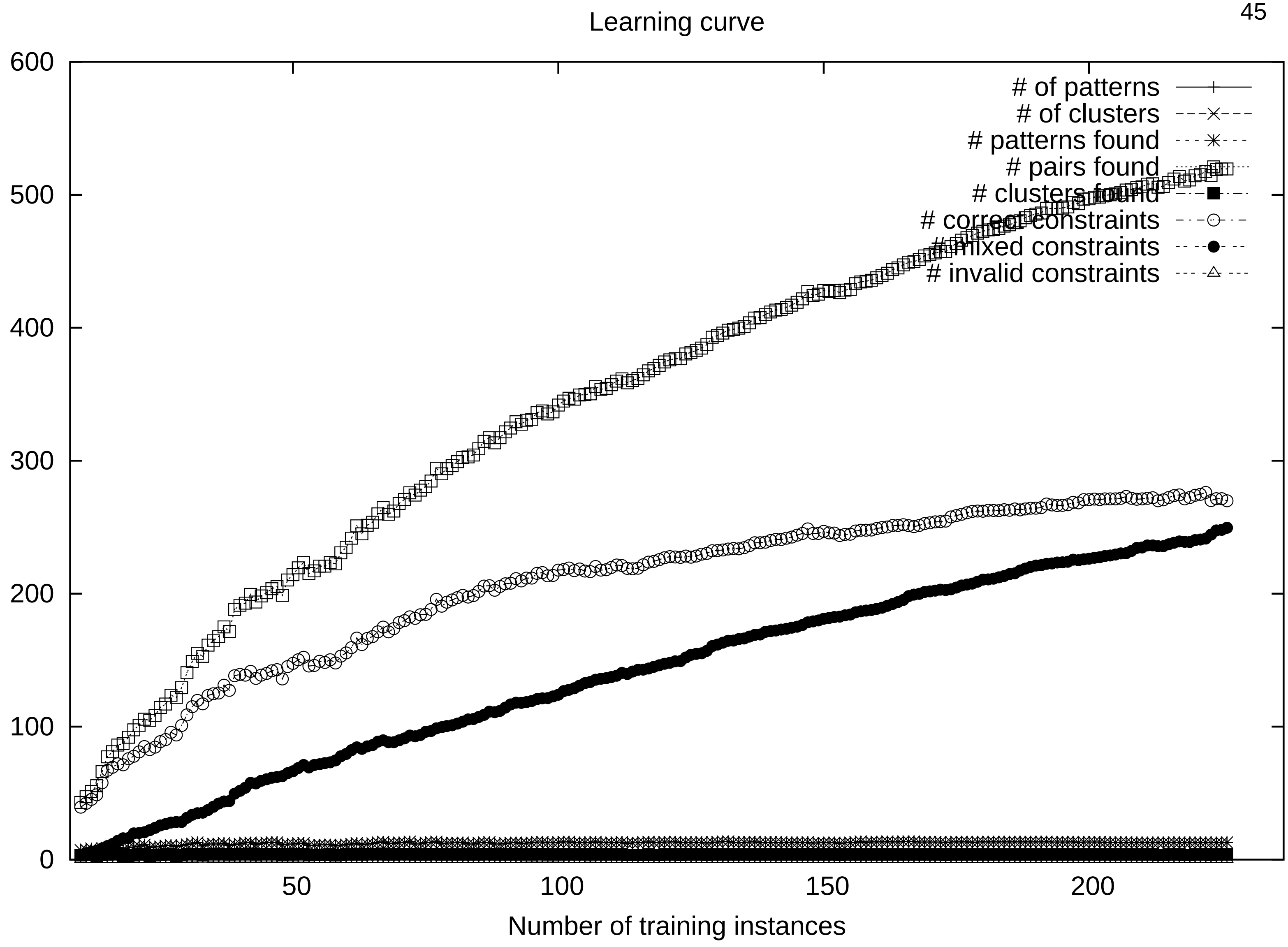
$$c(i,j) = \left\{ \begin{array}{ll} 1 & \text{if } v \in s \\ \text{-1} & v \notin s \end{array} \right.$$

(D) ProGenIE Order Constraints Results

- Obtained an average of 14.17 (± 1.81) patterns, clustered into 3.84 (± 0.38) clusters.

