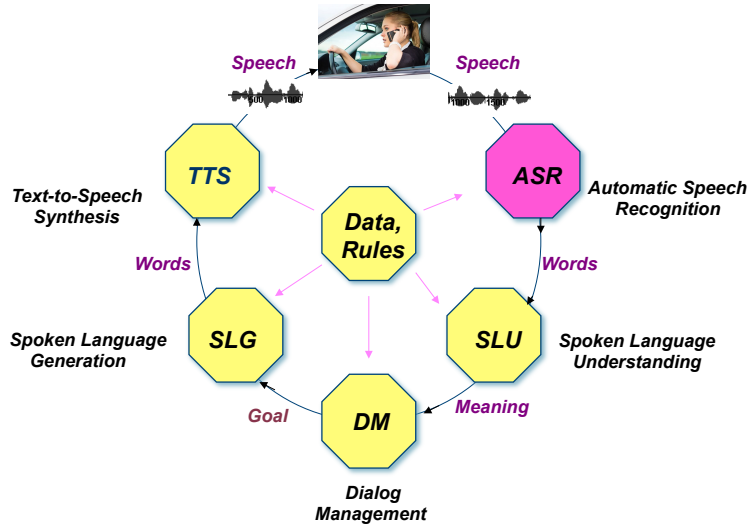
Statistical Generation 2-3: Trainable Sentence Planning for Spoken Dialogue Systems

Lecture 13 & 14
March 22,26 2013

Reading: M. Walker, A. Stent, F. Mairesse, and R. Prasad (2007). "Individual and Domain Adaptation in Sentence Planning for Dialogue." *Journal of Artificial Intelligence Research* **(30):** 413-456.

Anatomy of a Spoken Dialogue System

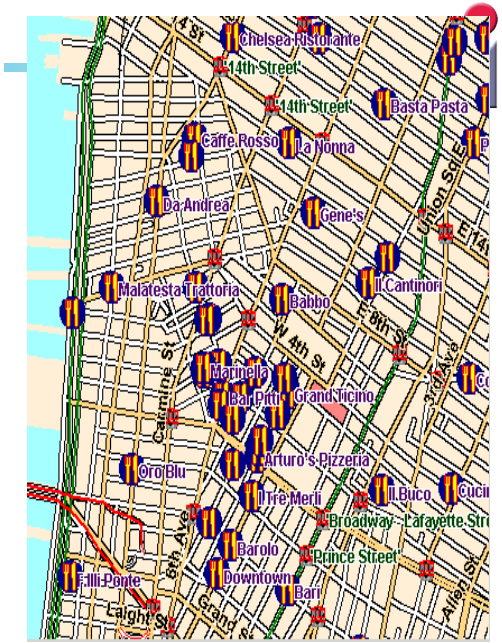


Example: MATCH Multimodal Dialogue System

- 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

Johnston, M., et al., "MATCH: An Architecture for Multimodal Dialogue Systems", ACL 2002

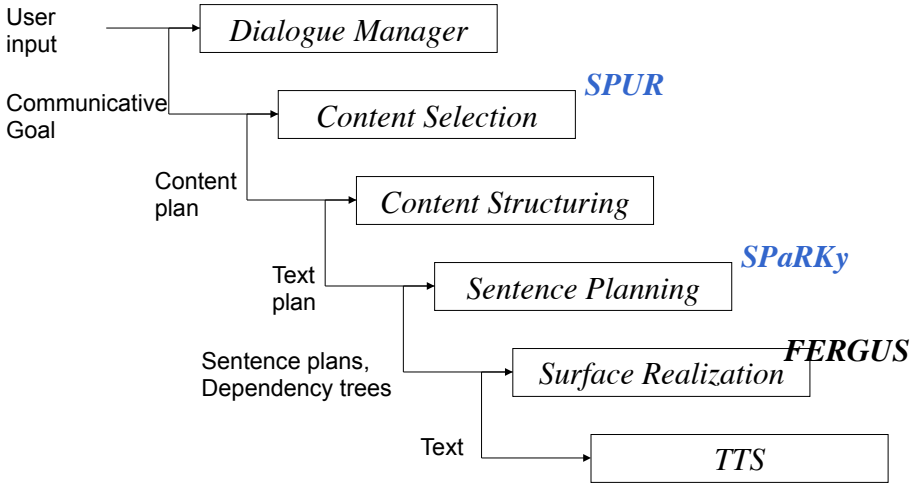
"Show me Italian restaurants in the West Village".



