Thesis Defense

Indirect Supervised Learning of Strategic Generation Logic

Pablo Ariel Duboue

Committee
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
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.
(A) Content Selection

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

<table>
<thead>
<tr>
<th>Key</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>award → name</td>
<td>“Oscar”</td>
</tr>
<tr>
<td>award → name</td>
<td>“MTV”</td>
</tr>
<tr>
<td>relative → type</td>
<td>c-son</td>
</tr>
<tr>
<td>relative → name → first</td>
<td>“Steve”</td>
</tr>
<tr>
<td>relative → type</td>
<td>c-step-cousin</td>
</tr>
<tr>
<td>relative → name → first</td>
<td>“Martin”</td>
</tr>
</tbody>
</table>
(A) Content Selection

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

| {name → first} | “John” |
| {weight}       | 150Kg |
| {award → name} | “Oscar” |
| {award → name} | “MTV”  |
| {relative → type} | c-son |
| {relative → name → first} | “Steve” |
| {relative → type} | c-step-cousin |
| {relative → name → first} | “Martin” |

Always include {name → first}.
(A) Content Selection

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>name → first</strong></td>
<td>“John”</td>
</tr>
<tr>
<td><strong>weight</strong></td>
<td>150Kg</td>
</tr>
<tr>
<td><strong>award → name</strong></td>
<td>“Oscar”</td>
</tr>
<tr>
<td><strong>award → name</strong></td>
<td>“MTV”</td>
</tr>
<tr>
<td><strong>relative → type</strong></td>
<td>c-son</td>
</tr>
<tr>
<td><strong>relative → name → first</strong></td>
<td>“Steve”</td>
</tr>
<tr>
<td><strong>relative → type</strong></td>
<td>c-step-cousin</td>
</tr>
<tr>
<td><strong>relative → name → first</strong></td>
<td>“Martin”</td>
</tr>
</tbody>
</table>

*Never include* ⟨weight⟩ *or* ⟨height⟩.
(A) Content Selection

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

```
(name → first)   "John"
(weight)         150Kg
(award → name)   "Oscar"
(award → name)   "MTV"
(relative → type) c-son
(relative → name → first) "Steve"
(relative → type) c-step-cousin
(relative → name → first) "Martin"
```

Include only if ⟨award → name⟩ ∈ {"Oscar"}.
(A) Content Selection

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

<table>
<thead>
<tr>
<th>Key</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>award → name</td>
<td>“Oscar”</td>
</tr>
<tr>
<td>award → name</td>
<td>“MTV”</td>
</tr>
<tr>
<td>relative → type</td>
<td>c-son</td>
</tr>
<tr>
<td>relative → name → first</td>
<td>“Steve”</td>
</tr>
<tr>
<td>relative → type</td>
<td>c-step-cousin</td>
</tr>
<tr>
<td>relative → name → first</td>
<td>“Martin”</td>
</tr>
</tbody>
</table>

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

```
[ pred    education
  pred_0 person-32
  pred_1 "Columbia University"
  pred_2 "Computer Science"

  mods [ time [ start "1999/8/27" end "2005/1/17" ]
         place "New York, NY"
] ]
```
(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 \equiv person.education

**output**

```
[ pred  education
 pred_0 person
 pred_1 education-event \rightarrow teaching-agent
 pred_2 education-event \rightarrow subject-matter

[ mods
  time [ start education-event \rightarrow date-start
         end education-event \rightarrow date-end
  place education-event \rightarrow place
  reason education-event \rightarrow reason
```
A schema produces a sequence of messages.
Messages are instantiated a predicates.
A schema is a finite state automaton over the language of predicates

\[
\begin{align*}
\text{intro-person}(\text{self}), \\
\text{education}(\text{self}, \text{education})^* , \\
(\text{spouse}(\text{self}, \text{spouse}), \text{intro-person}(\text{spouse}) ; \\
\{ \text{child}(\text{spouse}, \text{self}, \text{child}), \\
\text{intro-person}(\text{child}) \} )^* \\
(\text{movie}(\text{self}, \text{movie}), \text{intro-movie}(\text{movie}) ; \\
\{ \text{award}(\text{movie}, \text{self}, \text{award}), \\
\text{intro-award}(\text{award}, \text{self}) \} )^*
\end{align*}
\]
(A) DS schemata

