Statistical Generation 2: Trainable Sentence Planning for Spoken Dialogue Systems

Lecture 13
March 22, 2013


Example: MATCH Multimodal Dialogue System

- Function: Provides information about restaurants in New York City
- Input:
  - User query: Typed and spoken language, gesture
  - User model
  - Restaurant database
- Output: Spoken, written and graphical output
- Developer: AT&T Research Labs
- Status: Research Prototype


“Show me Italian restaurants in the West Village.”
User can ask system to summarize, compare, or recommend!

User-Tailored Generation

- User model helps determine entities and attributes to include
  - Don’t mention options that rank low according to the user model
  - Don’t mention attributes the user doesn’t care about
- User model affects organization of content
  - Mention highest-ranking options first
  - Mention attributes that contribute significantly to rank of option first
  - Mention features user cares about first
- Evaluation of MATCH and other systems indicates user tailored generation leads to improved:
  - User satisfaction
  - Task efficiency
  - Task effectiveness

(Walker et al., 2004; Carenini and Moore, 2006)
**Example User Models**

- **CK** considers food type and food quality to be important:
  - \( U(\text{restaurant}) = 0.41V(\text{FoodQuality}) + 0.24V(\text{FoodType}) + 0.16V(\text{Cost}) + 0.10V(\text{Service}) + 0.06V(\text{Neighborhood}) + 0.03V(\text{Decor}) \)

- **OR** considers cost to be most important, likes many food types:
  - \( U(\text{restaurant}) = 0.41V(\text{Cost}) + 0.24V(\text{FoodQuality}) + 0.16V(\text{Decor}) + 0.10V(\text{Neighborhood}) + 0.06V(\text{Service}) + 0.03V(\text{FoodType}) \)

**Recommendations**

- **Recommend**
  - restaurant with highest overall user-model score.
  - mention attributes that contribute significantly to high score

- **Example**:
  - **CK**: Babbo has the best overall value among the selected restaurants. Babbo’s price is 60 dollars. It has superb food quality, excellent service and excellent decor.
  - **OR**: Uguale has the best overall value among the selected restaurants. Uguale’s price is 33 dollars. It has good decor and very good service. It’s a French, Italian restaurant.

**Comparison for users CK and OR**

- **CK**: Among the selected restaurants, the following offer exceptional overall value. Babbo’s price is 60 dollars. It has superb food quality, excellent service and excellent decor. Il Mulino’s price is 65 dollars. It has superb food quality, excellent service and very good decor. Uguale’s price is 33 dollars. It has excellent food quality, very good service and good decor.

- **OR**: Among the selected restaurants, the following offer exceptional overall value. Uguale’s price is 33 dollars. It has good decor and very good service. It’s a French, Italian restaurant. Da Andrea’s price is 28 dollars. It has good decor and very good service. It’s an Italian restaurant. John’s Pizzeria’s price is 20 dollars. It has mediocre decor and decent service. It’s an Italian, Pizza restaurant.

**Requirements for NLG in Spoken Dialogue**

- High quality generation in domain
- Efficient generation
- Flexible generation
Approaches to Generation in Spoken Dialogue

- **Template-based generation**
  - Conceptually simple
  - Tailored to domain — quality often high
  - Must create templates for each application
  - Tailoring greatly increases number of templates needed
  - Must repeatedly encode linguistic constraints
  - Difficult to extend/maintain

- **Natural language generation**
  - Portable, general
  - Tailoring easily supported
  - Quality within a domain may be poorer
  - Can be inefficient
  - Linguistic expertise required

Trainable Generation

- Train NLG modules automatically
  - Supervised learning using user ratings of text quality

- Benefits:
  - Speed of NLG module engineering
  - Requires less linguistic and domain expertise
  - Clear method for adaptation

- Open questions:
  - Does trainable generation work well for flexible generation tasks?
  - How does the output quality compare to that of template generation?

Content Plan for a Recommendation

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Recommend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Bar Pitti, Arlecchino, Babbo, Cent'anni, Cucina Stagionale, Grand Ticino, Il Mulino, John's Pizzeria, Marinella, Minetta Tavern, Trattoria Spaghetto, Vittorio Cucina</td>
</tr>
<tr>
<td>Relations</td>
<td>justify(nuc1; sat:2) justify(nuc1; sat:3) justify(nuc1, sat:4)</td>
</tr>
<tr>
<td>Content</td>
<td>1. assert(best (Babbo)) 2. assert(has-att (Babbo, food quality(superb))) 3. assert(has-att (Babbo, decor(excellent))) 4. assert(has-att (Babbo, service(excellent)))</td>
</tr>
</tbody>
</table>

Problem: How to Choose A Good Content Organization?

One content plan, multiple text plans

1. Babbo has superb food quality
2. Babbo has excellent service
3. Babbo has excellent décor
4. Babbo is the best
6. Chanpen Thai has the best overall quality among the selected restaurants since it is a Thai restaurant, with good service, its price is 24 dollars, and it has good food quality.

8. Chanpen Thai is a Thai restaurant with good food quality. It has good service. Its price is 24 dollars. It has the best overall quality among the selected restaurants.