USER: *Summarize (with appropriate pen gesture to select on map).*

User can ask system to summarize, compare, or recommend

Spoken Language Generation in MATCH



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 (restaurant) = .41 V(FoodQuality) + .24 V (FoodType) + .16 V(Cost) + .10 V(Service) + .06 V(Neighborhood) + .03 V(Decor)
- OR considers cost to be most important, likes many food types:
  - U (restaurant) = .41 V(Cost) + .24 V (FoodQuality) + .16 V(Decor) + .10 V(Neighborhood) + .06 V(Service) + .03 V(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

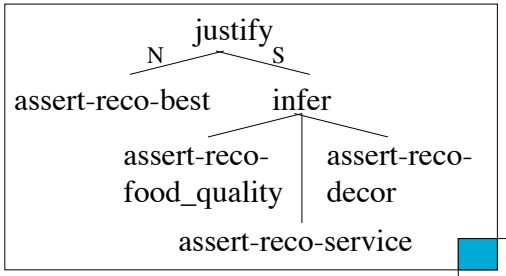
- Train NLG modules automatically
  - Supervised learning using user ratings of text quality
- Benefits:
  - ✓ Speed of NLG module engineering
  - ✓ Requires less linguistic and domain expertise
  - ✓ Clear method for adaptation
- Open questions:
  - Does trainable generation work well for flexible generation tasks?
  - How does the output quality compare to that of template generation?

Content Plan for a Recommendation

Strategy	Recommend
Items	Bar Pitti, Arlecchino, Babbo, Cent'anni, Cucina Stagionale, Grand Ticino, Il Mulino, John's Pizzeria, Marinella, Minetta Tavern, Trattoria Spaghetto, Vittorio Cucina
Relations	justify(nuc1; sat:2) justify(nuc:1; sat:3) justify(nuc:1, sat:4)
Content	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)))

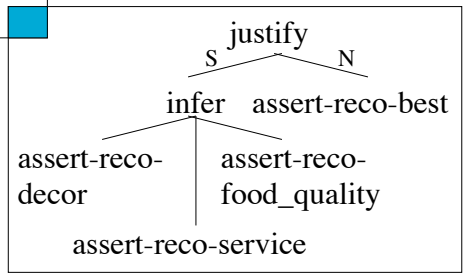
Problem: How to Choose A Good Content Organization?

One content plan, multiple text plans

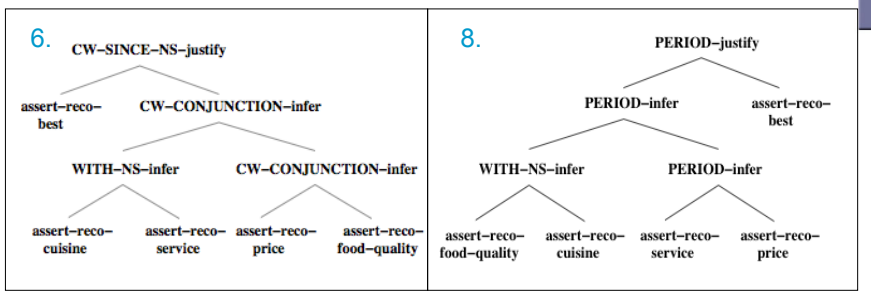


1. *Babbo is the best*
2. *Babbo has superb food quality*
3. *Babbo has excellent service*
4. *Babbo has excellent decor*

1. *Babbo has superb food quality*
2. *Babbo has excellent service*
3. *Babbo has excellent décor*
4. *Babbo is the best*



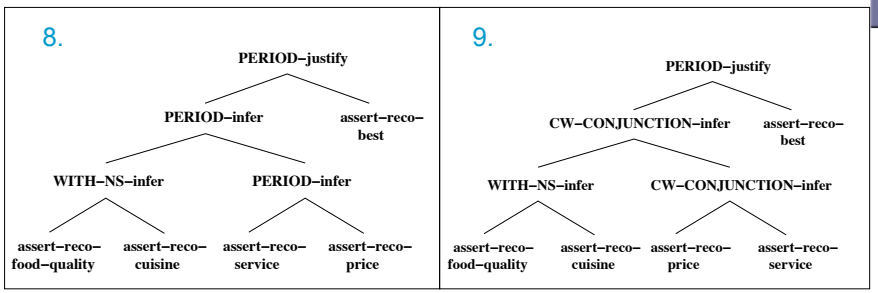
Leading to many sentence plans



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.

