Generating and Evaluating User-Tailored Responses

Matthew Marge

Marilyn Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, G. Vasireddy. Generation and Evaluation of User Tailored Responses in Multimodal Dialogue, Cognitive Science, 28: 811-840, 2004.

Why User Tailoring?

- People tailor their utterances to their conversational partners
  - Based on user model of partner

- User modeling improves:
  - Listeners’ comprehension
  - Satisfaction with interactive system
  - Efficiency at executing conversational tasks
  - Likelihood of changing listener’s beliefs/attitudes

MATCH System Focus

- **M**ultimodal **A**ccess **T**o **C**ity **H**elp System
  - Multimodal dialogue system giving info on NYC restaurants
- Tailoring content of dialogue system utterances for
  - Persuasion
  - Argumentation
  - Advice-giving

Overall Hypothesis

- Algorithms that adapt dialogue content for high-level discourse by referring to a user model will improve system **usability**, **efficiency**, and **effectiveness**

Shaping System Responses

- Dialogue systems should be concise when presenting information
- Present user with options in an easy-to-understand form
- **Recommendations** and **comparisons** among a set of options (i.e., restaurants) should also be concise

Guidelines for Evaluative Effective Arguments(Carenini & Moore, 2006)

1) Identify supporting/opposing evidence
2) Position the main claim (either first/last)
3) Select supporting/opposing evidence
4) Arrange supporting evidence
5) Address and order opposing evidence
6) Place order between supporting/opposing evidence

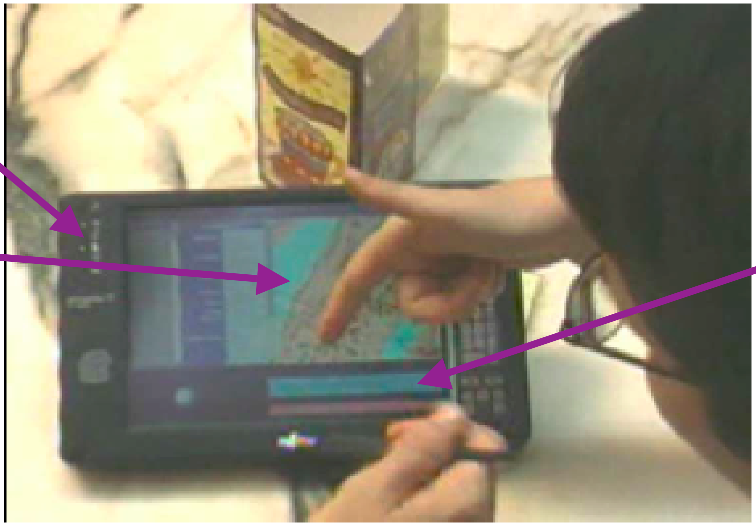
MATCH System Req's

- Based on GEA guidelines

1) Represent user's preferences
2) Measure strength of supporting & opposing evidence
3) Represent user awareness of facts
4) Order selected content into coherent, persuasive arguments

MATCH Dialogue System

- Screen contains table showing items matching user’s request



Buttons activate ASR

NYC street map

Feedback panel describing system state

MATCH Input/Output

- System input may be any combination of:
  - Speech
  - Pen gestures
  - Handwriting
- System output may be:
  - Speech
  - Changes in map display or feedback panel

MATCH System Capabilities

- Provides recommendations and comparisons of restaurants
- **Tailors** recommendations and comparisons to a model of user preferences
- Generates concise, easily understood responses

Applying NLG User Modeling Techniques to MATCH

- User models based on multi-attribute decision theory
- Dialogue planning algorithms for selecting, ranking and presenting of options
- Model specifies most important attributes, what is **worth telling** the user
- Makes predictions of which options user will prefer

