Content determination

NLG process division

1. Document planning
2. Microplanning
3. Surface realization

Further NLG process division

1. Document planning
   - Content determination
   - Document structuring
2. Microplanning
3. Surface realization

Content determination

- Content determination phase is responsible for deciding, what domain information should the output contain.
- Content determination is highly connected to application domain. The information available, and the genre of the intended output.
- Where as other parts of the NLG pipeline are easier to generalize, content determination is harder to do so.

Target text corpus

Example expert created input and output for the type of texts, which the NLG tries to target.

NLG's input

A **{k,c,u,d}** four-tuple:

- **Knowledge source** - domain information in encoded form (db, etc)
- **Communicative goal** - purpose of selected textual output properties, instance of typed process. Part of general goals. Can be split again into data comprehension goal.
- **User model** - model of audience; what user knows, how user interacts with the NLG’s parent system, etc
- **Discourse history** - current and previous outputs, and user interaction

Content determination process

- During content determination task, messages are created from knowledge source. Then the found messages are pruned, based on which of them satisfy the requirements of the communicative goal. Not all messages are wanted.
- Selecting, summarizing and reasoning with data.
- Essentially requires doing some kind of data analysis, and data mining, in order to obtain the interesting summarization of the data. Also being used: signal-processing techniques & complex planning algorithms.
- Content determination is usually based on the domain model mapping of domain elements. What entities, attributes, relationships and classes is wanted to pick from the data.

Document plan

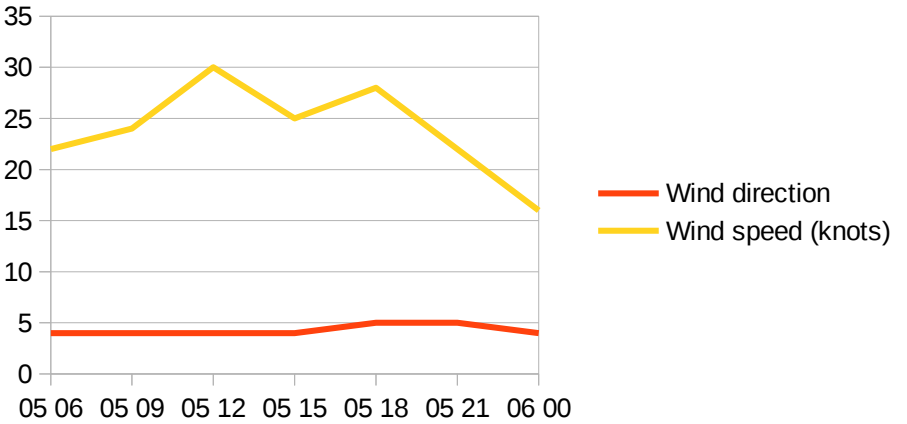
- Output of the document planner.
- A Structured tree. (from document structuring task)
- Content in leafs. These are called **messages**. (from content determination task)
- A message resembles a single content leaf in the tree.
- It's possible that a single document plan message maps to a single sentence in the final output of NLG, but in many cases, messages are smaller pieces of distinct data objects.

Domain model

Reiter & Dale's WeatherReporter:

- Time-span as entity
- Rainfall, wind, mist-and-fog as attribute
- Months and days as instances
- Possible relationships between the elements.
- Each instanced configuration of these domain elements is a message.
- Problem: What elements to map together in a configuration?
  - This needs to be somehow mapped from the corpus (= example inputs and outputs).

Time (day/hour)	Wind Direction	Wind Speed (Knots)
05/06	SE	22
05/09	SE	24
05/12	SE	30
05/15	SE	25
05/18	SSE	28
05/21	SSE	22
06/00	SE	16



[windSpell:

[timeSpan: 05/06 – 06/00]

[windSpeed: 16 – 30, average: xy,

direction: SE, 05/18 – 05/21: SSE]]

- **Selecting**
- Needs to reason what information is statistically significant, what is important and descriptive of the communicated domain element.
- **Summarizing**
- In what form the information is meaningful to be shown.

[windSpell:
[timeSpan: 05/06 – 06/00]
[windSpeed: 16 – 30, average: xy,
direction: SE, 05/18 – 05/21: SSE]]

References

- Book: Building Natural Language Generation Systems, Reiter & Dale (2000)
- Article: Building Natural Language Generation Systems, Reiter (1996)
- Article: A two-stage model for content determination, Sripada et al (2001)
- Article: SUMTIME-MOUSAM: Configurable Marine Weather Forecast Generator, Sripada et al (2003)