One text plan, many sentence plans



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.

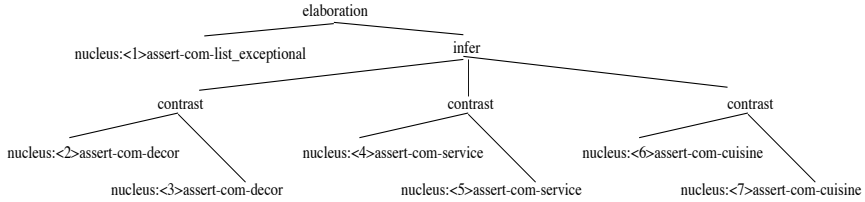
9. Chanpen Thai is a Thai restaurant with good food quality, its price is 24 dollars, and it has good service. It has the best overall quality among the selected restaurants.

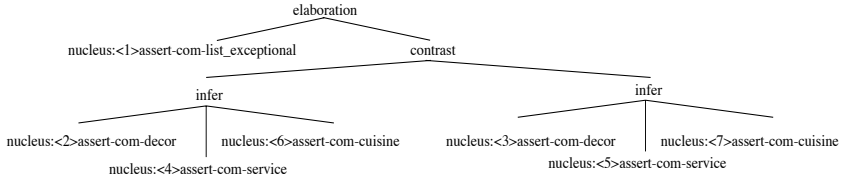
Alt	Realization	A	B	AVG
6	Chanpen Thai has the best overall quality among the selected restaurants since it is a Thai restaurant, with good service, its price is 24 dollars, and it has good food quality.	1	4	2.5
7	Chanpen Thai has the best overall quality among the selected restaurants because it has good service, it has good food quality, it is a Thai restaurant, and its price is 24 dollars.	2	5	3.5
4	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality, with good service, it is a Thai restaurant, and its price is 24 dollars.	2	4	3
9	Chanpen Thai is a Thai restaurant, with good food quality, its price is 24 dollars, and it has good service. It has the best overall quality among the selected restaurants.	2	4	3
5	Chanpen Thai has the best overall quality among the selected restaurants. It has good service. It has good food quality. Its price is 24 dollars, and it is a Thai restaurant.	3	2	2.5
3	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars. It is a Thai restaurant, with good service. It has good food quality.	3	3	3
10	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality. Its price is 24 dollars. It is a Thai restaurant, with good service.	3	3	3
2	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars, and it is a Thai restaurant. It has good food quality and good service.	4	4	4
1	Chanpen Thai has the best overall quality among the selected restaurants. This Thai restaurant has good food quality. Its price is 24 dollars, and it has good service.	4	3	3.5
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.	4	2	3

Content Plan for a Comparison (COMPARE-3)

strategy:	compare3
items:	Above, Carmine's
relations:	elaboration(nuc:1,sat:2); elaboration(nuc:1,sat:3); elaboration(nuc:1,sat:4); elaboration(nuc:1,sat:5); elaboration(nuc:1,sat:6); elaboration(nuc:1,sat:7); contrast(nuc:2,nuc:3); contrast(nuc:4,nuc:5); contrast(nuc:6,nuc:7)
content:	1. assert(exceptional(Above,Carmine's)) 2. assert(has-att(Above, decor(good))) 3. assert(has-att(Carmine's, decor(decent))) 4. assert(has-att(Above, service(good))) 5. assert(has-att(Carmine's, service(good))) 6. assert(has-att(Above, cuisine(New American))) 7. assert(has-att(Carmine's, cuisine(Italian)))

