Building Applied Natural Language Generation Systems

Robert Dale and Ehud Reiter
Overview

1. An Introduction to NLG
2. Requirements Analysis for NLG
3. NLG Architecture and System Design
4. A Case Study
5. A Closer Look at the Component Tasks
6. Conclusions and Pointers
Component Tasks in NLG

• Text Planning
  – Content determination
  – Discourse planning

• Sentence Planning
  – Sentence aggregation
  – Lexicalisation
  – Referring expression generation

• Linguistic Realization
  – Syntactic and morphological realization
  – Orthographic realization
Sentence Planning

Goal:
• To convert a text plan into a sequence of sentence definitions

Tasks:
• Aggregation
• Lexicalisation
• Reference
Aggregation

Sentence aggregation:
- combining messages into sentences

Paragraph aggregation:
- combining sentences into paragraphs

Paragraph aggregation is poorly understood
Ways of Aggregating Messages

Some possibilities:
- Simple conjunction
- Ellipsis
- Embedding
- Set introduction
Simple Conjunction

Without aggregation:
- The next train is the Caledonian Express.
- It leaves at 10AM

With aggregation:
- The next train is the Caledonian Express, and it leaves at 10AM
Ellipsis

Without aggregation:
- It leaves at 10AM
- It arrives at 12.30PM

With aggregation:
- It leaves at 10AM and Ø arrives at 12.30PM
Embedding

Without aggregation:
- The next train is the Caledonian Express.
- It leaves at 10AM

With aggregation:
- The next train, which leaves at 10AM, is the Caledonian Express.
Set Introduction

Without aggregation:
- It has a snack bar.
- It has a restaurant car.

With aggregation:
- It has \{a snack bar, a restaurant car\}.
- It has a snack bar and a restaurant car.

Caution: need to avoid changing the meaning:
- John bought a TV and Bill bought a TV.
- John and Bill bought a TV.
Choosing Aggregations

Usually many ways to aggregate a given message set:

- The next train is the Caledonian Express. It leaves at 10am
- The next train is the Caledonian Express, and it leaves at 10am
- The next train, which leaves at 10am is the Caledonian Express.
- The next train is the Caledonian Express, which leaves at 10am

Which is best?
Choice Heuristics

A few suggestions from the literature:
• conform to genre conventions and rules
• only aggregate messages which are siblings in the text plan tree
• only use relative clauses when the messages are linked by an Elaboration rhetorical relation

More research is needed
Lexicalisation

Converting concepts into words

- It **departs** at 10AM
- It **leaves** at 10AM
- It **has a departure time** of 10AM

Assume concepts are ‘close’ to words:

- Concept is DepartureTime, not a record of the train’s location and velocity at 0958, 0959, 1000, 1001, 1002
- latter case is very poorly understood
Lexicalisation Heuristics

Some suggestions from the literature

- Enforce variation - don’t overuse words
- Discourse focus:
  - John sold Mary an apple
  - Mary bought an apple from John
- Conform to level of formality and other pragmatic criteria:
  - father vs dad
- Conform to genre convention and rules:
  - Paracetamol vs Tylenol vs Acetaminophen
- Express multiple concepts with a single word:
  - He is a terrorist = kills_people & bad_person
Lexicalisation Algorithms

Decision trees:
- choose between ways of expressing a single concept, based on simple questions
- very simple and easy to implement, probably the most popular approach

Graph-rewriting:
- treat each message as a semantic network, and rewrite it according to a concept→word dictionary
- more powerful, theoretically elegant
- perhaps only justified in systems with multilingual output
A Discrimination Net

For realizing the CONTRAST rhetorical relation:

Clauses aggregated within one sentence?

Yes

"but" last used for CONTRAST?

Yes  lex = “although”

No  lex = “but”

"In contrast" last used for CONTRAST?

Yes  lex = “however”

No  lex = “in contrast”

No
Reference

How can we identify specific domain objects and entities?

- Initial introduction of an object
- Subsequent references to an already salient object
Initial Reference

Introducing an object into the discourse:
• use a full name
  – the Caledonian Express
• relate to an object that is already salient
  – the train’s restaurant car
• specify physical location
  – the train behind the Hertz billboard

Poorly understood; more research is needed
Some possibilities:

- **Pronouns**
  - _It_ leaves at 10am

- **Definite NPs**
  - _The train_ leaves at 10am

- **Proper names, possibly abbreviated**
  - _The Caledonian_ leaves at 10am
Choosing a Form of Reference

Some suggestions from the literature

• use a pronoun if it refers to an entity mentioned in the previous clause, and there is no other entity in the previous clause that the pronoun could refer to

• otherwise use a name, if a short one exists

• otherwise use a definite NP

• also important to conform to genre conventions -- for example, there are more pronouns in newspaper articles than in technical manuals
Example

I am taking the Caledonian Express tomorrow. It is a much better train than the Grampian Express. The Caledonian has a real restaurant car, while the Grampian just has a snack bar. The restaurant car serves wonderful fish, while the snack bar serves microwaved mush.
Sentence Planning: Choice

Choice may be the central issue in sentence planning:

- which aggregation, lexicalisation, or referring expression should be used?
- clear that genre/sublanguage is important in all of these choices
- evidence and support may come from psycholinguistics and practitioner’s wisdom
Another Choice Example

Which is better?
1. Press the button if the blue light comes on.
2. If the blue light comes on, press the button.

