Natural Language Generation

(http://www.dfki.de/~horacek/NLG.html)

Helmut Horacek

DFKI/Saarland University

DFKI, old building 1.16, Tel: 85775-2450

email:helmut.horacek@dfki.de

WHAT IS NL GENERATION?

Natural language generation is the process of

deliberatiely *constructing* a natural language text

out of available, *mostly non-linguistic* data

in order to meet specific *communication goals*

OBJECTIVES OF THE COURSE

- **State-of-the-art in NLG**
- **Overview of the more prominent NLG systems**
- **Current major issues in NLG research**

DECISION IN NATURAL LANGUAGE GENERATION

Deciding what to say

what the content of an utterance or a set of utterances should be

what information should be omitted

How to present this information effectively

how to organize that content in a coherent discourse

what tone or degree of formality should be adopted in the language used

how the material should be broken down into sentences and clauses

what syntactic constructions should be used

how entities should be described

what words should be used

AN EXAMPLE

Consider:

This course is being thaught by Helmut Horacek.
It is an introduction to natural language generation.

AN EXAMPLE

Consider:

This course is being thaught by Helmut Horacek.
It is an introduction to natural language generation.

This text embodies the following decisions:

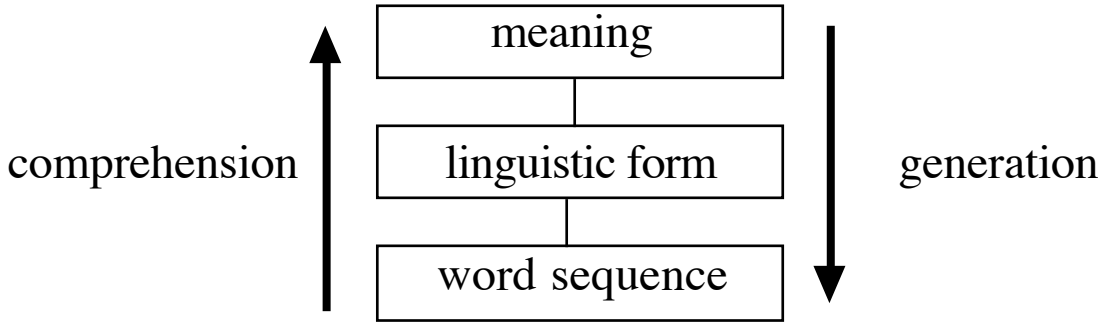
Of all the things known about the course, it states the lecturer's name and the topic of the course
It uses two simple sentences rather than one more complex sentence
Is uses a passive rather than an active sentence for the first piece of information
It uses the phrase *being thaught* rather than *being given*
It uses the pronoun *it* in the second sentence, in preference to a full noun phrase

The task of the NLG researcher is to explain

***how* and *why* particular choices are made**

DIFFERENCES BETWEEN NLG AND NLU

Is generation just the reverse of parsing?



Commonalities between NLG and NLU

some basic notion of a lexicon, using a taxonomy of basic word classes, word senses, and morphology

fairly shared notion of grammar as a means of describing the constructions available in a language

descriptions of various discourse phenomena – particularly anaphora – are important in both areas

FOCUS OF RESEARCH

NLU

The known is the text, perhaps with intonational information; the unknown is whatever the researcher chooses as a stopping point – typically some form of semantic representation with anaphors resolved

NLG

The known is the system's goals and intentions;
but at what level do you specify these?

This is why generation is in some sense more difficult. It leads to deeper thinking about the bigger picture, resulting in a view of language as a goal-oriented process, rather than simply as an information transferral process

NON-PROBLEMS IN COMPREHENSION

Deciding how much to say, and what not to say:

- **maintaining brevity**
- **avoiding stating the obvious**

Designing the text structure:

- **may need to add material to the basic subject matter**
- **controling the effects of the structure and ordering of the material**
- **making the text flow smoothly**

Problems in carrying out a detailed text plan once built:

- **determining the sentence boundaries and the use of conjunctions**
- **deciding when to use anaphora**
- **lexical selection**
- **use of marked syntactic structures for particular rhetorical effects**

COMPARING COMPREHENSION AND GENERATION

	Comprehension	*Generation*
the known	**the wording of the text**	**speaker's intentions, content/perspective selected**
primary effort	**to scan the text, during which its linguistic form and meaning gradually become apparent**	**choosing from alternatives, constructing specifications and then realizing them**
algorithms	**based on hypothesis management**	**planning by progressive refinement**
major problems	**ambiguity and underspecification**	**the process is oversupplied with source information and must decide what to highlight and what to omit**

PROBLEMATIC SITUATION OF NL GENERATION

- **Limited value of return**
  simple solutions often (almost) sufficient
- **Limited task agreement**
  initial specifications widely unconstrained
- **Limited systematicity**
  architecture, expressibility
- **Limited activity**
  as compared to work in analysis
- **Generation (sub)systems hard to evaluate**
  accepted metrics tend to penalize the use of “imperfect” methods

APPLICATIONS OF NL GENERATION

- **Report generation**
  (weather, business, …)
- **Flexible hypertext presentations**
  (museums, encyclopedic data, …)
- **Multi-lingual generation**
- **User- and context-adaptive presentations**
  (patient reports, …)
- **Multi-modal presentations**
- **Statistics-based generation**
  (machine-translation, function-driven applications, …)

SYSTEM ARCHITECTURE

Decomposition

- ***What* is said** **Course-graind planning (text planning)**
- ***When* it is said** **Fine-grained planning (sentence planning)**
- ***How* it is said** **Realization (syntactic generation)**

Interfaces of central importance

Precise decomposition into subprocesses unclear

Architectural models

- **Integrated – uniform, inefficient (historic)**
- **Sequential – practical, simplifying (currently the *standard* architecture)**

  However, no standards about the order of sentence planning tasks
- **Revision-based – theoretically best, hard to handle**

  Dedicated approaches according to demands of the genre

OVERVIEW

Content determination

Choosing and accommodating information

Document structuring

Ordering and rhetorical relations between pieces of content

Lexicalisation

Choice of words for pieces of content

Generating referring expressions

Descriptions of objects

Aggregation

Sentence constructions, compositions

Linguistic and structural realisation

Mapping specifications onto pieces of text