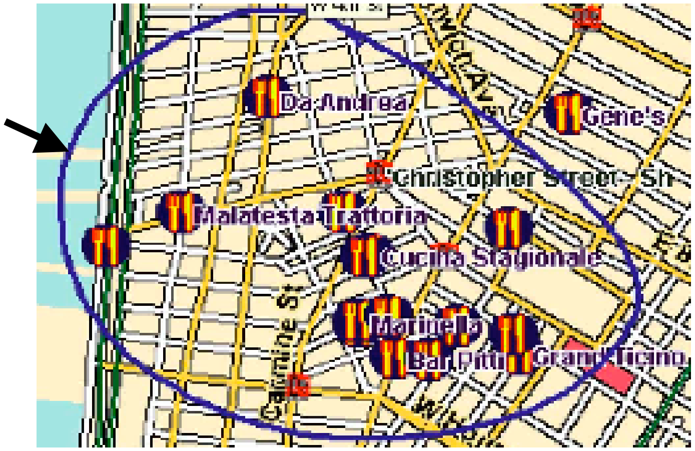
Example MATCH Interaction

User: *“Show me Italian restaurants in the West Village.”*

System: [zoom to West Village, display Italian restaurants]

U: [circles selection] “*Compare.”*

S: *“Among the selected restaurants, the following offer exceptional value. Uguale’s price is $33. It has excellent food quality and good décor. Da Andrea’s price is…[3 restaurants provided]”*

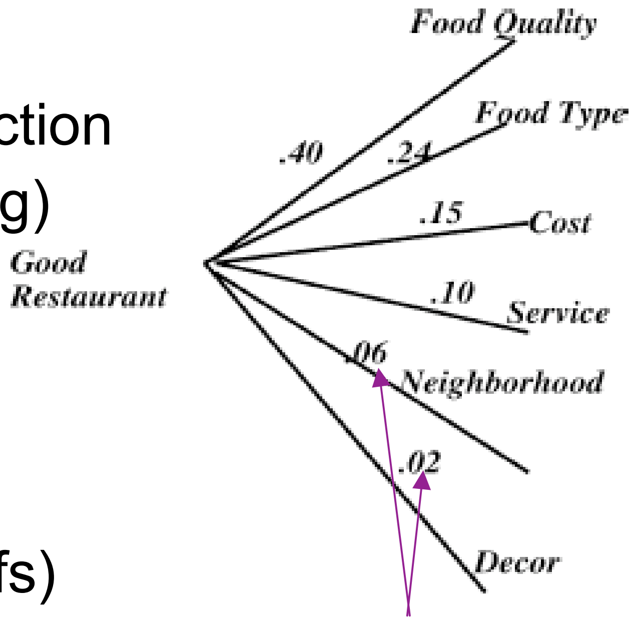


U: [writes word “recommend”]

S: *“Uguale has the best overall value among the selected restaurants. Uguale’s price is $33.”*

User-Model Attributes

- Goal: Select a good restaurant
- Factors involved in restaurant selection
  - **Food Quality** (quantitative rating)
  - **Cost** (quantitative scale)
  - **Service** (quantitative rating)
  - **Décor** (quantitative rating)
  - **Type of food**: French, Italian, Vietnamese (based on user prefs)
  - **Neighborhood**: Greenwich Village, etc. (based on user prefs)



