



Ontologies in Natural Language Generation

Catalina Hallett

ITRI, University of Brighton

catalina.hallett@itri.bton.ac.uk

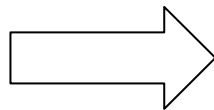
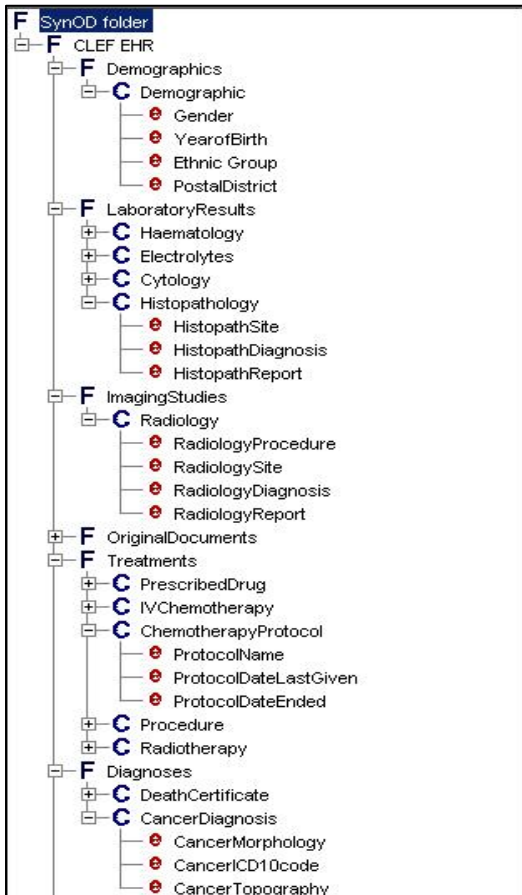


Overview

- Goals and architecture of NLG systems
- Use of domain ontologies in various components of NLG systems
- Shortcomings of domain ontologies: what does NLG really need?
- NLG in biomedicine
- An example: querying a repository of patient records using NLG techniques
- Conclusions

Natural Language Generation

Goal: to produce understandable and appropriate texts in a human language starting from some underlying non-linguistic representation of information



50 years old woman, diagnosed with breast cancer in 1999 at the age of 48. A fine needle biopsy of the axillary node was performed which demonstrated poorly differentiated adenocarcinoma.

PET scan 12/2001: PET reveals extensive cancer and patient begins chemotherapy. Third chemotherapy cycle given on 04/01/2002



Architecture

Document planning

Content determination

Document structuring

Microplanning

Lexicalisation

Aggregation

Generating referring
expressions

Surface realisation

Linguistic realisation

Structure realisation



Document planning

- Determine what information to communicate and means of communicating it
- Use ontologies for analysing the types of information available to the system
- Text is generated by expressing the content of the assertional component (A-box) using descriptions in the terminological component (T-box)

Micro-planning

- Converts a document plan into a sequence of phrase specifications (main words are selected, but not inflected, function words are not generated yet)
- Lexical variation for avoiding repetition and increasing fluency:
 - Using synonyms ("breast tumour"/"breast neoplasm"/"neoplasm of the breast")
 - Using is-a relations for generalization ("the CT scan"->"the scan", "the mastectomy"->"the surgical procedure")
- Provide constraints for syntactic aggregation rules:

[Patient has_disorder "pulmonary embolism"]
[Patient has_disorder "angina"]

} Patient has pulmonary embolism and angina

[Patient has_disorder "pulmonary embolism"]
[Patient has_feature "blue eyes"]

} Patient has pulmonary embolism and blue eyes (!)



Surface realisation

- The process of applying grammatical rules to produce a text which is morphologically and syntactically correct (words are inflected, function words are added, orthographical symbols introduced)
- At this stage, the meaning of the text has already been fully specified
- There is no further need for conceptual ontologies



Problems with domain ontologies

- Domain ontologies are most of the time too abstract (language-independent) to have a direct relationship to required forms of expressions
- NLG systems need to augment domain ontologies with semantic features before being able to generate natural language



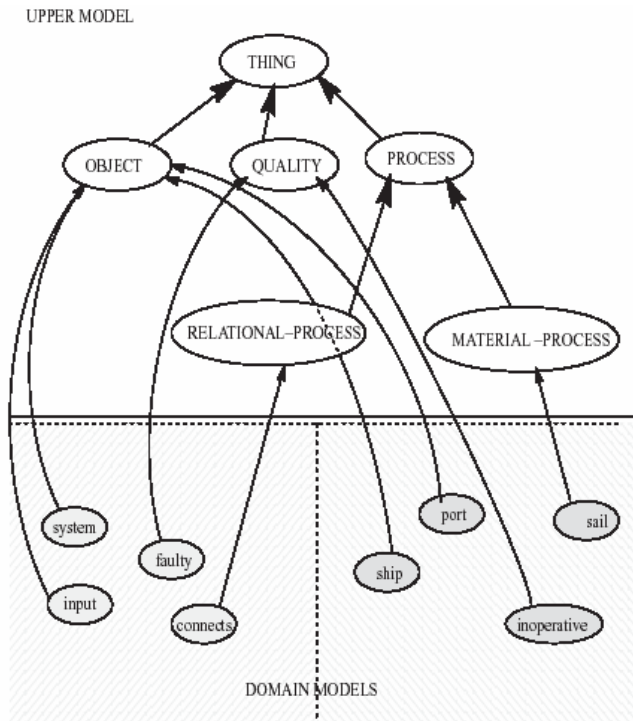
Types of ontologies

- Conceptual ontologies
 - “pure” language-independent ontologies describing the structure of a specific domain
- Linguistically-motivated ontologies
 - Provide a classification of any particular instance of facts, states of affairs, situations that occur in terms of a set of general objects and relations of specified types, that behave systematically with respect to their possible realizations
 - Concepts are classified according to their semantic function and choice of linguistic realisation
- Mixed ontologies
 - Combine domain knowledge and lexico-syntactical levels

