Empirically Estimating Order Constraints for Content Planning in Generation

Pablo A. Duboue and Kathleen R. McKeown

Computer Science Department
Columbia University
in the city of New York

A Natural Language Generation Pipeline

- Generation Pipeline from (Reiter 1994, Reiter and Dale 2000):

  1. Content Planning

     What to say

  2. Sentence Planning

     Division into sentences

  3. Surface Realisation

     How to say it

Our Task

- Applying Empirical Methods to Content Planning
  - Content Planning is Deeply Tied to Semantics
- Learning Backbone Ordering Constraints
  - Important in Practice
  - Dependent only on the Domain Semantics
- Easily Extendable

diabetic patients and past medical history

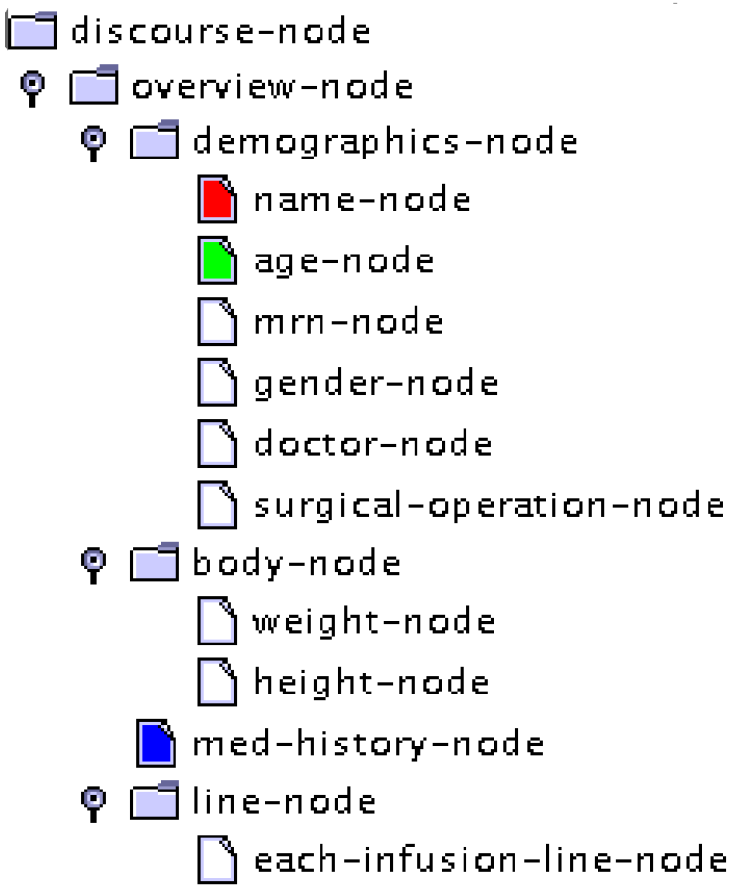
Task Specification

- Input
  - Set of Semantically Tagged Texts
- Output
  - Elements
    - Sequence of Semantic Tags
  - Global Ordering over Elements
- Methods
  - Apply Computational Biology over the Sequences of Tags

Our System: MAGIC

- MAGIC
  - Fully Developed
  - Intelligent Multimedia Presentation System
  - Medical Domain
- Task
  - Reporting Cardiac Surgery Patient Status
  - Time Critical

MAGIC: Example



"J. Doe is a seventy-eight year-old male patient of Doctor Smith undergoing aortic valve replacement. His medical history includes allergy to penicillin and congestive heart failure. He is sixty-six kilograms and one hundred sixty centimeters. . . . . ."