sample weights

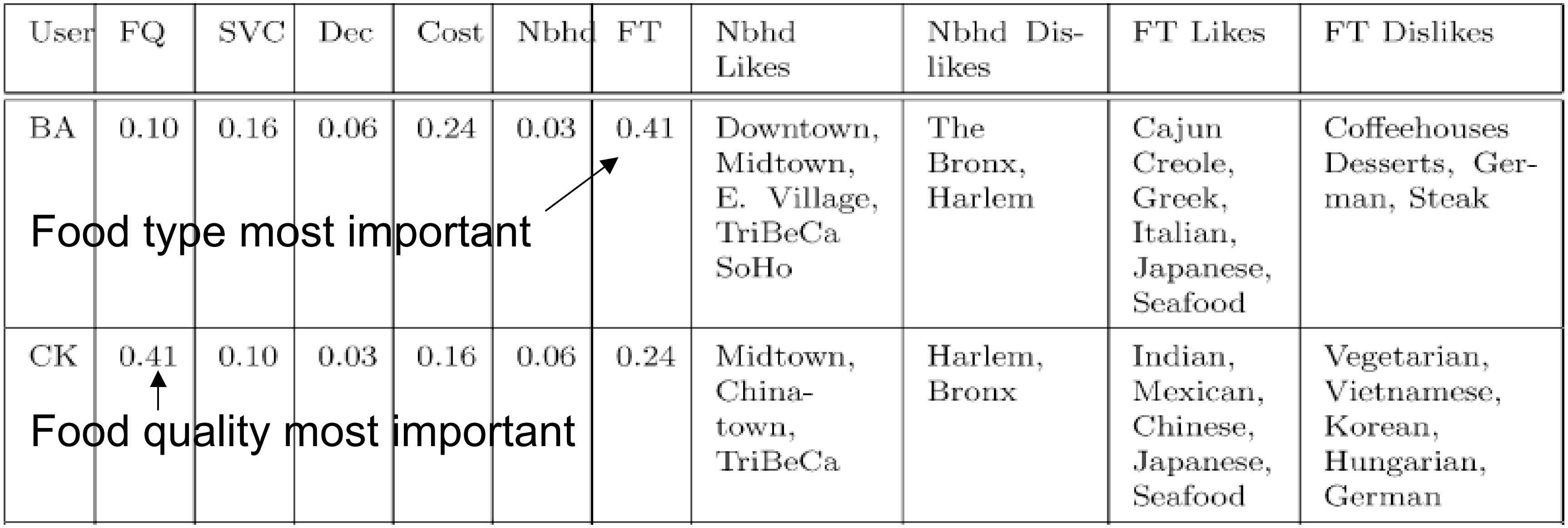
Applying Decision Theory to User-Model Attributes

- Utility(Option) = F(attributes of an option)
- Higher utility scores indicate greater desirability
- **Additive** Multi-attribute Value Function
  - U(option) = $w_1a_1 + w_2a_2 + \ldots w_na_n$
  - Weights (w_i) are user specific
  - a_i = Utility score for individual attribute *i* given currently selected restaurant

Assigning Weights to Attributes

- Use “SMARTER” procedure
  - Only needs user to specify **ranking** of attributes
  - Example SMARTER question obtaining user preferences (weights):
    - *“Imagine that for whatever reason you have had the horrible luck to have to eat at the worst possible restaurant in the city. The price is $100 per head, you do not like the type of food they have, you don’t like the neighborhood, the food itself is terrible, the decor is ghastly, and it has terrible service. Now imagine that a good fairy comes along…**What dimension would you choose? Food quality, service, décor, cost, neighborhood, or food type?***
  - Prompt repeated, removing selected attribute

Example User Models



User	FQ	SVC	Dec	Cost	Nbhd	FT	Nbhd Likes	Nbhd Dislikes	FT Likes	FT Dislikes
BA	0.10	0.16	0.06	0.24	0.03	0.41	Downtown, Midtown, E. Village, TriBeCa SoHo	The Bronx, Harlem	Cajun Creole, Greek, Italian, Japanese, Seafood	Coffeehouses Desserts, German, Steak
CK	0.41	0.10	0.03	0.16	0.06	0.24	Midtown, Chinatown, TriBeCa	Harlem, Bronx	Indian, Mexican, Chinese, Japanese, Seafood	Vegetarian, Vietnamese, Korean, Hungarian, German

Food type most important

Food quality most important

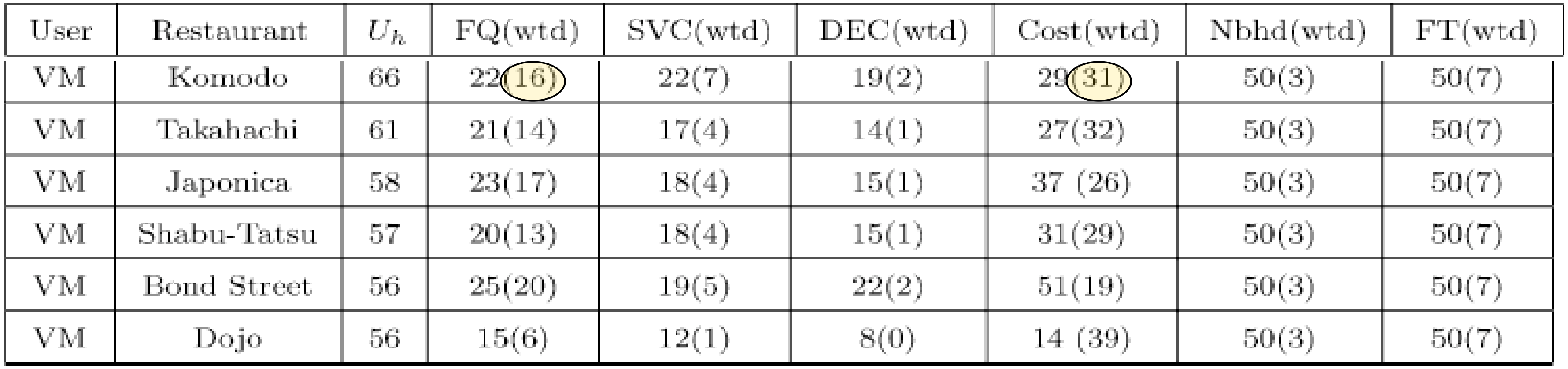
FQ: food quality, SVC: service, DEC: décor, Nbhd: neighborhood, FT: food type

