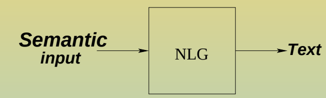
Statistical Generation

Regina Barzilay

`regina@cs.columbia.edu`

April 15, 2001

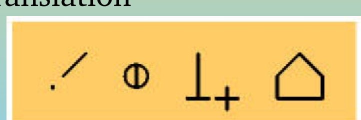
What is NLG?



- Input: databases, expert systems, log files.
- Output: reports, help messages, summaries.

Motivation

- Facilitate Information Access
  - Search engines and Q&A operate over text
  - Speech syntesizers operate over text
- Augmentative and Alternative communication
- Machine Translation



NLG Systems

- MAGIC — briefing of patient status (McKeown et al, 1996)
- ANA — stock market reports (Kukich, 1983)
- STREAK — basketball game reports (Robin & McKeown, 1993)

Input/Output Example

```
scoring((Shaquille, O'Neal), 37)
time(Friday, night)
team((Shaquille, O'Neal), (Orlando, Magic))
win(Orlando, Magic), (Toronto, Raptors)
score(101,89)
...
```

Orlando, FL — Shaquille O'Neal scored 37 points Friday night powering the Orlando Magic to a 101 89 victory over the Toronto Raptors, losers of seven in a row.

What to Say?

Game statistics: `win(Orlando, Magic), (Toronto, Raptors)`
Player's records: `team((Shaquille, O'Neal), (Orlando, Magic))`
Team's record: `lost(7,(Toronto, Raptors))`
Player's height: `height (O'Neal, 2m)`

Content Determination and Organization

- Goal: select predicates from the given knowledge base and order them
- Challenge: define text well-formedness
- Assumption: texts in the same ganre exhibit similarity in structure
- Method: schemata, text grammar, scripts

  Report $\longrightarrow$ Stats, Main events, History

How to Say: Aggregation

Before aggregation:

```
Shaquille O'Neal scored 37 points.  The
game was on Friday night.  Orlando
Magic defeated Toronto Raptors.
Raptors lost seven games in a row.
```

After aggregation:

```
Shaquille O'Neal scored 37 points
Friday night powering the Orlando Magic
to a 101 89 victory over the Toronto
Raptors, losers of seven in a row.
```

How to say: Lexicalization

Map domain concepts to words:

- Constraints: discourse focus, style constraints, syntactic environment.
- Implementation: decision tree.

Example: `win(X,Y)`

Verb:	X defeated Y, Y was defeated X, X won in the game against Y
Noun:	victory of X over Y, victory of X, defeat of Y, X's triumph

How to Say: Realization

- Insert function words.
- Choose correct specification of content words.
- Order words.

FUF/SURGE input for the sentence "John likes Mary now":

```
((cat clause)
  (proc   ((type mental) (tense present) (lex "like")))
  (partic ((processor  ((cat proper) (lex "John")))
           (phenomenon ((cat proper) (lex "Mary")))))
  (circum ((time ((cat adv) (lex "now"))))))
```


NLG Architecture

- Content Determination
- Discourse Planning
- Sentence Aggregation
- Lexicalization
- Syntactic and morphological realization

90's: Generation Renaissance

Traditional generation

- Most of the components are applications specific and not reusable
- A lot of hand-crafted rules

Statistical Generation

- 1995: Knight & Hatzivassiloglou developed first statistical surface realizer
- Today: Empirical methods are developed for several components of generation system

Statistical Content Planning

Duboue & McKeown ACL 2001

- Goal: learn ordering constraints
- Given: set of transcripts manually annotated with semantic units