- A schema produces a sequence of messages.
- Messages are instantiated as 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.
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 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, ...
(A) Optimization Approach

- Given training input $I$ and output $O$ pairs.
- To find the entity $e^* \in$ Schemas or Rules such that
  \[ e^* = \arg\max_e P(e|I, O) \]
- 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.
I use the weighted F-measure over the labels as fitness:

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

where

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

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

This function captures the problem well and allows selecting solutions that prefer precision or recall through the \( \alpha \) parameter.
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.
(A) Contributions and Results Highlights

- **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 $\tau$ 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.
(A) Structure of this Talk

(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)
<table>
<thead>
<tr>
<th>ALIASES</th>
<th>His aliases include Fadel Nazzal al-Khalayleh, Fadil al-Khalaylah, …</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDUCATION</td>
<td>He’s thought to be a high school dropout</td>
</tr>
<tr>
<td>TRANSLATION</td>
<td>Zarqawi went to Afghanistan to fight the Soviets in the late 1980s</td>
</tr>
<tr>
<td>JOIN_ORG</td>
<td>In Afghanistan, Zarqawi plugged into the al Qaeda</td>
</tr>
<tr>
<td>PRISON</td>
<td>In the early 1990s, he was jailed and spent seven years in jail</td>
</tr>
<tr>
<td>BIRTH_DATE</td>
<td>al-Zarqawi (Arabic: ) (possibly born on October 20, 1966)</td>
</tr>
<tr>
<td>BIRTH_COUNTRY</td>
<td>al-Zarqawi (Arabic: ) (possibly born on October 20, 1966) is a shadowy…</td>
</tr>
<tr>
<td>BIRTH_NAME</td>
<td>Ahmad Fadeel al-Nazal al-Khalayleh (Arabic: ), is believed to be his real name</td>
</tr>
<tr>
<td>DESCRIBED</td>
<td>Zarqawi is usually described as somber and unintelligent</td>
</tr>
<tr>
<td>PRISON</td>
<td>in 2001, al-Zarqawi was arrested again in Jordan</td>
</tr>
<tr>
<td>MASTERMIND</td>
<td>On July 11, 2004, Zarqawi claimed responsibility for a July 8 mortar attack in Samarra</td>
</tr>
<tr>
<td>MASTERMIND</td>
<td>Zarqawi has also claimed responsibility for the Canal Hotel bombing of the U.N. …</td>
</tr>
<tr>
<td>BIRTH_DATE</td>
<td>al-Zarqawi (possibly born on October 20, 1966)</td>
</tr>
<tr>
<td>BIRTH_COUNTRY</td>
<td>al-Zarqawi (possibly born on October 20, 1966) is a shadowy Jordanian national</td>
</tr>
<tr>
<td><strong>PRISON</strong></td>
<td>Zarqawi was jailed briefly in the 1980s for sexual assault</td>
</tr>
<tr>
<td><strong>TRASLATION</strong></td>
<td>In 1989, Zarqawi traveled to Afghanistan to fight against the Soviet invasion of ...</td>
</tr>
<tr>
<td><strong>TRASLATION</strong></td>
<td>In the mid-1990s, al-Zarqawi travelled to Europe</td>
</tr>
<tr>
<td><strong>PRISON</strong></td>
<td>he was arrested in Jordan in 1992</td>
</tr>
<tr>
<td><strong>CREATE_ORG</strong></td>
<td>In Afghanistan, al-Zarqawi established a terrorist training camp</td>
</tr>
<tr>
<td><strong>PRISON</strong></td>
<td>in 2001, al-Zarqawi was arrested again in Jordan</td>
</tr>
<tr>
<td><strong>MASTERMIND</strong></td>
<td>On July 11, 2004, Zarqawi claimed responsibility for a July 8 mortar attack in Samarra</td>
</tr>
<tr>
<td><strong>OCCUPATION</strong></td>
<td>al-Zarqawi is a Palestinian jihadi leader</td>
</tr>
<tr>
<td><strong>ALIAS</strong></td>
<td>al-Zarqawi, A.K.A. Fedel Nazzel Khalayleh,</td>
</tr>
<tr>
<td><strong>JOIN_ORG</strong></td>
<td>He is from the Beni Hassan tribe</td>
</tr>
<tr>
<td><strong>MASTERMIND</strong></td>
<td>Zarqawi has been implicated in terrorist activity worldwide</td>
</tr>
<tr>
<td><strong>MASTERMIND</strong></td>
<td>He has also been implicated in a foiled chemical weapons attack against Jordan’s ...</td>
</tr>
<tr>
<td><strong>MASTERMIND</strong></td>
<td>Zarqawi was behind the assassination of US diplomat Lawrence Foley in Amman, ...</td>
</tr>
<tr>
<td><strong>OCCUPATION</strong></td>
<td>Al-Qaeda Zarqawi has been named as the leader of Jund al-Shams</td>
</tr>
</tbody>
</table>
(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