Solution: Trainable Sentence Planning

- SPaRKy (Sentence Planning with Rhetorical Knowledge)
  - trainable sentence planner for information presentation in MATCH multi-modal dialogue system
- Two-stage approach to sentence planning
  - Sentence plan generator (SPG) generates possible sentence plans from text plans
  - Sentence plan ranker (SPR), which is trained on human judgments, ranks sentence plans
- Used for complex user-tailored presentations
  - Recommendations, comparisons

Sentence Plan Generation

- Input: Set of text plan trees
- Output: A set of sentence plan trees, each with an accompanying dependency tree
- Steps:
  1. Group assertions in content plan using principles from Centering Theory
     - Group assertions that talk about the same thing, e.g., about same restaurant, or same attribute
  2. Use 6 (domain-independent) clause combining operations to assign assertions to sentences and insert discourse cues
     - Chosen randomly according to a probability distribution
  3. Generate referring expressions
     - proper names replaced by pronouns based on recency
Clause Combining Operations: Examples

- **Merge:** (contrast, infer)
  - Babbo has superb décor AND Babbo has mediocre food quality \(\Rightarrow\) Babbo has superb décor and mediocre food quality.

- **Relative-clause:** (infer, justify)
  - Baluchi’s has the best overall quality among the selected restaurants AND Baluchi’s is located in uptown Manhattan \(\Rightarrow\) Baluchi’s, which is located in uptown Manhattan, has the best overall quality among the selected restaurants.

- **Cue-word-conjunction but:** (contrast, infer, justify)
  - Above has decent décor AND Carmine’s has good décor \(\Rightarrow\) Above has decent décor but Carmine’s has good décor.

- **With-reduction:** (infer, justify)
  - Above is an Italian restaurant AND Above has good décor \(\Rightarrow\) Above is an Italian restaurant with good décor.

Sentence Plan Tree for One Recommend Alternative

Content Plan: Recommend(Chanpen Thai)

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Recommend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relations</td>
<td>justify(nuc1; sat:2); justify(nuc:1; sat:3); justify(nuc:1, sat:4); justify(nuc:1, sat:5)</td>
</tr>
</tbody>
</table>
| Content   | 1. assert(best (Chanpen Thai))  
            2. assert(is (Chanpen Thai, cuisine(Thai)))  
            3. assert(has-att (Chanpen Thai, food-quality(good)))  
            4. assert(has-att (Chanpen Thai, service(good)))  
            5. assert(is (Chanpen Thai, price(24 dollars))) |
6. Chanpen Thai has the best overall quality among the selected restaurants since it is a Thai restaurant, with good service, its price is 24 dollars, and it has good food quality.

8. Chanpen Thai is a Thai restaurant with good food quality. It has good service. Its price is 24 dollars. It has the best overall quality among the selected restaurants.

### Features for Sentence Plan Ranking

- Represent a declarative encoding of the decision in context
- N-gram features (1-3)
  - Information about lexical selection and ordering
  - Replace names with types, e.g., Babbo with RESTNAME
- Concept features
  - Concept (1-3)-grams generated from named entities labelled on the SPG outputs, e.g., CONC-DÉCOR-CLAIM = 1 if claim is expressed after decor
- Tree features
  - Count structural configurations in the sentence plans and dependency trees
  - Types of tree feature:
    - Ancestor
    - Preorder traversal
    - Sister
    - Leaf
    - Global

### Example Tree features

- CW_BECAUSE_NS_justify
- assert-reco-best
- CW_CONJUNCTION_infer
- WITH_NS_infer
- assert-reco-food_quality
- assert-reco-service
- assert-reco-decor

### Exp 1: Which features are best predictors?

- **Method:** 10-fold cross-validation
  - Repeatedly train SPR on 90% of the corpus of labeled sentence plan trees, test on remaining 10%
- **Results**
  - Using ALL features produces best results, but not always statistically significant
  - N-gram features as good as ALL for COMPARE-2 and RECOMMEND
  - **Why?**
    - Hypothesis: individual lexical items are uniquely associated with many of the combination operators
    - E.g., “with” for WITH-NS operator
    - N-gram features equivalent to tree features for this domain

### Human vs. SPR scores

<table>
<thead>
<tr>
<th>Realization</th>
<th>Human (AVG)</th>
<th>RankBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babbo has the best overall quality among the selected restaurants because</td>
<td>1.5</td>
<td>0.45</td>
</tr>
<tr>
<td>it has superb food quality, with excellent service, and it has excellent decor.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Babbo has excellent service. It has superb food quality. It has excellent decor.</td>
<td>2</td>
<td>0.21</td>
</tr>
<tr>
<td>It has the best overall quality among the selected restaurants.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Since Babbo has excellent service and superb food quality, with excellent decor, it has the best overall quality among the selected restaurants.</td>
<td>3.5</td>
<td>0.77</td>
</tr>
<tr>
<td>Babbo has excellent service and superb food quality, with excellent decor.</td>
<td>4</td>
<td>0.88</td>
</tr>
<tr>
<td>It has the best overall quality among the selected restaurants.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>With excellent decor, excellent service and superb food quality, Babbo has the best overall quality among the selected restaurants.</td>
<td>5</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Performance of SPR

- **Evaluation:**
  - Exp 2: Can SPaRKy select a high quality sentence plan from set of randomly generated sentence plans?
  - Exp 3: How does the output from SPaRKy compare with the output from a template-based generator?

