CENTRE FOR LANGUAGE TECHNOLOGY
MACQUARIE UNIVERSITY ~ SYDNEY

Roadmapping for Natural Language Generation

Robert Dale

rdale@ics.mq.edu.au
www.clt.mq.edu.au

Underlying Premise

- The problem: Current NLG research delivers solutions that are looking for problems
- The disconnect: areas where NLG might be used but isn't:
  - Spoken language dialog systems
  - Text summarisation systems
  - Machine translation systems
  - Grammar-checking systems
- Consequence: NLG needs a phased series of realistic outcomes that demonstrate the value of the technology

#1: A standardised architecture for summarising tabular data structures in a specific domain

- Basic idea: One of the most obvious areas where the linguistic sophistication of NLG techniques can be demonstrated is in the use of aggregation to provide concise descriptions of sets of similar or related facts. A common source of such facts is in tables.
- Outcome by 2007: the development of an API that enables generation of texts from 80% of the simple tables that appear in a widely used domain, such as financial reporting. Likely to be available as a plug-in for a product such as Microsoft Excel.

#2: Extension of table summarisation to a wide range of domains and multiple languages

- Basic idea: The success of the subgoal #2 would provoke the development of similar technologies and techniques for other domains and languages.
- Outcome by 2008: This subgoal would likely result in tabular summarisation being available in five major European languages, plus Japanese and Mandarin, in three other high value domains.

#3: A rich markup language that enables high level control of the prosody in text to speech

- Basic idea: We need to go beyond standards like SSML.
- Outcome by 2007: Higher-level control of prosody that SSML provides, and hooks that can be used appropriately by concept to speech systems.

#4: Syntactic smoothing of sentence-extraction based summarisation

- Basic idea: NLG makes it possible to produce smoother summaries by reconstructing sentences from parts of sentences.
- Major outcome by 2008: one or more products on the market that produce appreciably improved summaries of input documents.

#5: Shallow Semantic Summarisation

- The aim: to improve the quality of output that is possible by introducing a more sophisticated approach to the analysis of the source text.
- Basic idea: the quality of summarisation will be improved if the text reconstruction mechanism has some idea of the meaning of the text, even if only at a superficial level.
- Major outcome by 2010: market leadership of a technology that improves upon the products deriving from subgoal #4, at least in some high-value domains.

#6: A standardised architecture for adding natural language generation capabilities to relational databases

- Basic idea: as we begin to see useful results in generating, for example, summaries of information in spreadsheets, more complex underlying datasets will begin to look worth attacking.
- Major outcome by 2009: We might expect the outcome here to be the provision of plug-ins by major database vendors such as Oracle that provide NLG reporting and summarisation functionalities for databases in a range of supported domains, probably based on the development of relevant XML-based standards.

