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
An Introduction to NLG

• What is Natural Language Generation?
• Some Example Systems
• Types of NLG Applications
• When is NLG an Appropriate Technology?
What is NLG?

Natural language generation is the process of deliberately constructing a natural language text in order to meet specified communicative goals.

[McDonald 1992]
What is NLG?

- **Goal:**
  - computer software which produces understandable texts in English or other human languages

- **Input:**
  - some underlying non-linguistic representation of information

- **Output:**
  - documents, reports, explanations, help messages, and other kinds of texts

- **Knowledge sources required:**
  - knowledge of language and of the domain
NLP = NLU + NLG

Natural Language Understanding

Natural Language Generation

Meaning

Text

Text
Example 1: FoG

- **Function:**
  - Produces textual weather reports in English and French
- **Input:**
  - Graphical weather depiction
- **User:**
  - Environment Canada (Canadian Weather Service)
- **Developer:**
  - CoGenTex
- **Status:**
  - Fielded, in operational use since 1992
FoG: Output

FPCN20 Status: CURRENT-HOT RELEASED

FPCN20 CHWG 15Z300
MARINE FORECASTS FOR ARCTIC WATERS ISSUED BY THE ARCTIC WEATHER CENTRE OF ENVIRONMENT CANADA AT 05.00 PM MDT SATURDAY 15 APRIL 1995 FOR TONIGHT AND SUNDAY WITH AN OUTLOOK FOR MONDAY.
THE NEXT SCHEDULED FORECAST WILL BE ISSUED AT 05.00 AM MDT.
WINDS ARE IN KNOTS.
FOG IMPLIES VISIBILITY LESS THAN 5/8 NM.
HIST IMPLIES VISIBILITY 5/8 TO 6 NM.

GREAT SLAVE LAKE.
WINDS LIGHT TONIGHT AND SUNDAY. SNOW ENDING NEAR MIDNIGHT. VISIBILITIES NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... LIGHT WINDS.

GREAT DEAR LAKE.
FREEZING SPRAY WARNING ISSUED.
WINDS EAST 20 TO 25 TONIGHT AND SUNDAY. FREEZING SPRAY.
OUTLOOK FOR MONDAY... WINDS EASTERLY 20 TO 25.

MACKENZIE RIVER FROM MILE 0 TO MILE 100.
WINDS LIGHT TONIGHT AND SUNDAY. SNOW ENDING THIS EVENING. VISIBILITIES NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... LIGHT WINDS.

MACKENZIE RIVER FROM MILE 100 TO MILE 300.
WINDS LIGHT STRENGTHENING TO SOUTHEAST 15 SUNDAY AFTERNOON. SNOW ENDING EARLY THIS EVENING. VISIBILITIES NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... WINDS SOUTHEASTERLY 15.
Example 2: PlanDoc

- **Function:**
  - Produces a report describing the simulation options that an engineer has explored

- **Input:**
  - the simulation log file

- **User:**
  - Southwest Bell

- **Developer:**
  - Bellcore and Columbia University

- **Status:**
  - Fielded, in operational use since 1996
PlanDoc: Input

RUNID fiberall FIBER 6/19/93 act yes
FA 1301 2 1995
FA 1201 2 1995
FA 1401 2 1995
FA 1501 2 1995
ANF co 1103 2 1995 48
ANF 1201 1301 2 1995 24
ANF 1401 1501 2 1995 24
END. 856.0 670.2
This saved fiber refinement includes all DLC changes in Run-ID ALLDLC. RUN-ID FIBERALL demanded that PLAN activate fiber for CSAs 1201, 1301, 1401 and 1501 in 1995 Q2. It requested the placement of a 48-fiber cable from the CO to section 1103 and the placement of 24-fiber cables from section 1201 to section 1301 and from section 1401 to section 1501 in the second quarter of 1995. For this refinement, the resulting 20 year route PWE was $856.00K, a $64.11K savings over the BASE plan and the resulting 5 year IFC was $670.20K, a $60.55K savings over the BASE plan.
Example 3: AlethGen

- **Function:**
  - Produce a letter (in French) to a customer from a customer-service representative

- **Input:**
  - Customer database, plus information entered by the service rep with a GUI

- **User:** La Redoute (French mail-order company)

- **Developer:** ERLI

- **Status:**
  - Passed acceptance test, waiting for customer to finish an upgrade of their IT system
<table>
<thead>
<tr>
<th>Client no.</th>
<th>Complaint date</th>
<th>Complaint received by</th>
<th>Personal event</th>
<th>Personal event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9999987</td>
<td>7/12/94</td>
<td>Videotex</td>
<td>Holiday</td>
<td>Holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fax</td>
<td>Wedding</td>
<td>Wedding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Letter</td>
<td>Hospital</td>
<td>Hospital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Death</td>
<td>Death</td>
</tr>
<tr>
<td>Tone of the complaint</td>
<td>Date of personal event</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>21/12/94</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chère Madame,

Je suis désolée que vous ayez rencontré différents problèmes.

Les taies vertes et l'enveloppe d'édredon imprimée que vous attendez ont été expédiées le 4 Novembre 1994, et notre service de livraison aurait déjà dû les livrer à votre relais de DUNKERQUE. Je vais me renseigner pour savoir ce qui a pu se produire.

Malheureusement, je ne peux pas renouveler ce linge de lit, car nous ne pourrions pas vous le livrer avant votre départ. Dans votre intérêt, je préfère l'annuler.