Solutions

Common approaches:

- attaching links between concepts and lexical items
 - taking the lexical subsumption hierarchy and fitting this to a non-linguistic ontology
- > rigid and difficult to port to other domains
- Provide mapping rules from domain ontology concepts to linguistic ontology concepts



Bateman's Upper Model :

An interface ontology that is:

- abstract enough to allow domain switching
- constructed on linguistic principles to allow easy realization

Domain model concepts are classified in terms of the categories from the Upper Model ontology



NLG in bio-medicine

- Large number and range of applications:
 - Generating personalised written materials for health promotion: HealthDoc, STOP
 - Generating informative texts for patients based on their medical record: TIGGER, PERSIVAL (UMLS terms), PIGLET (handcrafted taxonomy based on Read; Grail)
 - Generating on-line personalised explanations for patients: MIGRAINE (UMLS-based KB)
 - Generating explanatory material for prescriptions
(de Rossis & Grasso, 1995)
 - Generating patient information leaflets for drugs: PILLS (UMLS)
 - Generating coherent critiques in initial definitive management of multiple trauma: TraumaGEN



Medical ontologies in NLG: the challenge

- Medical ontologies are poor in linguistic information
 - Some attempts to provide linguistic information: the UMLS SPECIALIST lexicon)
- Lack of concepts not directly part of the medical domain, but used in medical texts (e.g., colors, sizes, textures, agents)
- Lack of concepts representing processes (and relevant semantic constraints): diagnose, test, consult, perform
- Errors in classification may cause poor quality text



...

- Insufficient granularity (for linguistic realization purposes):
 - “cancer” and “headache” are both types of disorders
- ... but “a patient is diagnosed with cancer”, while “a patient complains of/suffers from headaches”



- NLG systems produce text tailored to the discourse consumer (different words for same concept)
- Ontologies of medical terms are highly specialised and the terminology is not adequate for presenting information to patients



Use of medical ontologies in practical NLG

- Medical ontologies seem to be used mostly as terminologies and thesauri
- Most relations in ontologies are ignored
- Most NLG applications do not make use of the inferential capabilities of ontologies



CLEF - Background

- The aim of the CLEF project is to create a scalable, generic architecture for capture and management of clinical data in the area of cancer
- Data is extracted from various types of document using Information Extraction techniques and stored in a central repository
- The repository can be accessed either through querying or by a direct summarisation or report request.
- Access to the CLEF repository is primarily intended for clinicians, medical researchers and other health professionals, in order to improve clinical care and postgenomic research.
- Our task is to provide NLG tools for:
 - easy and comprehensive querying of the HER
 - automatic generation of EPRs and summaries of EPRs in text format
 - Editing EPRs



CLEF – Querying the repository

Goals:

- Help formulating complex queries in natural language
- Restrict queries in such a way that they are meaningful and unambiguous

Method:

- users edit the conceptual content of a query by making choices over a feedback text
- the user starts by editing a simple query frame, where concepts to be instantiated are clickable spans of text (anchors)
- once a value for the concept represented by an anchor is selected, the system updates the semantic representation of the query and re-generates the feedback text
- Ontological constraints direct the user towards building semantically valid queries

CLEF Querying walkthrough

The screenshot shows a window titled "Query Patient Record" with a menu bar containing "File", "Options", "Help", and "Query PR". Below the menu bar, the status is "not connected". The main area contains the following text:

Assessment query

Relevant subjects: [Some type of patients][of a certain age description] diagnosed with mixed acidophil-basophil carcinoma [in some stage] of [some organ(s)]

[Treatment].

[Outcome].

At the bottom, there is a "Status:" label followed by a text box containing "Ready" and a "Submit" button.



CLEF Querying and ontologies

- Taxonomies for providing key query concepts to be selected by the user (several taxonomies used according to the type of data encoding in the repository: ICD-10, SNOMED, UMLS)
- Concept relations for restricting the user choice
 - Once the user has selected the value "acral lentiginous melanoma, malignant" for the anchor "tumour_name", the list of available choices for the concept "tumour_locus" is restricted to ["Skin, face", "Skin, trunk", "Eyelid NOS", "Skin, limb", "upper External ear", "Skin, scalp,neck", "Skin limb, lower", "Skin lip, NOS"]
- Problems:
 - Mapping domain concepts into lexical forms: general rules that map, for example, diagnoses into nouns
 - Choosing the appropriate realisation for domain concepts:
 - Producing "surgery on the pancreas" and "surgery on a leg" requires knowledge about the fact that there is just one pancreas, but more than one leg
 - At the moment, all domain concepts in the same class are treated uniformly



Conclusions

- NLG applications are dependent on domain specific taxonomies and terminologies
- However, most domain ontologies are too language-independent to be directly usable in an NLG system
- There is no practical evidence that deep domain ontologies are more beneficial to NLG applications than simple taxonomies