PROGENIE: Biographical descriptions for Intelligence Analysis

Pablo Duboue, Kathleen McKeown and Vassileios Hatzivassiloglou

Computer Science Department
Columbia University
in the city of New York

Goals

- Provide final users with quick and concise descriptions
  - Foreign military personnel
  - Foreign political personnel
  - Terrorists
  - Criminal
- Customizable
  - Different users
  - Different scenarios
  - Different requirements
- PROGENIE's approach

  On the fly generation of person's descriptions

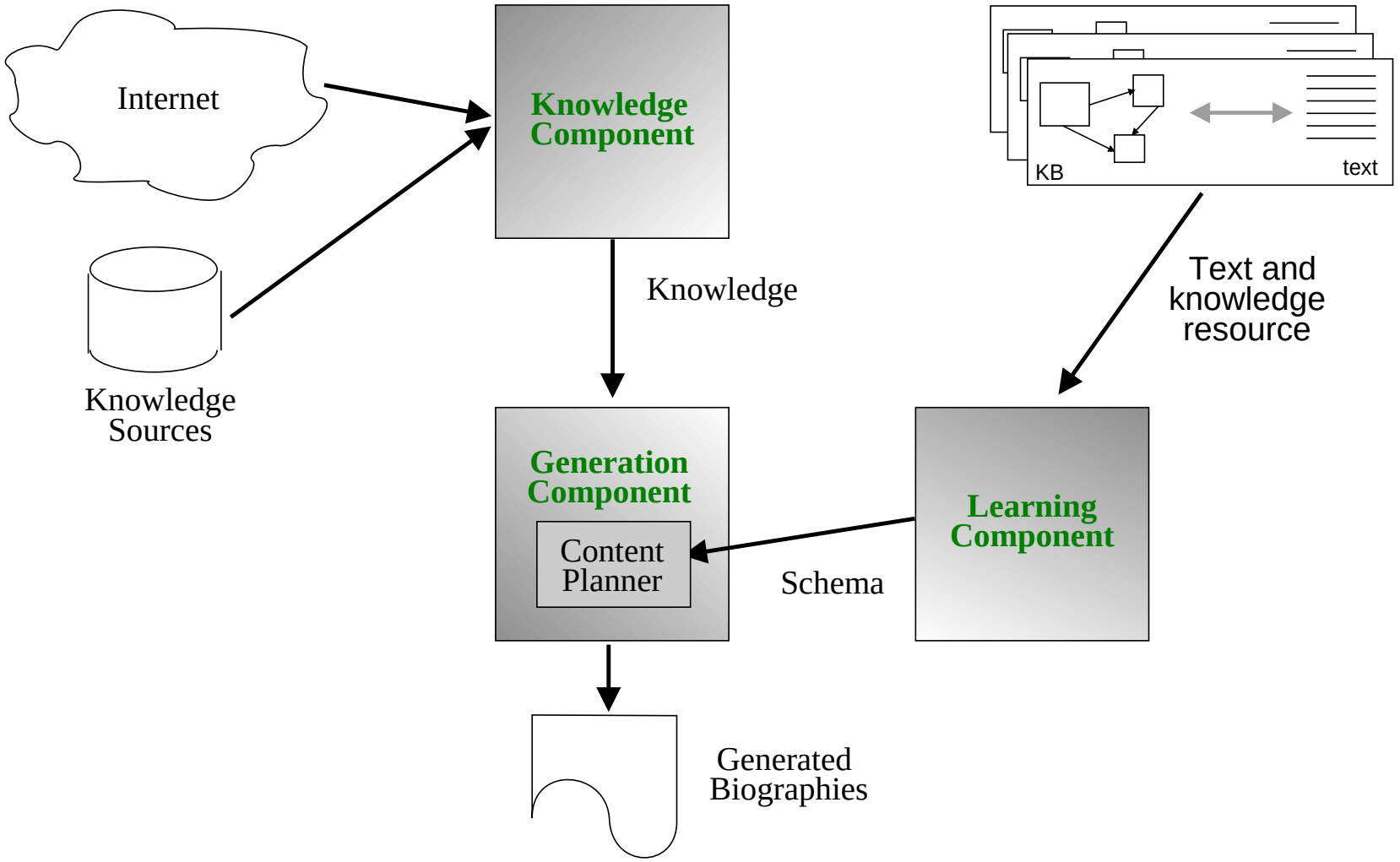
Motivation and Relevance

- Information Retrieval
  - Look for existing biographies
- Summarization
  - Integrate pieces of text from various textual sources
- Natural Language Generation (NLG)
  - Create text from structured information sources