The Data

- From the Evaluation Described in (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 ($\min = 13$, $\max = 66$, $\sigma = 11.6$)

The Data: Example

"He is 58-year-old age male. gender History is significant for Hodgkin's disease, pmh
treated with . . . to his neck, back and chest. Hyperspadias, pmh BPH, pmh
hiatal hernia pmh and proliferative lymph edema in his right arm. pmh No IV's
or blood pressure down in the left arm. Medications — Inderal , med-preop
Lopid , med-preop Pepcid , med-preop nitroglycerine drip-preop and heparin. med-preop EKG has PAC's. ekg-preop
His Echo showed AI, MR of 47 cine amps with hypokinetic basal region.
echo-preop
Hematocrit 1.2, otherwise his labs are unremarkable. Went to OR for what was
hct-preop
felt to be 2 vessel CABG off pump both mammaries . . . . . "
procedure

Analysis of the Problem

- Focus on the **Sequence** of Semantic Tags:

age, gender, pmh, pmh, pmh, pmh, med-preop, med-preop, med-preop, drip-preop, med-preop, ekg-preop, echo-preop, hct-preop, procedure, . . .

- Find Regularities in Sequences
- Biological Sequence Analysis Techniques
  - Similar problems
  - Scalability

How to Learn Order Constraints

- Measure the Frequency of Possible Orderings
  - Ordering of Elements Built over Semantic Tags
- Reject Incorrect Orderings
- Build Table of Counts, Compute Probabilities
  - Similar to Shaw and Hatzivassiloglou ’99
- Suitable Elements:
  - Increase Regularity in the Input

More Regularity: Motif Detection

- Motifs
  - A small subsequence, highly conserved through evolution
  - A Fixed-length Pattern

	I	II	III	IV	V	VI
HT13	pvk**K**a--	t-**IDL**kdaf	-**LPQG**-fk	q**YMDDI**ll	sh**GL**--	k**FLG**qii
NVV0	ikk**K**---	ti**LDI**gday	-**LPQG**-wk	-**YMDDI**yi	qy**GFM**-	k**WLG**fel
SFV1	pvp**K**p--	tt**LDL**tngf	-**LPQG**-fl	a**YVDDI**yi	na**GYV**v	e**FLG**fni
HERVC	pvp**K**p--	tc**LDL**kdaf	-**LPQR**-fk	q**YVDDL**ll	tv**GIR**c	c**YLG**fti
GMG1	mvr**K**a--	tk**VDV**raaf	-**CPFG**-la	a**YLDDI**li	--**GLN**-	k**YLG**fiv

- Motif Detection Algorithms
  - TEIRESIAS

TEIRESIAS

- Pattern Discovery Algorithm
- Benefits
  - Swapped Domains

    a–b–c

    c–b–a
  - Hand-tunable Parameters
- Algorithm Sketch
  - Identify Basic Patterns
  - Grow Patterns (“Convolution”)
  - Find Patterns with Enough **Support**

More Regularity: Clustering

- Capturing Further Regularities

  intraop-problems, **intraop-problems**, **?**, **drip**

  intraop-problems, **?**, **drip**, **drip**

- Solution: Clustering
  - Agglomerative Clustering
  - Approximate Matching Distance
    - Measures Similarity Related to the Training-set
  - Parameterized with a Distance Threshold

Final Algorithm

Sequences

⇓

Motif (Pattern) Detection

⇓

Patterns

⇓

Clustering

⇓

Cluster of Patterns

⇓

Constraints Inference

⇓

Order Constraints over Clusters

Results

- Evaluation Settings
  - Using the 24 transcripts
  - 3-fold Cross Validation
  - Hand-tuning of Parameters
- **Constraint Accuracy:** 89.45%

Qualitative Evaluation

- Evaluation Setting
  - Using All Available Data
  - Same Parametric Settings
  - 29 constraints, out of 23 clusters
- Comparison to the Existing Content Planner
  - All the Constraints Found were Validated
  - Gained Placement Constraints for 2 Pieces of New Information
  - Learned Minor Order Variations in the Placement of 2 Rules

Conclusion

- A Novel Empirical Method for Learning of Content Planning Elements
  - Relating the Problem to Biological Sequence Analysis
- Successful Results
  - Feasibility of the Task
  - High Precision and Increased Variability of the Plan

Further Work

- Integrate Results
  - Genetic Search over the Planners Space
  - Alignment Scores as a Measure of Similarity
- Automatic Tagging
- Explore Other Alternatives
  - Pattern Expressibility
  - Other Techniques, both in NLP and Bioinformatics