Knowledge

Content Selection

Relevant Knowledge

Text Structuring

Document Plan

Aggregation & Lexicalization

Text
(B) Graphical Model

Knowledge

Relevant Knowledge

Content Selection

Text
(B) Graphical Model

Knowledge

Relevant Knowledge

Content Selection

Text
(B) 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**

  $H_0 : P(p \in \mathcal{P} | c \in \mathcal{C}) = p_0 = P(p \in \mathcal{P})$  if $p \notin \mathcal{V}(c)$

  $H_1 : P(p \in \mathcal{P} | c \in \mathcal{C}) = p_1 \gg p_2 = P(p \in \mathcal{P})$  if $p \in \mathcal{V}(c)$
(B) Example

- **Given:**
  - \((KB_1, Bio_1), (KB_2, Bio_2), (KB_3, Bio_3), (KB_4, Bio_4)\)

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

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

- \(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.
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
## (B) Corpora

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Prec.</th>
<th>Rec.</th>
<th>$F^*$</th>
<th>selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>biography.com</td>
<td>0.74</td>
<td>0.64</td>
<td>0.69</td>
<td>297</td>
</tr>
<tr>
<td>s9.com</td>
<td>0.51</td>
<td>0.53</td>
<td>0.52</td>
<td>184</td>
</tr>
<tr>
<td>imdb.com</td>
<td>0.71</td>
<td>0.53</td>
<td>0.61</td>
<td>295</td>
</tr>
<tr>
<td>wikipedia.org</td>
<td>0.70</td>
<td>0.47</td>
<td>0.56</td>
<td>420</td>
</tr>
</tbody>
</table>

- **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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F*</th>
<th>selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>biography.com</td>
<td>0.75</td>
<td>0.67</td>
<td>0.71</td>
<td>300</td>
</tr>
<tr>
<td>s9.com</td>
<td>0.52</td>
<td>0.55</td>
<td>0.54</td>
<td>181</td>
</tr>
<tr>
<td>imdb.com</td>
<td>0.68</td>
<td>0.59</td>
<td>0.59</td>
<td>284</td>
</tr>
<tr>
<td>wikipedia.org</td>
<td>0.65</td>
<td>0.52</td>
<td>0.57</td>
<td>481</td>
</tr>
</tbody>
</table>

**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
### Order

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>U</td>
<td>T</td>
<td>D</td>
<td>B</td>
<td><strong>name-1$first-name$0</strong> frame</td>
</tr>
</tbody>
</table>
|   |   |   |   |   | name first-name "Shirley"
|1  | U | T | D | B | **name-1$last-name$0** frame |
|   |   |   |   |   | name last-name "Jones"
|2  | U | T | D | B | **date-1$month$0** frame |
|   |   |   |   |   | birth date-instant month 3
|3  | U | T | D | B | **date-1$day$0** frame |
|   |   |   |   |   | birth date-instant day 31
|4  | U | T | D | B | **date-1$year$0** frame |
|   |   |   |   |   | birth date-instant year 1934
|5  | U | T | D | B | **place-2$city$0** frame |
|   |   |   |   |   | birth place city "Smithton"
(B) Document Structuring results

- **Kendall’s \( \tau \)**

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

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

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.47</td>
<td>0.94 ± 0.10</td>
</tr>
<tr>
<td>Variant 4</td>
<td>0.52 ± 11.43</td>
<td>0.89 ± 0.12</td>
</tr>
</tbody>
</table>
(B) Learned Words Examples

(⟨birth date month⟩, 3) March.