PROGENIE's Approach

- Builds on the NLG tradition
  - Diverges from it, automatically construct content plans
- Combine a generator with an agent-based infrastructure
- Mix textual with non-textual sources

System Description



Learning Component

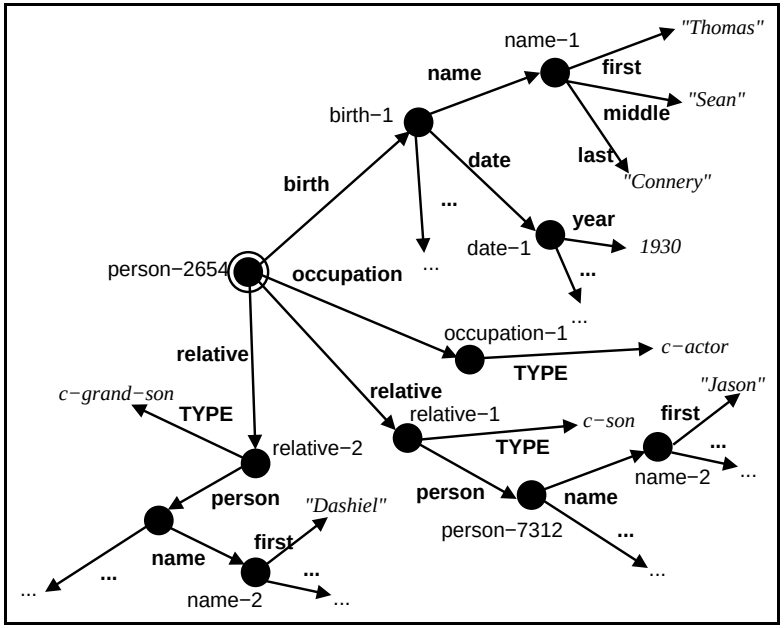
- Content Planner
  - **Structuring:** Distribution of the information among textual elements
  - **Selection:** Filtering of the available data
- Schemas
  - An implementation for Content Planners (McKeown, 1983)
- Construct Content Planning Schemas, from training data
  - Training material: data and biographies
  - The learned schemas will be used with new, unseen people

Text and Knowledge Resource

- Celebrities
  - Easily available
  - Representative of the learning issues
  - Possibility of corpus re-distribution
- Size
  - Data frames for 1,100 different celebrities
  - assorted biographies, ranging from 110 to 500 words
  - Data and biographies crawled from independent web sites

Example of Text and Knowledge Resource

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. . . .



Learning of Content Selection Rules (1)

- To appear
  - Duboue and McKeown, “Statistical Acquisition of Content Selection Rules for Natural Language Generation”, EMNLP 2003
- Goals
  - Analyze how variation on the data influence variations in the text
  - Obtain high-level content selection rules, to filter out the input

Learning of Content Selection Rules (2)

- Example

  Given:

  – (KB-1,Bio-1),(KB-2,Bio-2),(KB-3,Bio-3),(KB-4,Bio-4)

  If:

  – KB-$\{1, 2\}$ contain $\langle$`birth place state` $`MD'\rangle$

  – KB-$\{3, 4\}$ contain $\langle$`birth place state` $`NY'\rangle$

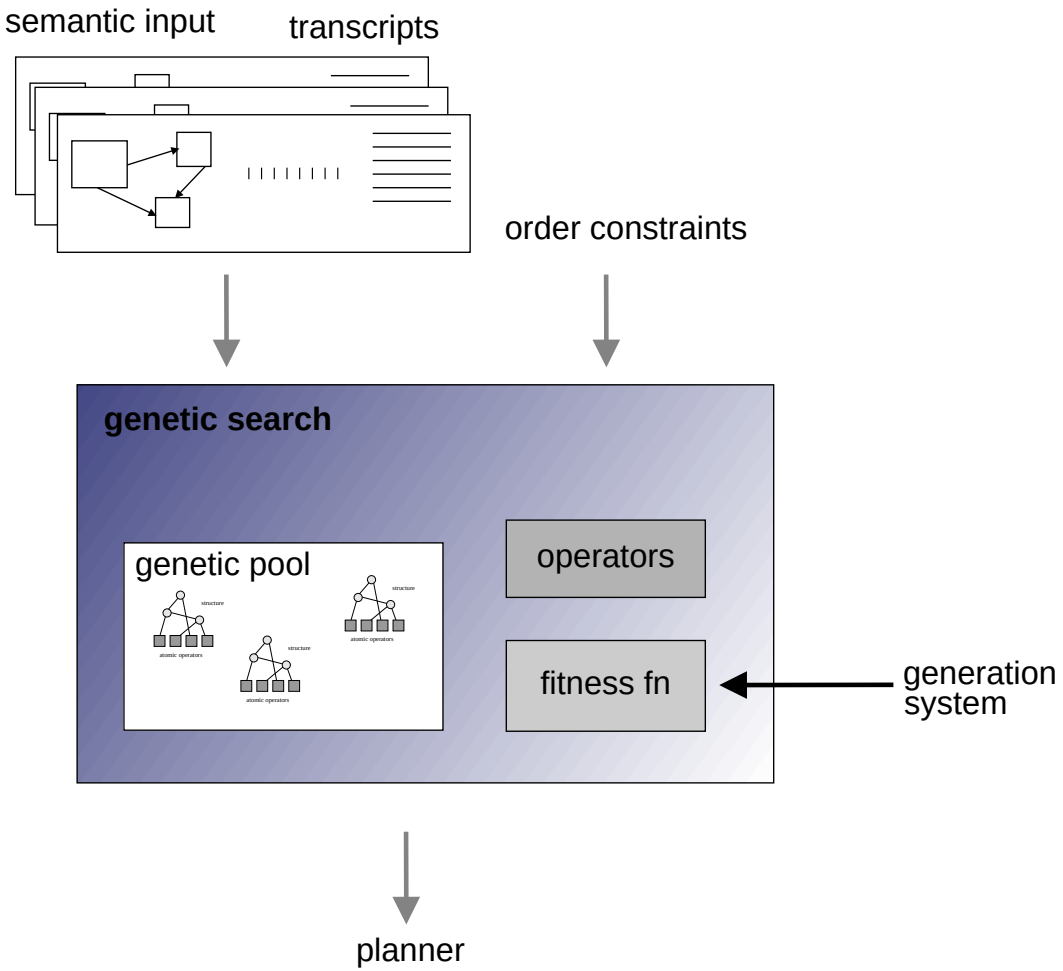
  Then:

  – Compare the language models of Bio-$\{1, 2\}$ against Bio-$\{3, 4\}$.

  – If the models differ (cross entropy), content select $\langle$`birth place state`$\rangle$.

Learning of Content Planning Schemas

- Earlier experiments performed in a medical domain.
- Corpus collected during the evaluation described in McKeown et al. (2001).
- In Duboue and McKeown (2001), we mined the corpus to extract ordered constraints between semantic elements.
- In Duboue and McKeown (2002), we used the corpus to learn content planning schemas using an alignment-based fitness function.



Knowledge Component

- Data for **Learning**
  - Supplied by internal databases and networks
  - E.g., Intelink, IAFIS
- Data for **Execution**
  - Information Extraction Agents on the Internet
  - Publicly available data as a test bed
  - Data represented in RDF (Semantic Web)

Generation Component

1. ***Inference Module*** Limited world knowledge inferencing
2. **Content Planner** McKeown's schemas
3. **Text Planner** Splits a rhetorical tree into paragraphs
4. **Referring Expression Generator** Handles pronominalization
5. ***Aggregation*** Mixes together clauses with similar structure
6. **Lexical Chooser** Selects words for concepts
7. **Surface Realizer** FUF/SURGE unification based realizer

Generated Example

Osama Bin Laden

- **overview:**
  - **name of the person:**
    - He is Usama Bin Laden.
  - **place of birth:**
    - He was born in Saudi Arabia.
  - **nationality of the person:**
    - He was a national of Saudi Arabia.
    - He does not currently have a nationality.
  - **occupation:**
    - He is a terrorist.
    - He is the leader of Al-Qaeda.
    - He is a civil engineer.
    - He is a constructor.
  - **education received:**
    - He attended the primary school in Jeddah, Saudi Arabia.
    - He attended the secondary school in Jeddah, Saudi Arabia.
    - To study security, the CIA gave him training according to Hazhir Teimourian.

Conclusions

- PROGENIE
  - Solves an existing requirement for intelligence and law enforcement personnel
- Status
  - Prototype **Learning Component** implemented in an earlier domain
    - New version, acquired Content Selection rules
  - **Generation Component**, five operational modules
  - **Knowledge Component**, under construction