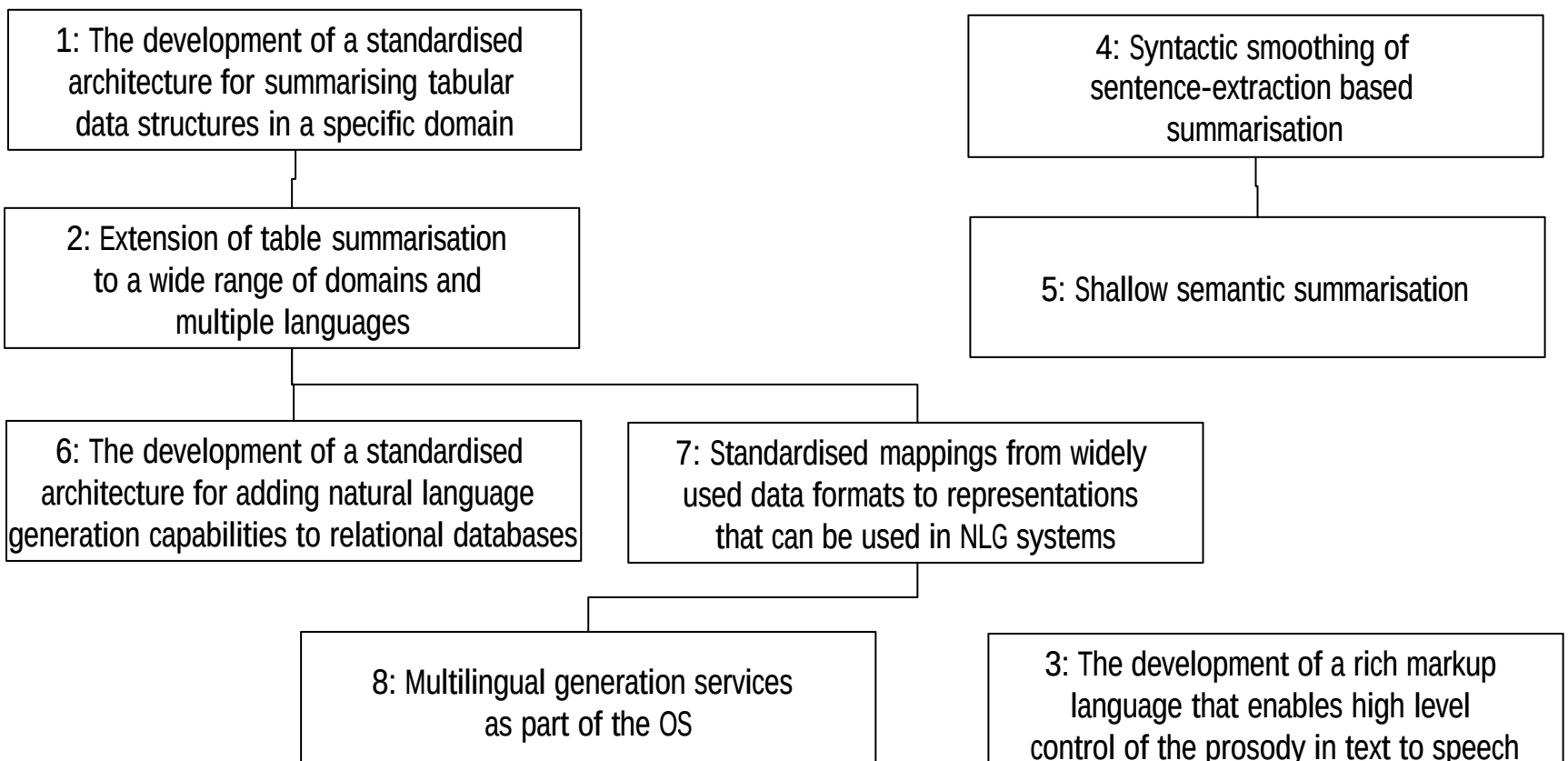
#7: Standardised mappings from widely used data formats to representations that can be used in NLG systems

- Basic idea: while database vendors will be interested in how they can make the contents of databases more accessible, the vendors of desktop office productivity applications will have a similar concern for their applications.
- Outcome by 2009: the development of a level of representation that can be used in conjunction with NLG technologies to provide such outputs.

#8: Multilingual generation services as part of the OS

- Basic idea: As the benefit of NLG technologies here is appreciated and as the technology becomes better understood, we can expect to see the services required become part of the underlying operating system.
- Major outcome by 2011: a widely understood NLG API that can be used by program developers to provide multilingual NLG reporting and output facilities in their applications.

The Subgoals



Dale's Subgoals

1	A standardised architecture for summarising tabular data structures in a specific domain	2007
2	Extension of table summarisation to a wide range of domains and multiple languages	2008
3	A rich markup language that enables high level control of prosody in TTS	2007
4	Syntactic smoothing of sentence-extraction based summarisation	2008
5	Shallow semantic summarisation	2010
6	A standardised architecture for adding NLG capabilities to relational DBs	2009
7	Standardised mappings from widely used data formats to representations that can be used in NLG systems	2009
8	Multilingual generation services as part of the OS	2011

Reiter's Subgoals

1	Experimental evaluation methodology for NLG	2006
2	Empirical lexicons	2007
3	Text Summaries of Complex Data	2009
4	Personal simplified web pages	2014

Compatibility

		Experimental evaluation methodology for NLG	2006
		Empirical lexicons	2007
A standardised architecture for summarising tabular data structures in a specific domain	2007		
A rich markup language that enables high level control of prosody in TTS	2007		
Extension of table summarisation to a wide range of domains and multiple languages	2008		
Syntactic smoothing of sentence-extraction based summarisation	2008		
A standardised architecture for adding NLG capabilities to relational DBs	2009		
Standardised mappings from widely used data formats to NLG representations	2009	Text Summaries of Complex Data	2009
Shallow semantic summarisation	2010		
Multilingual generation services as part of the OS	2011	Personal simplified web pages	2014