(⟨birth date day⟩, 17) 17.


(⟨significant-other #TYPE⟩, c-fiancée) dated, engaged.

(⟨occupation #TYPE⟩, c-job-comedian) comic, stand, Comedian, Comedy, comedian, comedy, comedic, Comedians.
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 children including one half-brother and one half-sister from her mother’s first marriage. Her parents Michael and Cindy Horowitz were hippie intellectuals, and family friends included the likes of Beat poet Allen Ginsberg and interculture guru Timothy Leary who was a key influence. Her 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 1984, Ryder was performing a monologue chosen from J.D. Salinger’s “Franny & Zooey” at ACT when Deborah Lucchessi, a talent scout,...
(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 \( \neg sel \).

<table>
<thead>
<tr>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>“John”</td>
</tr>
<tr>
<td>weight</td>
<td>150Kg</td>
</tr>
<tr>
<td>award</td>
<td>“Oscar”</td>
</tr>
<tr>
<td>award</td>
<td>“MTV”</td>
</tr>
<tr>
<td>relative</td>
<td>c-son</td>
</tr>
<tr>
<td>relative</td>
<td>“Steve”</td>
</tr>
<tr>
<td>relative</td>
<td>c-step-cousin</td>
</tr>
<tr>
<td>relative</td>
<td>“Martin”</td>
</tr>
</tbody>
</table>

Always include \( \text{name} \rightarrow \text{first} \).
- **CS is labelling atomic pieces of knowledge**
  - Labelling with two labels, select \((sel)\) or omit \((-sel)\).

<table>
<thead>
<tr>
<th>Relation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{name} \rightarrow \text{first})</td>
<td>“John”</td>
</tr>
<tr>
<td>weight</td>
<td>150Kg</td>
</tr>
<tr>
<td>(\text{award} \rightarrow \text{name})</td>
<td>“Oscar”</td>
</tr>
<tr>
<td>(\text{award} \rightarrow \text{name})</td>
<td>“MTV”</td>
</tr>
<tr>
<td>(\text{relative} \rightarrow \text{type})</td>
<td>c-son</td>
</tr>
<tr>
<td>(\text{relative} \rightarrow \text{name} \rightarrow \text{first})</td>
<td>“Steve”</td>
</tr>
<tr>
<td>(\text{relative} \rightarrow \text{type})</td>
<td>c-step-cousin</td>
</tr>
<tr>
<td>(\text{relative} \rightarrow \text{name} \rightarrow \text{first})</td>
<td>“Martin”</td>
</tr>
</tbody>
</table>

**Include only if** \(\text{award} \rightarrow \text{name} \in \{“Oscar”\}\.}
CS is labelling atomic pieces of knowledge

- Labelling with two labels, select \( sel \) or omit \( \neg sel \).

<table>
<thead>
<tr>
<th>(name → first)</th>
<th>“John”</th>
</tr>
</thead>
<tbody>
<tr>
<td>(weight)</td>
<td>150Kg</td>
</tr>
<tr>
<td>(award → name)</td>
<td>“Oscar”</td>
</tr>
<tr>
<td>(award → name)</td>
<td>“MTV”</td>
</tr>
<tr>
<td>(relative → type)</td>
<td>c-son</td>
</tr>
<tr>
<td>(relative → name → first)</td>
<td>“Steve”</td>
</tr>
<tr>
<td>(relative → type)</td>
<td>c-step-cousin</td>
</tr>
<tr>
<td>(relative → name → first)</td>
<td>“Martin”</td>
</tr>
</tbody>
</table>

Include only if \( \langle \text{relative} \rightarrow \text{type} \rangle \in \{ \text{c-son} \} \).
Include only if \( \langle \text{relative} \rightarrow \text{name} \leftarrow \text{type} \rangle \in \{ \text{c-son} \} \).
(C) Approach: Content Selection Rules