Two text plans for COMPARE-3





Alternative Realizations for COMPARE-3

Alt	Realization	A	B	AVG
11	Above and Carmine's offer exceptional value among the selected restaurants. Above, which is a New American restaurant, with good decor, has good service. Carmine's, which is an Italian restaurant, with good service, has decent decor.	2	2	2
12	Above and Carmine's offer exceptional value among the selected restaurants. Above has good decor, and Carmine's has decent decor. Above and Carmine's have good service. Above is a New American restaurant. On the other hand, Carmine's is an Italian restaurant.	3	2	2.5
13	Above and Carmine's offer exceptional value among the selected restaurants. Above is a New American restaurant. It has good decor. It has good service. Carmine's, which is an Italian restaurant, has decent decor and good service.	3	3	3
14	Above and Carmine's offer exceptional value among the selected restaurants. Above has good decor while Carmine's has decent decor, and Above and Carmine's have good service. Above is a New American restaurant while Carmine's is an Italian restaurant.	4	5	4.5
20	Above and Carmine's offer exceptional value among the selected restaurants. Carmine's has decent decor but Above has good decor, and Carmine's and Above have good service. Carmine's is an Italian restaurant. Above, however, is a New American restaurant.	2	3	2.5
25	Above and Carmine's offer exceptional value among the selected restaurants. Above has good decor. Carmine's is an Italian restaurant. Above has good service. Carmine's has decent decor. Above is a New American restaurant. Carmine's has good service.	NR	NR	NR

(25 not produced because it violates centering constraints)

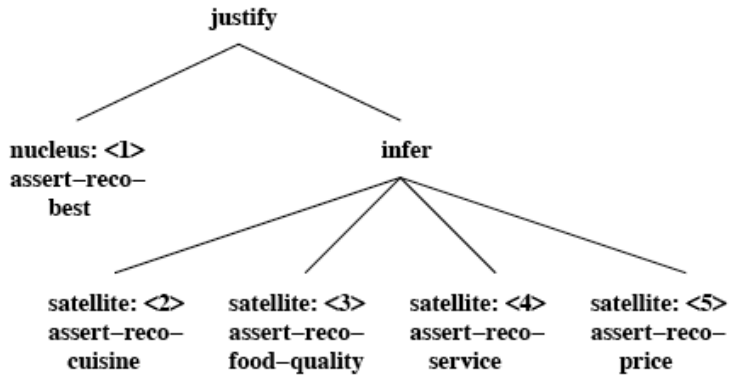
Solution: Trainable Sentence Planning

- 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 content plans
- **Output:** A set of sentence plan trees, each with an accompanying dependency tree
- **Steps:**
  1. Group content items 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

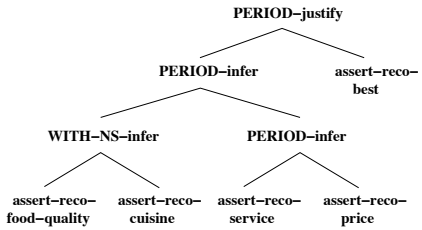
Input: Set of Text Plan trees



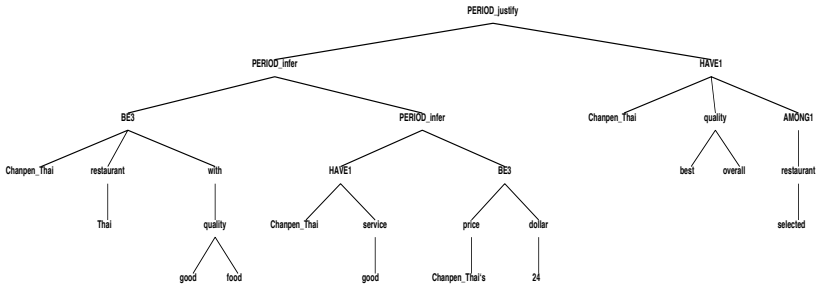
Clause Combining Operations: Examples

- Merge: (contrast, infer)
  - Babbo has superb décor AND
    Babbo has mediocre food quality ==>
    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 ==>
    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 ==>
    Above has decent décor *but* Carmine's has good décor.
- With-reduction: (infer, justify)
  - Above is an Italian restaurant AND
    Above has good décor ==>
    Above is an Italian restaurant with good *décor*.