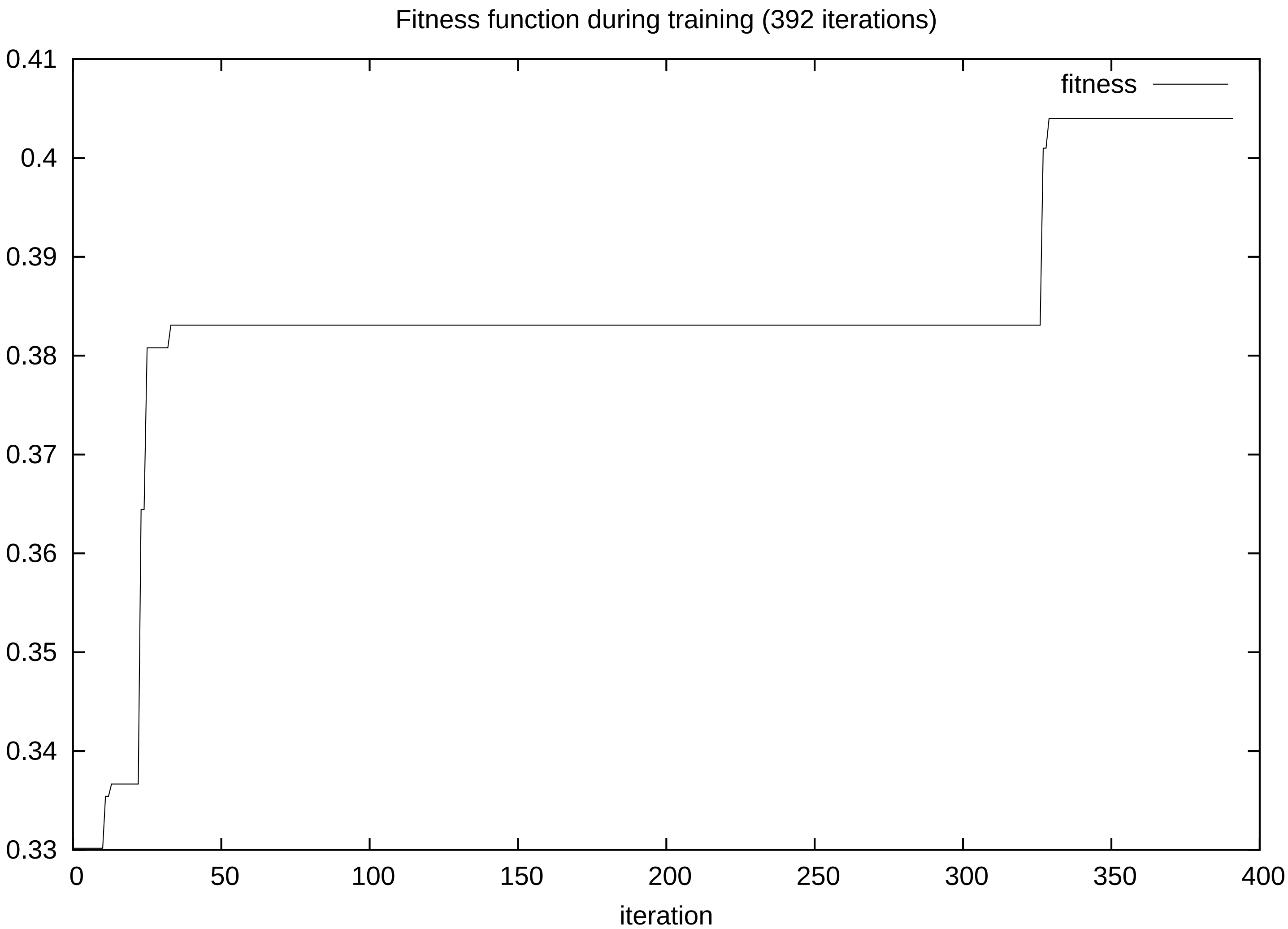
- An average of 537.04 (± 18.43) constraints are found from which
  - 276.64 ± 15.50 (a 51.50%) are always correct,
  - 260.41 ± 12.38, (a 48.51%) sometimes contain errors.

(D) ProGenIE Order Constraints Results



(D) ProGenIE Problem

- Search does not make progress.



(D) Document Structuring Conclusions

- Proposed fitness function
  - Corpus-based.
  - Allows for learning in simpler domain.
- Search process
  - Good for simpler domains.
  - Progresses too slow in more complex domains.
  - Need corpus-based search operators.

(D) Contributions

- Indirect Supervised Learning contributions
  - Devising, implementing and testing a system for the automatic construction of training material for learning CS and DS logic.
- Content Selection contributions
  - The proposal and study of techniques to learn CS logic from a traning material consisting of structured knowledge and selection labels.
- Document Structuring contributions
  - Defined the problem of learning DS schemata from indirect observations, proposing, implementing and evaluating two different, yet similar techniques in two different domains.

(D) Limitations and Further Work

- General Limitations
  - Text-Knowledge corpus requirement.
  - + *Use a small knowledge set to bootstrap the whole process.*
- Limitations of the *matched text* construction process
  - Model limitations.
  - + *Improve the model using EM.*
- Limitations of the learning of Content Selection rules
  - Captures only paradigmatic information.
  - + *Complement the approach with summarization techniques.*
- Limitations of the learning of Document Structuring schemata
  - Requires communicative predicates.
  - + *Learn statistical predicates for a fully statistical system.*

(D) Thesis Conclusions

“The main effort in porting a generator to a new domain is in the adaptation of its discourse planning component.”

(Bontcheva and Wilks, 2004)