Annotated Transcript

He is $\underline{\text{58-year-old}}_{\text{age}}$ $\underline{\text{male}}_{\text{gender}}$. History is significant for $\underline{\text{Hodgkin's disease}}_{\text{pmh}}$, treated with ...to his neck, back and chest. $\underline{\text{Hyperspadias}}_{\text{pmh}}$, $\underline{\text{BPH}}_{\text{pmh}}$, $\underline{\text{hiatal hernia}}_{\text{pmh}}$ and $\underline{\text{proliferative lymph edema in his right arm}}_{\text{pmh}}$. No IV's or blood pressure down in the left arm. Medications — $\underline{\text{Inderal}}_{\text{med-preop}}$, $\underline{\text{Lopid}}_{\text{med-preop}}$, $\underline{\text{Pepcid}}_{\text{med-preop}}$, $\underline{\text{nitroglycerine}}_{\text{drip-preop}}$ and $\underline{\text{heparin}}_{\text{med-preop}}$. $\underline{\text{EKG has PAC's}}_{\text{ekg-preop}}$. ...

Semantic Sequence

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

Pattern Detection

Analogous to motif detection

T_1: $\underline{\text{A B C D}}$ F A $\underline{\text{A B F D}}$

T_2: F C $\underline{\text{A B D D}}$ F F

- Scanning
- Generalizing
- Filtering

Example of Learned Pattern

intraop-problems

intraop-problems

$$\left\{ \begin{array}{ll} \text{operation} & 11.11\% \\ \text{drip} & 33.33\% \\ \text{intraop-problems} & 33.33\% \\ \text{total-meds-anesthetics} & 22.22\% \end{array} \right\}$$

drip

Evaluation

Pattern confidence: 84.62%

Constraint accuracy: 89.45%

Learning Lexical Choice

Barzilay & Lee, EMNLP 2002

- Goal: induce the mapping between semantic concepts and their verbalization
- Given: Semantic input and sentences generated by humans
- Implementation: Verbalization of Nuprl Proofs

[Parent [sex:female]]	mother
[love(x,y)]	x loves y, x is in love with y

Learning a Lexicon

- Automatically align semantic inputs and their verbalizations

Semantic input	`Instruct(McKeown,CS422,12:00pm)`
Verbalization input	Prof. McKeown teaches the NLP class at noon

- Induce lexicon entries:

`Instruct`$(arg_1, arg_2, arg_3) \rightarrow arg_1$ teaches arg_2 at arg_3

CS422 $\rightarrow$ NLP class

...

Challenges of Alignment

- Hard to match semantics with a single verbalization

Semantic input:

`show-from(a=0,b=0,a*b=0)`

Verbalizations:

Given a and b as in the theorem statement, prove that $a * b = 0$.
Suppose that a and b are equal to zero. Prove that their product is also zero.
Assume that $a = 0$ and $b = 0$.

Our Approach: Multi-Sequence Alignment

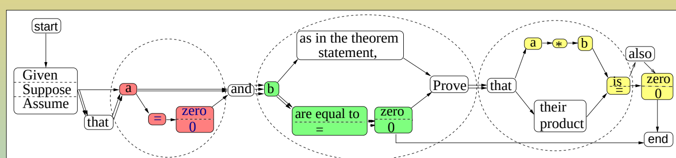
- MSA in biology – find commonalities in bilogical sequences of the same family

a	b	c	–	a
a	b	a	b	a
a	c	c	b	–
c	b	–	b	c

- We wish to compare semantics with all verbalizations *simultaneously*
  - Solution: Use MSA to build composite of verbalizations
  - Rationale: ameliorate "mutations" within individual verbalizations
  - Gains: accuracy and expressiveness

MSA Lattice

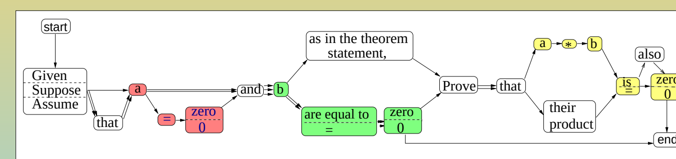
`show-from(a=0,b=0,a*b=0)`

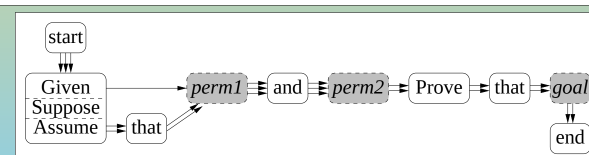


Given a and b as in the theorem statement, prove that $a * b = 0$.
Suppose that a and b are equal to zero. Prove that their product is also zero.
Assume that $a = 0$ and $b = 0$.