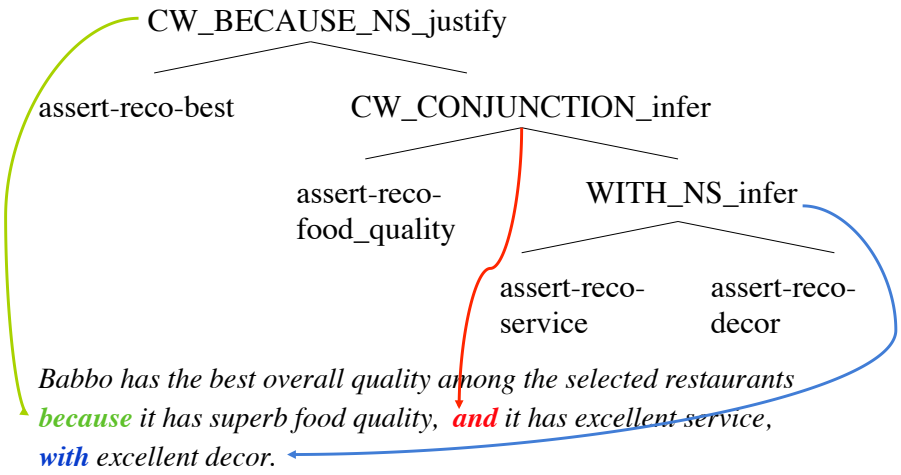
Output: Set of SP-tree & D-tree pairs, e.g.,



SP-tree & D-tree for Alternative 8



Sentence Plan Tree for One Recommend Alternative



Sentence Plan Ranking

- **Input:** Set of sentence plan trees (+ d-trees)
- **Uses:** Set of rules learned from labeled set of sentence plan training examples
- **Output:** Ranked list of sentence plan trees

Training the Sentence Plan Ranker - 1

- 30 text plans for each type of information presentation (recommend, compare)
- Sentence plan generator generated 20 (sp-tree, d-tree) pairs for each (total 1800)
- Template generator produced 1 realization for each text plan (30)
- Two judges rated realized text of each variant on a scale from 1(worst) - 5 (best)
  - Organization, ease of understanding
- Features automatically generated from realizations and sp-tree/ dependency tree pairs
  - 7024 features

Training the Sentence Plan Ranker - 2

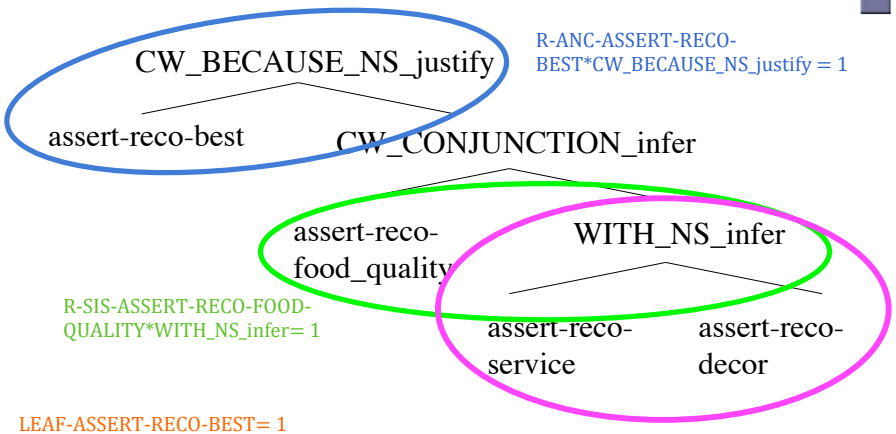
- Use RankBoost to learn a function from features to ratings
  - Given features & ranking as input
    - (sp-tree, d-tree, realization) triples are examples
    - Ratings are feedback
  - Produces a ranking over alternatives, not just the best alternative
  - Can handle many sparse features
  - Learns a rule-based model indicating the effects of features on ranking (allows qualitative analysis of models)

Freund, Y., et al. (1998). An efficient boosting algorithm for combining preferences. In *Machine Learning: Proc. of the 15th Int'l Conference*.