Considerations
- The psychologist’s view: (1) is read quicker
- The technical author’s view: (2) is less likely to be misinterpreted
Research Issues

- How do we make choices?
- How do we perform higher-level aggregation, such as forming paragraphs from sentences?
- How do we lexicalise if domain concepts do not easily map into words?
- What is the best way of making an initial reference to an object?
Sentence Planning

- Examine the target texts to determine the range of expression required
- Construct mappings from the text plan to sentence plans that can be given to a realiser
Sentence Planning

- Sentence aggregation rules work within substructures of the text plan
- The rules identify structures that can be combined
- Rules may combine entire messages or parts of messages
A Text Plan

TEXTPLAN

SATELLITE-01 [SEQUENCE]
- NUCLEUS: cooler than average

SATELLITE-02 [SEQUENCE]
- NUCLEUS: drier than average

SATELLITE-01 [ELABORATION]
- NUCLEUS: average # raindays
- SATELLITE-01 [CONTRAST]
  - NUCLEUS: rain so far

SATELLITE-02 [ELABORATION]
- NUCLEUS: rainspell
- SATELLITE-01 [CONTRAST]
  - NUCLEUS: rain amounts
A Text Plan

((type textplan)
 (relations ((sequence satellite- 01 satellite02)))
 (satellite- 01 ((nucleus ⟨MONTHLYTEMPMSG⟩)))
 (satellite- 02 ((nucleus ⟨MONTHLYRAINFALLMSG⟩))
   (relations ((elaboration nucleus satellite- 01)
               (elaboration nucleus satellite- 02)))
 (satellite- 01 ((relations ((contrast nucleus satellite- 01))
               (nucleus ⟨RAINYDAYSMSG⟩)
               (satellite- 01 ((nucleus ⟨RAINSOFARMSG⟩))))))
 (satellite- 02 ((relations ((contrast nucleus satellite- 01))
               (nucleus ⟨RAINSPELLMSG⟩)
               (satellite- 01 ((nucleus ⟨RAINAMOUNTSMMSG⟩)))))))
Sentence Planning Rules

Sentence planning rules are sensitive to rhetorical relations:

- If two messages are in a SEQUENCE relation they can be conjoined at the same level.
- If one message is an ELABORATION of another it can either be conjoined at the same level or embedded as a minor clause or phrase.
- If one message is a CONTRAST to another it can be conjoined at the same level or embedded as a minor clause or phrase.
A Sentence Planning Rule

Given the structure: 

- **NUCLEUS**
  - Message1

- **SATELLITE-01 [CONTRAST]**
  - **NUCLEUS**
    - Message2

The transformation rule results in:

- **S+**
  - Conj
    - although
  - S
    - Message2

Thus, the extracted message is:

**Message1**

although

**Message2**
A Sentence Planning Rule

NUCLEUS
Message1

SATELLITE-01 [CONTRAST]

NUCLEUS
Message2

S
S+

Message1

Conj
but
Message2
Variations in Describing Rainfall

Variations in syntactic category:
S: [rainfall was very poor indeed]
NP: [a much worse than average rainfall]
AP: [much drier than average]

Variations in semantics:
Absolute: [very dry]
[a very poor rainfall]
Comparative: [a much worse than average rainfall]
[much drier than average]
A Sentence Planning Rule

\[ \neg \text{MonthlyTempMsg}(x) \]
\[ \text{absolute- or- relative} \]
\[ \text{relative- to- average} \]
\[ + \]

\[ \neg \text{MonthlyRainfallMsg}(y) \]
\[ \text{absolute- or- relative} \]
\[ \text{relative- to- average} \]
A Sentence Planning Rule

\[
\begin{align*}
\text{MonthlyTempMsg}(x) & \quad \text{absolute- or- relative} \\
\text{relative- to- average} & \\
\text{+} & \\
\text{MonthlyRainfallMsg}(y) & \quad \text{absolute- or- relative} \\
\text{absolute} &
\end{align*}
\]
Many different results are possible:

- The month was cooler and drier than average. There were the average number of rain days, but the total rain for the year so far is well below average. There was rain on every day for 8 days from 11th to 18th, but rainfall amounts were mostly small.

- The month was cooler and drier than average. Although the total rain for the year so far is well below average, there were the average number of rain days. There was a small amount of rain on every day for 8 days from 11th to 18th.
Component Tasks in NLG

• Text Planning
  – Content determination
  – Discourse planning

• Sentence Planning
  – Sentence aggregation
  – Lexicalisation
  – Referring expression generation

• Linguistic Realization
  – Syntactic and morphological realization
  – Orthographic realization