- Learning from
  - Input
  - Output

- Representation (rules)

| $C_1$: constraints in node | $\mathcal{P}$: path to other node | $C_2$: constraints in other node |
(C) Approach: Content Selection Rules

- Each rule is executed and its output compared to the automatically obtained reference
I use the weighted F-measure over the labels as fitness:

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

where

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

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

This function captures the problem well and allows selecting solutions that prefer precision or recall through the \( \alpha \) parameter.
• **Combining two rules**

![Diagram](image)

- The new rule shares 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
(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**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Variant 0</th>
<th>Variant 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>biography.com</td>
<td>0.60</td>
<td>0.61</td>
</tr>
<tr>
<td>s9.com</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>imdb.com</td>
<td>0.58</td>
<td>0.32</td>
</tr>
<tr>
<td>wikipedia.org</td>
<td>0.85</td>
<td>0.18</td>
</tr>
</tbody>
</table>

- **Tri-partite (CS) Rules.**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>P</th>
<th>R</th>
<th>$F^*$</th>
<th>sel</th>
</tr>
</thead>
<tbody>
<tr>
<td>biography.com</td>
<td>0.58</td>
<td>0.72</td>
<td>0.64</td>
<td>410</td>
</tr>
<tr>
<td>s9.com</td>
<td>0.34</td>
<td>0.49</td>
<td>0.40</td>
<td>248</td>
</tr>
<tr>
<td>imdb.com</td>
<td>0.50</td>
<td>0.46</td>
<td>0.48</td>
<td>338</td>
</tr>
<tr>
<td>wikipedia.org</td>
<td>0.52</td>
<td>0.37</td>
<td>0.43</td>
<td>433</td>
</tr>
</tbody>
</table>
• **Only performed in biography.com**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Prec.</th>
<th>Rec.</th>
<th>$F^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>j48 (C4.5)</td>
<td>0.68</td>
<td>0.49</td>
<td>0.57</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.62</td>
<td>0.49</td>
<td>0.55</td>
</tr>
<tr>
<td>SMO (SVM)</td>
<td>0.61</td>
<td>0.50</td>
<td>0.55</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.62</td>
<td>0.57</td>
<td>0.59</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.60</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>SELECT-ALL/SELECT-NONE</td>
<td>0.58</td>
<td>0.66</td>
<td>0.62</td>
</tr>
<tr>
<td>CS Rules</td>
<td>0.58</td>
<td>0.72</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Experiment: Prec. and Rec. at different values of $\alpha$
## (C) Experiment: Cross-Corpora application of the rules

<table>
<thead>
<tr>
<th>Tested on</th>
<th>Trained on</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>biography.com</td>
<td>s9.com</td>
<td>imdb.com</td>
<td></td>
</tr>
<tr>
<td>biography.com</td>
<td>P:0.58</td>
<td>F*: 0.64</td>
<td>P:0.17</td>
<td>F*: 0.28</td>
</tr>
<tr>
<td></td>
<td>R:0.72</td>
<td></td>
<td>R:0.79</td>
<td></td>
</tr>
<tr>
<td>s9.com</td>
<td>P:0.66</td>
<td>F*: 0.46</td>
<td>P:0.34</td>
<td>F*: 0.40</td>
</tr>
<tr>
<td></td>
<td>R:0.35</td>
<td></td>
<td>R:0.49</td>
<td></td>
</tr>
<tr>
<td>imdb.com</td>
<td>P:0.56</td>
<td>F*: 0.44</td>
<td>P:0.23</td>
<td>F*: 0.33</td>
</tr>
<tr>
<td></td>
<td>R:0.37</td>
<td></td>
<td>R:0.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P:0.50</td>
<td>F*: 0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R:0.46</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(C) Example Rules

\[
\langle \text{person} \rightarrow \text{name} \rightarrow \text{first} \rangle:
(\text{-}, \text{-}, \text{-}). \quad \text{;TRUE}
\]

Always say the first name of the person being described.