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



R-ANC-ASSERT-RECO-BEST*CW_BECAUSE_NS_justify = 1

R-SIS-ASSERT-RECO-FOOD-QUALITY*WITH_NS_infer= 1

LEAF-ASSERT-RECO-BEST= 1

LEAF-ASSERT-RECO-BEST*ASSERT-RECO-FOOD-QUALITY= 1

ETC.

TRAV-WITH_NS_infer*ASSERT-RECO-SERVICE*ASSERT-RECO-DECOR= 1

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

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 best sentence plan as selected by human judges
    - RANDOM – score of a sentence plan randomly selected from alternatives

Experiment 2: Results

User	Strategy	SPaRKy	Human	Random
AVG	RECOMMEND	3.6 (0.77)	3.9 (0.55)	2.8 (0.81)
AVG	COMPARE-2	4.0 (0.66)	4.4 (0.54)	2.8 (1.30)
AVG	COMPARE-3	3.6 (0.68)	4.0 (0.49)	2.7 (1.20)

- 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: SPaRKy vs. Templates

- Method: For each content plan, compare
  - SPaRKy -- human rater 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

User	Strategy	SPaRKy	Human	Template
AVG	RECOMMEND	3.6 (0.59)	4.4 (0.37)	4.2 (0.74)
AVG	COMPARE-2	3.9 (0.52)	4.6 (0.39)	3.6 (0.75)
AVG	COMPARE-3	3.4 (0.38)	4.6 (0.35)	4.1 (1.23)

- HUMAN significantly better than TEMPLATE only for COMPARE-2
  - Human raters did not like template for COMPARE-2
- Standard Deviation for TEMPLATE very large
  - So, good overall, but does poorly on some inputs
- TEMPLATE significantly better than SPaRKy for RECOMMEND and COMPARE-3
- SPaRKy better than TEMPLATE for COMPARE-2 (trend)

But remember …

Alt	Realization	A	B	AVG
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.	1	4	2.5
7	Chanpen Thai has the best overall quality among the selected restaurants because it has good service, it has good food quality, it is a Thai restaurant, and its price is 24 dollars.	2	5	3.5
4	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality, with good service, it is a Thai restaurant, and its price is 24 dollars.	2	4	3
9	Chanpen Thai is a Thai restaurant, with good food quality, its price is 24 dollars, and it has good service. It has the best overall quality among the selected restaurants.	2	4	3
5	Chanpen Thai has the best overall quality among the selected restaurants. It has good service. It has good food quality. Its price is 24 dollars, and it is a Thai restaurant.	3	2	2.5
3	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars. It is a Thai restaurant, with good service. It has good food quality.	3	3	3
10	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality. Its price is 24 dollars. It is a Thai restaurant, with good service.	3	3	3
2	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars, and it is a Thai restaurant. It has good food quality and good service.	4	4	4
1	Chanpen Thai has the best overall quality among the selected restaurants. This Thai restaurant has good food quality. Its price is 24 dollars, and it has good service.	4	3	3.5
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.	4	2	3

Training Rankers for Individual Users

Alt	Realization	A	B	SPR_A	SPR_B	SPR_{AVG}
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.	1	4	0.16	0.65	0.58
7	Chanpen Thai has the best overall quality among the selected restaurants because it has good service, it has good food quality, it is a Thai restaurant, and its price is 24 dollars.	2	5	0.38	0.54	0.42
4	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality, with good service, it is a Thai restaurant, and its price is 24 dollars.	2	4	0.53	0.62	0.53
9	Chanpen Thai is a Thai restaurant, with good food quality, its price is 24 dollars, and it has good service. It has the best overall quality among the selected restaurants.	2	4	0.47	0.53	0.63
5	Chanpen Thai has the best overall quality among the selected restaurants. It has good service. It has good food quality. Its price is 24 dollars, and it is a Thai restaurant.	3	2	0.59	0.32	0.46
3	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars. It is a Thai restaurant, with good service. It has good food quality.	3	3	0.64	0.40	0.62
10	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality. Its price is 24 dollars. It is a Thai restaurant, with good service.	3	3	0.67	0.46	0.58
2	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars, and it is a Thai restaurant. It has good food quality and good service.	4	4	0.75	0.50	0.74
1	Chanpen Thai has the best overall quality among the selected restaurants. This Thai restaurant has good food quality. Its price is 24 dollars, and it has good service.	4	3	0.64	0.52	0.45
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.	4	2	0.81	0.29	0.73