**Experiment 2**

- **Method:** 2-fold cross-validation
  - Repeatedly train SPaRKy on randomly selected 50% of corpus of labeled sentence plan trees, test on remaining 50%
  - Evaluate SPaRKy on test set by comparing 3 data points for each content plan:
    - SPaRKy -- score of SPR’s top-ranked sentence plan
    - HUMAN -- score of the sentence plan rated highest by human judges
    - RANDOM -- score of a randomly selected sentence plan

**Experiment 2: Results**

<table>
<thead>
<tr>
<th>Strategy</th>
<th>System</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommend</td>
<td>SPaRKy</td>
<td>3.6</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>HUMAN</td>
<td>3.9</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>RANDOM</td>
<td>2.9</td>
<td>0.88</td>
</tr>
<tr>
<td>Compare-2</td>
<td>SPaRKy</td>
<td>3.9</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>HUMAN</td>
<td>4.4</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>RANDOM</td>
<td>2.9</td>
<td>1.30</td>
</tr>
<tr>
<td>Compare-3</td>
<td>SPaRKy</td>
<td>3.4</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>HUMAN</td>
<td>4.0</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>RANDOM</td>
<td>2.7</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- For all three information presentation types
  - HUMAN significantly better than SPaRKy (paired t-test, p < .001)
  - SPaRKy significantly better than RANDOM (paired t-test, p < .001)
  - SPaRKy can generate high quality output from a random set of sentence plans

**Experiment 3**

- **Method:** For each content plan, compare
  - SPaRKy -- score of SPR’s top-ranked sentence plan
  - HUMAN -- score of sentence plan rated highest by the human judges
  - TEMPLATE -- human rater score of sentence plan produced by template-based generator used in MATCH system

(Walker et al., Cognitive Science, 2004)
Experiment 3: Results

<table>
<thead>
<tr>
<th>Strategy</th>
<th>System</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommend</td>
<td>TEMPLATE</td>
<td>4.2</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>SPaRKy</td>
<td>3.6</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>HUMAN</td>
<td>4.4</td>
<td>0.37</td>
</tr>
<tr>
<td>Compare-2</td>
<td>TEMPLATE</td>
<td>3.6</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SPaRKy</td>
<td>3.9</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>HUMAN</td>
<td>4.6</td>
<td>0.39</td>
</tr>
<tr>
<td>Compare-3</td>
<td>TEMPLATE</td>
<td>4.1</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>SPaRKy</td>
<td>3.4</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>HUMAN</td>
<td>4.6</td>
<td>0.35</td>
</tr>
</tbody>
</table>

- HUMAN significantly better than TEMPLATE only for COMPARE-2
- TEMPLATE significantly better than SPaRKy for RECOMMEND and COMPARE-3
- SPaRKy better than TEMPLATE for COMPARE-2 (trend)
  - Human raters did not like template for COMPARE-2

Type of Rules Learned

- If leaf_#assert-reco-best > 0 then increase ranking by 0.5 => Put recommendation before supporting information
  - Babbo has the best overall quality among the selected restaurants because it has good service.
  - Because Babbo has good service it has the best overall quality among the selected restaurants.
- rule-anc-assert-com-price*CW_CONJUNCTION-infer*PERIOD-justify > -infinity, then increase ranking by .53 => Justifications involving price should be merged with other information using a conjunction
  - Le Madeleine has the best overall quality among the selected restaurants. It has very good food quality and its price is 40 dollars.
  - Le Madeleine has the best overall quality among the selected restaurants. It has very good food quality. Its price is 40 dollars.

Summary

- SPaRKy, a trainable sentence planner for complex information presentations in spoken dialogue
- Trainable sentence planning can produce
  - high quality output
  - with less human effort than rule-based NLG
- Gap between HUMAN scores and TEMPLATE scores indicates
  - SPG produces sentence plans as good as those of template generator
  - Accuracy of SPR needs to be improved

References

RankBoost Algorithm (Schapire 99, Iyer et al. 98)

- Each example $x$ represented as sum of $m$ indicator functions:
  \[ h_s(x) = 1 \text{ if rule-sisters_{MERGE}_infer} > 1, \text{ else } 0 \]
- Each function $h_s(x)$ has single $\alpha_s$ parameter
- Ranking Score: $F(x) = \sum_s \alpha_s h_s(x)$
- Score used to rank competing sp-trees for same text plan
- Training data is a set of pairs $(x,y)$ for each example $x$ rated higher than $y$ (so 20 candidates could generate as many as 20*19/2 training pairs)
- Training: set the parameters $\alpha_s$ to minimize the loss function
  \[ \text{Loss} = \sum_{(x,y)} e^{- (F(x) - F(y))} \]
- As Loss is minimized, $(F(x) - F(y))$ where $x$ is preferred to $y$ is pushed to positive and ranking errors will tend to be reduced