Natural Language Generation
A survey

Chrysanne DiMarco

David R. Cheriton School of Computer Science
and
Department of English
University of Waterloo

September 24, 2010
Outline

1. What Natural Language Generation is
2. Macroplanning
3. Microplanning
4. Realization
5. Examples
Natural Language Generation (NLG) involves automatically producing human (natural) language from a computational representation of information.

Information can be linguistic, knowledge about the world, knowledge about the speaker/audience, ... .

Traditionally, NLG systems were concerned with:

- What should I say?
- How should I say it?

Many interrelated planning processes:

- Decide on information to be said.
- Construct discourse plan.
- “Chunk” information into units of discourse.
- Select appropriate phrases and words.
- Produce grammatically correct output.
Present-day systems

- Macroplanning (what to say).
- Microplanning (how to say it).
- A third stage: Realization (Reiter 1994).
Macroplanning (what to say)

- Macroplanning involves content choice and organization.
- Input is one or more communicative goals:
  - Explain some state;
  - Describe some object;
  - Relate some events; and so on.
- States, objects, events, etc. are defined in the NLG system’s knowledge base.
- Output is a (tree) structure:
  - Leaves are sentence-sized groupings of representation items.
  - Internal nodes are inter-sentence relations (e.g., Cause, Sequence, Elaboration, ...).
  - This tree represents the discourse structure of the text.
Macroplanners (how they work)

- To construct the discourse structure, a macroplanner requires a set of *templates* to organize the knowledge representation items.
  - Schemas. Topic structure of stereotypical paragraphs.
  - Relations/plans. More flexible relations among discourse units.
- Macroplanning has used various theories of text linguistics (e.g., Rhetorical Structure Theory).
- Also various AI planning methods for constructing the discourse tree.
Microplanning (how to say it)

- Microplanning (*sentence planning*) involves sentence-sized groupings of info in leaves of the discourse tree.
- Various issues:
  - *Sentence scoping*: Whether and how to combine two leaves.
    
    *I was hungry today. I went to Burger King.*
    
    *I was hungry today so I went to Burger King.*
  - *Pronominalization*.
  - *Aggregating*: Removing repetitions.
    
    *Smoking is bad for you. Smoking shortens your life. Smoking makes you smell bad.*
    
    *Smoking is bad for you, shortens your life, and makes you smell bad.*
  - *Thematization*.
  - *Topic ordering*.
- Special-purpose microplanners exist, but so far no general-purpose microplanner.
Realization (how to say it)

- Realization involves generation of surface level of text.
- Input is underlying deep-syntactic representation.
- Realizer needs lexicons, grammars, lexical ontologies.
- Output is grammatical, fluent (?) sentence.
## Types of realizers

- Canned text.
- Template-filling.
- Linguistics-based.
- Statistics-based.
Linguistic generators

- **FUF/SURGE (Elhadad, 1992):**
  - Input, lexicon, grammar in same notation.
  - Uses unification to merge all relevant info, produce sentence specification.

- **Penman/KPML (ISI/USC, late 1980s, Matthiessen and Bateman, 1991):**
  - Large interrelated grammar blends lexis, syntax, semantics.
  - Involves pragmatic functions of language: speaker attitude, relationship to intended audience.
Statistical generator

- **Nitrogen (Langkilde and Knight, 1998):**
  - Statistically-based generation.
  - Uses “overgenerate-and-prune” strategy.
  - Weak phrase structure grammar generates literally millions of possible sentences corresponding to input specification.
  - Statistical ranker chooses most fluent sequence at each choice point. Uses three-word window.
  - Determines fluency by consulting table of hundreds of millions of trigrams from Wall Street Journal.
The Virtual Human Project

http://www.youtube.com/uscict
What is missing?

- Fluent expressive text.
- Affect (emotion).
- Subtleties and nuances of language and gesture.
- Personalization.