Matching against Semantics

`show-from(a=0,b=0,a*b=0)`





Evaluation of the Generated Text

- Baseline: traditional generation system
  (Holland-Minkley et al., AAAI '99)
- Fidelity:
  - 20 proofs judged by Nuprl expert
  - Binary judgment — correct, incorrect
  - Results
    MSA: 20(100%) correct
    (Holland-Minkley et al.): 17(85%) correct

Readability Results

Judge	Lemma a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
A	■	■	■	■	◆	■	■	■	■	■	■	◆	■	■	■	■	■	◆	■	◆
B	□	□	■	□	■	■	◆	■	■	◆	◆	◆	□	■	■	■	■	■	□	■
C	◆	◆	□	□	◆	■	□	■	□	□	■	◆	◆	■	■	◆	□	■	◆	◆
D	□	■	■	◆	◆	◆	■	□	■	□	□	■	◆	□	■	■	□	■	◆	■
E	□	◆	■	◆	◆	◆	■	◆	■	□	□	◆	◆	□	◆	□	■	◆	□	◆
F	■	■	■	■	□	◆	□	■	■	■	■	■	■	■	□	■	■	■	■	■
G	□	■	■	■	◆	□	■	■	◆	■	■	◆	■	■	□	◆	◆	■	◆	◆
H	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
I	□	□	■	◆	◆	□	□	■	■	□	■	□	□	■	◆	◆	■	■	■	□
J	◆	◆	■	◆	■	◆	◆	■	◆	◆	◆	◆	■	■	■	■	□	□	□	■
K	■	■	■	□	■	■	■	□	□	◆	□	□	■	■	■	■	□	□	◆	■
■ > 50%?		✓	✓				✓	✓	✓		✓		✓	✓	✓	✓	✓	✓		✓

preference for: ■ — MSA, □ — hand crafted, ◆ — no preference.

Generation for MT

(Knight, Langkilde, Hatzivassiloglou)

- Input: "semantic" interlingua
  semantic objects and relations between them
- Symbolic generator: transform input into word lattices
- Statistical linearizer: select the best path in the lattice

Interlingua example

```
(A / |workable|
      :DOMAIN (A2 / |sell<cozen|
      :AGENT I
      :PATIENT (T / |trust, reliance|
                   :GPI THEY))
:POLARITY NEGATIVE)
```

Symbolic Generator

```
((x1 :agent) (x2 :patient) (x3 :rest)
 ->
 (s (seq (x1 np nom-pro) (x3 v-tensed) (x2 np acc-pro)))
 (s (seq (x2 np nom-pro) (x3 v-passive) (wrd ''by'')
         (x1 np acc-pro)))
 (np (seq (x3 np acc-pro nom-pro) (wrd ''of'')
          (x2 np acc-pro) (wrd ''by'') (x1 np acc-pro)))
 ...)
```

Lexicon (110, 000):

```
(<word> <pos> <rank> <concept>)
(''eat'' VERB 1 |eat, take in|)
```


Resulted lattice

Statistical Component

- Input
  I cannot betray their task.
  I will not be able to betray their trust.
  I am not able to betray their trust.
  ...
  I cannot betray trust of them.
  I will not be able to betray trust of them.
- Scoring:
  $P(w_1, \ldots, w_n) =$
  $P(w_1|start) * P(w_2|w_1) * \ldots * P(w_n|w_{n-1}) * P(end|w_n)$

Limitations

- Bigrams are too simplistic.
  "`Ann are`" from "`Joe and Ann are in love`"
- Long distance relations.
- Nitrogen prefers sequences of simple words like "`was`", "`of`", "`the`".