SPUR Dialogue Planner

- **S**peech **P**lanning with **U**tilities for **R**estaurants (SPUR) uses user model for:
  - Ranking options returned from DB query
    - SPUR selects subset of restaurants to recommend or compare
      - e.g. *{Uguale, Da Andrea}* from "Italian restaurants in West Village"
  - Selects which attributes are to be mentioned for that option
    - Based on **conciseness** parameter
      - Taken as input, determines conciseness of system responses

DB Query Results used by SPUR (Influenced by User Model)

- User VM believes *cost* is most important, *food quality* is second-most important
  - Looking for “Japanese Restaurant in East Village”
  - Komodo highest-ranked because of its low cost and high quality (FQ)



User	Restaurant	U_h	FQ(wtd)	SVC(wtd)	DEC(wtd)	Cost(wtd)	Nbhd(wtd)	FT(wtd)
VM	Komodo	66	22(16)	22(7)	19(2)	29(31)	50(3)	50(7)
VM	Takahachi	61	21(14)	17(4)	14(1)	27(32)	50(3)	50(7)
VM	Japonica	58	23(17)	18(4)	15(1)	37 (26)	50(3)	50(7)
VM	Shabu-Tatsu	57	20(13)	18(4)	15(1)	31(29)	50(3)	50(7)
VM	Bond Street	56	25(20)	19(5)	22(2)	51(19)	50(3)	50(7)
VM	Dojo	56	15(6)	12(1)	8(0)	14 (39)	50(3)	50(7)

SPUR Dialogue Strategies

- Two types
  1) **Recommend** one of a set of restaurants
  2) **Compare** 3 or more selected restaurants
- Determining **conciseness** of responses
  - Controlled with an “outlier” parameter *z*
  - SPUR uses this to choose **which** restaurants and their attributes are worth mentioning (i.e., outliers)

Defining Outliers

- Use *z-score*: how many standard deviations a value *v* is away from mean of a population of values *V*

$$z(v) = \frac{v - \mu_V}{\sigma_V}$$

- 2 value populations:
  - Other attributes for same restaurant
    - Used for **recommendations**
  - The same attribute for other restaurants
    - Used for **comparisons**
- Now a **variety** of attributes/restaurants can be deemed worth mentioning (based on *z*)
  - e.g., *z* = *1.0* --> weighted attribute values must be **more than 1 s.d. away** from mean to be mentioned