\[
\langle \text{education} \rightarrow \text{place} \rightarrow \text{country} \rightarrow \text{name} \rightarrow \text{last} \rangle:
(\text{value} \in \{\text{“Scotland”, “England”}\}, \text{-}, \text{-}).
\]

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

\[
\langle \text{significant-other} \rightarrow \#\text{TYPE} \rangle:
(\text{value} \in \{\text{c-husband, c-wife}\}, \text{-}, \text{-}).
\]

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

\[
\langle \text{relative} \rightarrow \text{name} \rightarrow \text{last} \rangle:
(\text{-}, (\langle \text{-last -name -relative \#TYPE} \rangle, \text{value} \in \{\text{c-father}\}).
\]

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

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-1, surgerylen-1

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

- Semantic input
- Transcripts
- Order constraints
- Genetic search
- Genetic pool
- Mutations crossover
- Fitness function
- Planner
(D) MAGIC Approach

- Order Constraints

\[ \text{Sequences} \downarrow \quad \text{Pattern Detection} \downarrow \quad \text{Patterns} \downarrow \quad \text{Clustering} \downarrow \quad \text{Cluster of Patterns} \downarrow \quad \text{Constraints Inference} \downarrow \quad \text{Order Constraints over Clusters} \]
• $F_O$ 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.

\[
\text{plan} \quad \begin{array}{ccccccc}
\text{A} & \text{A} & \text{B} & \text{A} & \text{C} & \text{D} & \text{C} & \text{F}
\end{array}
\]

\[
\text{order constraint} \quad \begin{array}{cc}
\text{A} & ? \quad \text{A} \\
\text{D} & \text{C}
\end{array}
\]
(D) MAGIC Approach
(D) MAGIC Approach

- Alignment

This alignment produces a 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
Baseline
- The first generation of three runs (a total of 6,000 random instances).
- Scored using Kendall’s \(\tau\) against the MAGIC planner, they had an average \(\tau\) of 0.0952 ± 0.1144.

Learned Planners
- The best instance for each run at each iteration step is scored against the sequence obtained from the MAGIC planner.
- The average over the three runs gave \(\tau\) of 0.2288 ± 0.0342.
Fitness function during training (50 iterations)
(D) MAGIC Document Structuring Results

Fitness function during training (22 iterations)
(D) MAGIC Document Structuring Results

Overall Population Goodness

iteration

bad
average
good
(D) ProGenIE Data

- wikipedia.org

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Average</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># pairs</td>
<td>361</td>
<td>-</td>
<td>341</td>
<td>20</td>
</tr>
<tr>
<td># frames</td>
<td>58,387</td>
<td>161.737</td>
<td>55,326</td>
<td>3,061</td>
</tr>
<tr>
<td># triples</td>
<td>108,009</td>
<td>299.194</td>
<td>102,297</td>
<td>5,712</td>
</tr>
<tr>
<td># words</td>
<td>68,953</td>
<td>191.006 ± 55.17</td>
<td>64,784</td>
<td>4,169</td>
</tr>
<tr>
<td># chars</td>
<td>418,035</td>
<td>1,157.992 ± 334.01</td>
<td>392,925</td>
<td>25,110</td>
</tr>
</tbody>
</table>

- Orderings Quality

<table>
<thead>
<tr>
<th>avg. length</th>
<th>(\tau)</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.35 ± 11.4260</td>
<td>0.8909 ± 0.1154</td>
</tr>
</tbody>
</table>
(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).

\[
T(i, j) = \max \begin{cases} 
T(i - 1, j) & \text{if } T(i - 1, j) \text{ was a mismatch} \\
T(i - 1, j - 1) & \text{(skip)} \\
T(i, j - 1) & \text{(match)} \\
\end{cases} + c(i, j) \\
\]

\[
c(i, j) = \begin{cases} 
1 & \text{if } v \in s \\
-1 & \text{if } v \notin s \\
\end{cases} 
\]
(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

Learning curve

- # of patterns
- # of clusters
- # patterns found
- # pairs found
- # clusters found
- # correct constraints
- # mixed constraints
- # invalid constraints

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

(Bontcheva and Wilks, 2004)