Dès aujourd'hui, je régularise votre compte Paiement Confiance de 350,00 F.

D'autre part, vous n'avez pas reçu les chaussures noires. Malheureusement, nous attendons toujours leur rentrée en stock.

Je ne peux que vous demander à nouveau de patienter ou vous conseiller de reporter votre choix sur un autre article de notre catalogue.

Je vous prie de bien vouloir accepter toutes mes excuses.

Je vous souhaite de bonnes vacances.

Bien cordialement,

Nadia, rédactrice
Why Use NLG?

• Important information is stored on computers in ways which are not comprehensible to the end user:
  – graphical weather maps
  – simulation log files
  – databases and spreadsheets
  – expert-system knowledge bases

• NLG systems can present this information to users in an accessible way
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
Requirements Analysis

The developer needs to:

• understand the client’s needs
• propose a functionality which addresses these needs
Corpus-Based Requirements Analysis

A corpus

- consists of examples of output texts and corresponding input data
- specifies ‘by example’ the functionality of the proposed NLG system
- is a very useful resource for design as well as requirements analysis
Corpus-Based Requirements Analysis

Four Steps:

• assemble an initial corpus of (human-authored) output texts and associated input data
• analyze the information content of the corpus texts in terms of the input data
• develop a target text corpus
• create a formal functional specification
Step 1: Creating an Initial Corpus

- Collect a corpus of input data and associated (human-authored) output texts
- One source is archived examples of human-authored texts
- If no human-authored examples of the required texts exist, ask domain experts to produce examples
- The corpus should cover the full range of texts expected to be produced by the NLG system
Problems with the Initial Corpus

A quick scan of the corpus may reveal that

- it is impossible to automatically generate some of the texts in the initial corpus
- the corpus texts appear sub-optimal and open to improvement
- texts written by different human authors have very different structures and styles
- it is not obvious why a particular text was generated for a particular input
Step 2: Analyzing the Content of the Texts

• Goal:
  - to determine where the information present in the texts comes from, and the extent to which the proposed NLG system will have to manipulate this information

• Result:
  - a detailed understanding of the correspondences between the available input data and the output texts in the initial corpus
An Example

• Context: we want to build an NLG system which replies to rail-travel inquiries.
• The system has access to train timetable information
An Example Text

There are 20 trains daily from Aberdeen to Glasgow. The next train is the Caledonian Express. It leaves Aberdeen at 10am. This train has been running for 55 years.

Thank you for considering rail travel.
Units of Information

• The first step in analyzing the corpus for information content is to identify the individual units of information that make up each of the texts.

• A general rule of thumb: units of information correspond to individual clauses or sentences.
Units of Information

1. There are 20 trains daily from Aberdeen to Glasgow.
2. The next train is the Caledonian Express.
3. It leaves Aberdeen at 10am.
4. This train has been running for 55 years.
5. Thank you for considering rail travel.
Information Types in Text

- Unchanging text
- Directly-available data
- Computable data
- Unavailable data
There are 20 trains daily from Aberdeen to Glasgow. The next train is the Caledonian Express. It leaves Aberdeen at 10am. This train has been running for 55 years.

Thank you for considering rail travel.
Directly- Available Data

There are 20 trains daily from Aberdeen to Glasgow. The next train is the Caledonian Express. It leaves Aberdeen at 10am. This train has been running for 55 years.

Thank you for considering rail travel.
There are 20 trains daily from Aberdeen to Glasgow. The next train is the Caledonian Express. It leaves Aberdeen at 10am. This train has been running for 55 years. Thank you for considering rail travel.
There are 20 trains daily from Aberdeen to Glasgow. The next train is the Caledonian Express. It leaves Aberdeen at 10am. This train has been running for 55 years.

Thank you for considering rail travel.
Solving the Problem of Unavailable Data

• More information can be made available to the system: this may be expensive
• If the system is an authoring-aid, a human author can add this information
• The corpus can be revised to eliminate clauses that convey this information
Step 3: Building the Target Text Corpus

Mandatory changes:
• eliminate unavailable data
• specify what text portions will be human-authored

Optional changes:
• simplify the text to make it easier to generate
• improve human-authored texts
• enforce consistency between human authors
Step 4: Functional Specification

- Based on an agreed target text corpus
- Explicitly states role of human authoring, if present at all
- Explicitly states structure and range of inputs to be used
Research Issues

- Development of an appropriate corpus-analysis methodology
- Using expert-system knowledge-acquisition techniques
- Automating aspects of corpus analysis
- Integrating corpus-analysis with standard requirements analysis procedures
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
Applied NLG in 1997

- Huge increase in the number of NLG applications being investigated or developed
- Slower increase in the number of applications which are actually fielded
  - but at least there are some: in 1987 there were none
- We are beginning to see reusable software, and specialist software houses
- Applied NLG is young, but the future looks bright
Resources: SIGGEN

SIGGEN (ACL Special Interest Group for Generation)

- Web site at http://www.siggen.org/
  - papers and bibliographies
  - conference and workshop announcements
  - software
  - job announcements
  - related pages
Resources: Conferences and Workshops

- International Workshop on NLG: every two years
- European Workshop on NLG: every two years, alternating with the International Workshops
- NLG papers at ACL, ANLP, IJCAI, AAAI ...
- See SIGGEN Web page for announcements
Resources: Books

Currently being written:
Building Applied Text Generation Systems
Cambridge University Press

Don’t miss it!