Week 4: NLG Pipeline for weather reporting Lexicalization and Aggregation

MYRIAM MUNEZERO
7TH FEBRUARY 2017
Outline

- **Forecast Generator**
- **Lexical choice in SumTime-Mousam**

**Refresher**

- **lexicalization**
  - choosing which words should be used to express non-linguistic data
- **aggregation**
  - deciding how to distribute information among sentences (for example, many short sentences or a few longer sentences)
FOG – Forecast Generator
Uses rules and natural-language generation to convert weather maps into forecast text

Introduction

- FOG has been producing marine forecasts since 1991-2
  - Handles 4 other marine type forecasts and one public forecast

- The system is composed of a Text Planner and a Text Realizer
  - Linguistic processing (includes lexicalization and aggregation), takes place over both modules.
Linguistic Processing Technique

"Winds southwest 15 to 20 knots diminishing to light late this evening."

Table 1. Time-merging data to concepts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Sample Data</th>
<th>Wind Direction</th>
<th>Wind Speed</th>
<th>Concepts</th>
<th>Wind Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 a.m.</td>
<td>223</td>
<td>13</td>
<td>southwest</td>
<td>15–20</td>
<td>↓</td>
</tr>
<tr>
<td>7 a.m.</td>
<td>235</td>
<td>17</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>9 a.m.</td>
<td>231</td>
<td>21</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>9 p.m.</td>
<td>280</td>
<td>12</td>
<td>(west)</td>
<td>light</td>
<td>↓</td>
</tr>
<tr>
<td>10 p.m.</td>
<td>307</td>
<td>11</td>
<td>(northwest)</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>11 p.m.</td>
<td>182</td>
<td>8</td>
<td>(south)</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Midnight</td>
<td>246</td>
<td>10</td>
<td>(southwest)</td>
<td>↓</td>
<td>↓</td>
</tr>
</tbody>
</table>

Data

Text output
Text planner

- Builds concepts to describe transitions between states
  - The planner chooses to use the concept 'diminish' because of the transition from "southwest 15 to 20" to "light winds"
- Then groups and structures the conceptual content into sentence-sized chunks, and outputs the intelingua.
- Wind events are cast into a dependency trees that represents one sentence.

<table>
<thead>
<tr>
<th>Time</th>
<th>Sample Data</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wind direction</td>
<td>Wind speed</td>
</tr>
<tr>
<td>6 a.m.</td>
<td>223</td>
<td>13</td>
</tr>
<tr>
<td>7 a.m.</td>
<td>235</td>
<td>17</td>
</tr>
<tr>
<td>9 a.m.</td>
<td>231</td>
<td>21</td>
</tr>
<tr>
<td>9 p.m.</td>
<td>250</td>
<td>12</td>
</tr>
<tr>
<td>10 p.m.</td>
<td>307</td>
<td>11</td>
</tr>
<tr>
<td>11 p.m.</td>
<td>182</td>
<td>8</td>
</tr>
<tr>
<td>Midnight</td>
<td>246</td>
<td>10</td>
</tr>
</tbody>
</table>
In the intelingua, nodes = names of abstract concepts
- Points to English and French lexical entries

Arcs = indicate the dependency types
- I and II represent a predicate's first and second arguments
- ATTR = the attributive relation.
Text realizer

- Transforms the Interlingua into English and French text
  - Each language requires its own grammar and dictionary: language model

- Makes use of Meaning <-> Text theory as the language model
  - Based on a large structured dictionary (lexicon)
  - Set of correspondence rules that provide mappings between several representation levels

- While MTT has 7 representation levels, FOG uses 5:
  - Intelingua
  - Two syntactic levels (deep and surface)
  - A morphological level
  - The final text level

With MTT’s language model:

- All dependency tree nodes and leaves are lexical items
- Linked by oriented dependency relations
- Grammatical relations between words are explicitly expressed by labels on the arcs of linking the words
- A constituent of the sentence is a word and all the words that hang from it.
Taking the interlingua structure, the text realizer replaces the abstract interlingual concepts (nodes with like ‘wind’, ‘diminish’) with language-specific equivalents from its lexicon to obtain the deep syntactic representations.
So 'diminish' becomes DIMINISH, and DIMINUER

The lexicon stores the rules that govern this replacement

- DIMINISH's lexical entry contains a rule that states that DIMINISH can replace 'diminish' if the interlingua node has 'wind' or 'gale' as dependent II.

- If dependency tree involved temperatures, then the above rule would not be satisfied, and the 'diminish' node would become the verb LOWER

The nodes are labeled with full lexemes

- Each lexeme has its syntactic class and morphosyntactic features:
  - cn = common noun, prep = preposition, etc.
Text realizer cont.

- Makes the transition from deep syntax to surface syntax
  - New nodes, e.g., auxiliary lexemes, and new dependency relations might be produced.
  - For example the addition of TO with DIMINISH, but not with BECOMING
    - “Winds southwest 15 to 20 knots **becoming** southeast 15 late this evening.”
    - “Winds southwest 15 to 20 knots **diminishing to** light late this evening.”

- The transformation from surface syntactic representation to morphological representation involves linearizing the surface syntactic tree and propagating grammatical features from head words to dependents for gender agreement in French and number agreement in both languages.

- The realizer then produces text using information from the lexicon. It takes each lexeme’s surface form directly from the lexicon, or computes it according to the root form and the morphological features and description.
Lexical Choice in SumTime-Mousam

Introduction

- SumTime-Mousam – NLG system that generates weather forecast texts from numerical weather prediction data.

- SumTime-Mousam generates texts in three stages: Document planning, microplanning, and surface realization.

- Problem: Used a consistent set of data-to-word rules, which avoided words that only occurred in a few idiolects and words that had idiolect-dependent meanings.
Lexical choice problem

- Take for example:
  - "W 8–13 backing SW by mid afternoon and S 10–15 by midnight."

- Generating the text requires making several word choices, e.g.,
  - Direction: should West be expressed as W or W’LY?
  - Speed: should 8 knots be expressed as 8 or 08?
  - Verb: should backing or becoming be used to describe the change in wind direction?
  - Time phrase: should by evening, by late evening, or by midnight be used to express the time 0000?
Lexical choice rules technique

- Developed lexical choice rules for time phrases, verbs, numbers, directions, connectives, and adverbs, in wind descriptions.

- First analyze how people write time phrases, verbs, etc.
  - collected and analyzed a corpus of human written forecasts, parsed the forecast into phrases and aligned each phrase to a corresponding data file.

- Learned a classifier
  - Used the machine learning algorithm C4.5 (as implemented in Weka’s J4.8 classifier) to learn classifiers.
  - The classifiers, for instance predicted which time phrase would be used in wind phrases extracted from the corpus;
  - in other words, the class being predicted by the classifier was by evening, by midday, and so forth
  - The classifier was trained only on phrases which had been successfully aligned with the data file, e.g., 2359 phrases that referred to a known time.
Lexical choice rules technique cont.

- Also experimented with giving it other features; like author, collocation, surface, temporal feature, etc.

- Used as a baseline a classifier which always chose the most common phrase for a time; for example, it always chose by midday for 1200.
  - This classifier had a 67% error rate (all error rates are calculated using 10-fold cross-validation).

- The lowest error rate, their J48 classifier achieved with their set of features was 48%
Their method was a way to avoid the choice of words which only occurred in one idiolect and words whose meanings varied in different idiolects.
Lexicalization and aggregation in weather forecast NLG systems can be combined or done both in the planner and realizer.

Machine learning can be used for learning rules for lexical choice.