Why "averages" can hurt

- Compare training and testing on individual judgments with training and testing on averaged judgments …
- Random baseline has average error rate of 0.5

RECOMMEND	A's model	B's model	AVG model
A's test data	0.17	0.52	0.29
B's test data	0.52	0.17	0.27
AVG's test data	0.31	0.31	0.20

- Best results:
  - Minimise ranking error by avoiding compromises such as mixtures of learned rules

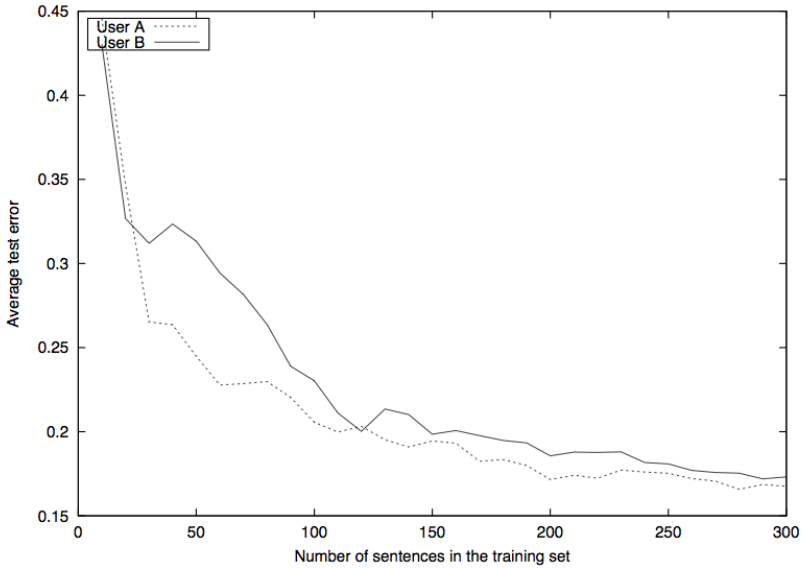
Type of Rules Learned – Insight into user preferences

- *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.*

Individual differences

- Users have different perceptions of the quality of alternative realizations of a content plan
- Individualized models perform better than those trained for groups of users.
- Qualitative analysis indicates that trainable sentence generation is sensitive to variations in
  - presentation type
  - individual human preferences about interaction between domain specific content and syntactic structure
- Note that generation effectively builds a new, artificial corpus, from which elements are sampled to be rated by users.

Individual preferences can be learned (relatively) quickly



Templates can be beaten (for some of the people, some of the time)

User	Strategy	SPaRKy	Human	Template
A	RECOMMEND	3.5 (0.87)	3.9 (0.61)	3.9 (1.05)
A	COMPARE-2	3.8 (0.98)	4.3 (0.73)	4.2 (0.64)
A	COMPARE-3	3.1 (1.02)	3.6 (0.80)	3.9 (1.19)
B	RECOMMEND	4.4 (0.70)	4.7 (0.46)	4.5 (0.76)
B	COMPARE-2	4.4 (0.69)	4.7 (0.53)	3.1 (1.21)
B	COMPARE-3	4.4 (0.62)	4.8 (0.40)	4.2 (1.34)

SPaRKy can produce output => TEMPLATE in many cases

BUT still significant gap between SPaRKy and Humans

Moving to a new domain

- Add new rhetorical relations
- Add domain assertions (messages)
- Map domain assertions to **D-trees for** input to RealPro surface realizer
- Modify probability distribution of clause combining operators (may be learned from corpora)
- Generate alternative realizations and collect user ratings

Summary

- **SPaRKy**, a trainable sentence planner for complex information presentations in spoken dialogue
- Trainable sentence planning can produce output of quality equal to or better than template-based generator
  - with less programming effort and more flexibility
- Gap between HUMAN scores and TEMPLATE scores indicates
  - SPG produces sentence plans as good as those of template generator
  - Accuracy of SPR can be improved