SPUR Recommendation Strategy

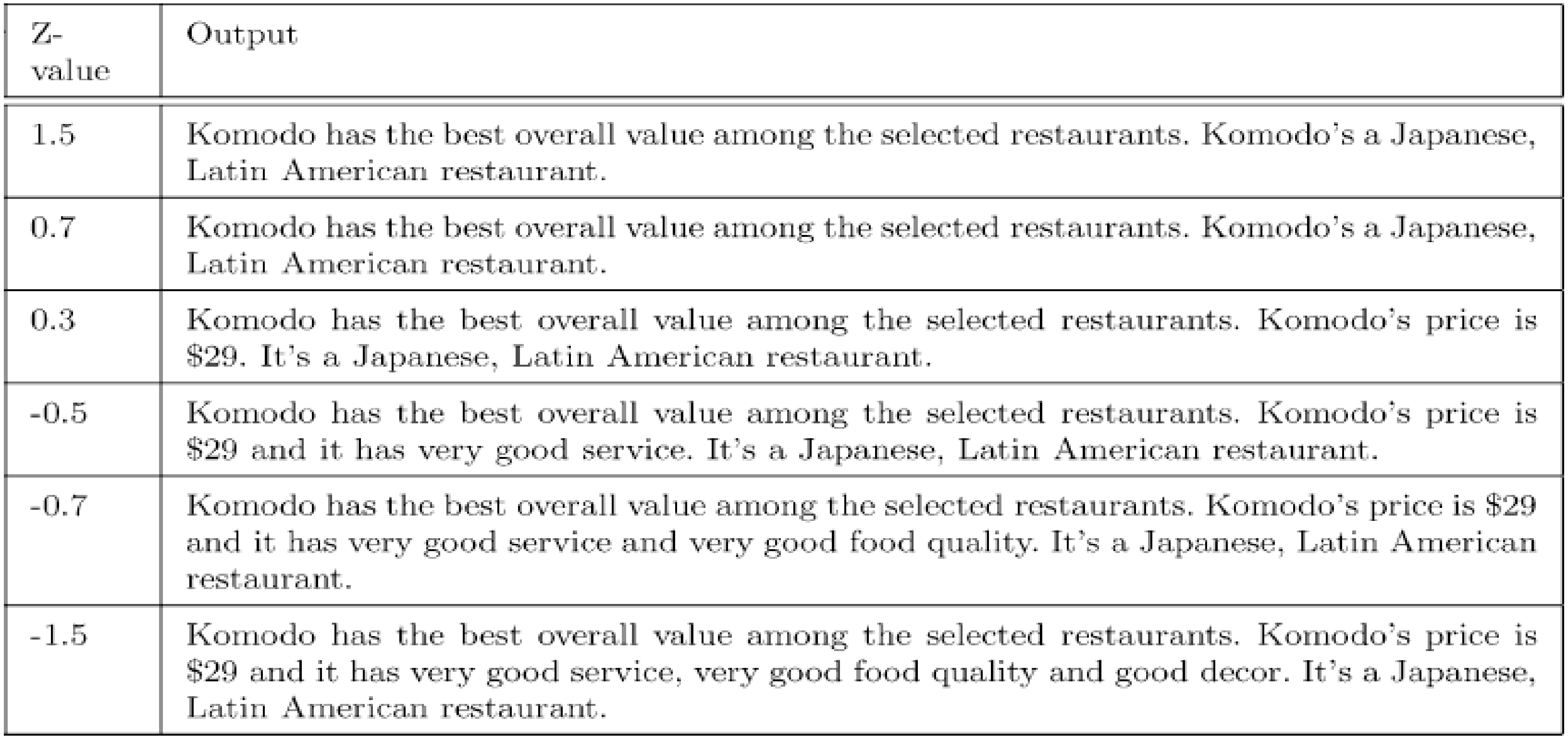
1. Select the best restaurant
   - Based on highest overall utility
2. Provide convincing reasons for user to choose it
   - Use *z-score* to identify attributes whose weighted attribute values are **outliers**
3. Build content plan
   - To be realized into natural language text

Varying Conciseness for Recommendations

Concise

↓

Verbose



Z-value	Output
1.5	Komodo has the best overall value among the selected restaurants. Komodo's a Japanese, Latin American restaurant.
0.7	Komodo has the best overall value among the selected restaurants. Komodo's a Japanese, Latin American restaurant.
0.3	Komodo has the best overall value among the selected restaurants. Komodo's price is $29. It's a Japanese, Latin American restaurant.
-0.5	Komodo has the best overall value among the selected restaurants. Komodo's price is $29 and it has very good service. It's a Japanese, Latin American restaurant.
-0.7	Komodo has the best overall value among the selected restaurants. Komodo's price is $29 and it has very good service and very good food quality. It's a Japanese, Latin American restaurant.
-1.5	Komodo has the best overall value among the selected restaurants. Komodo's price is $29 and it has very good service, very good food quality and good decor. It's a Japanese, Latin American restaurant.

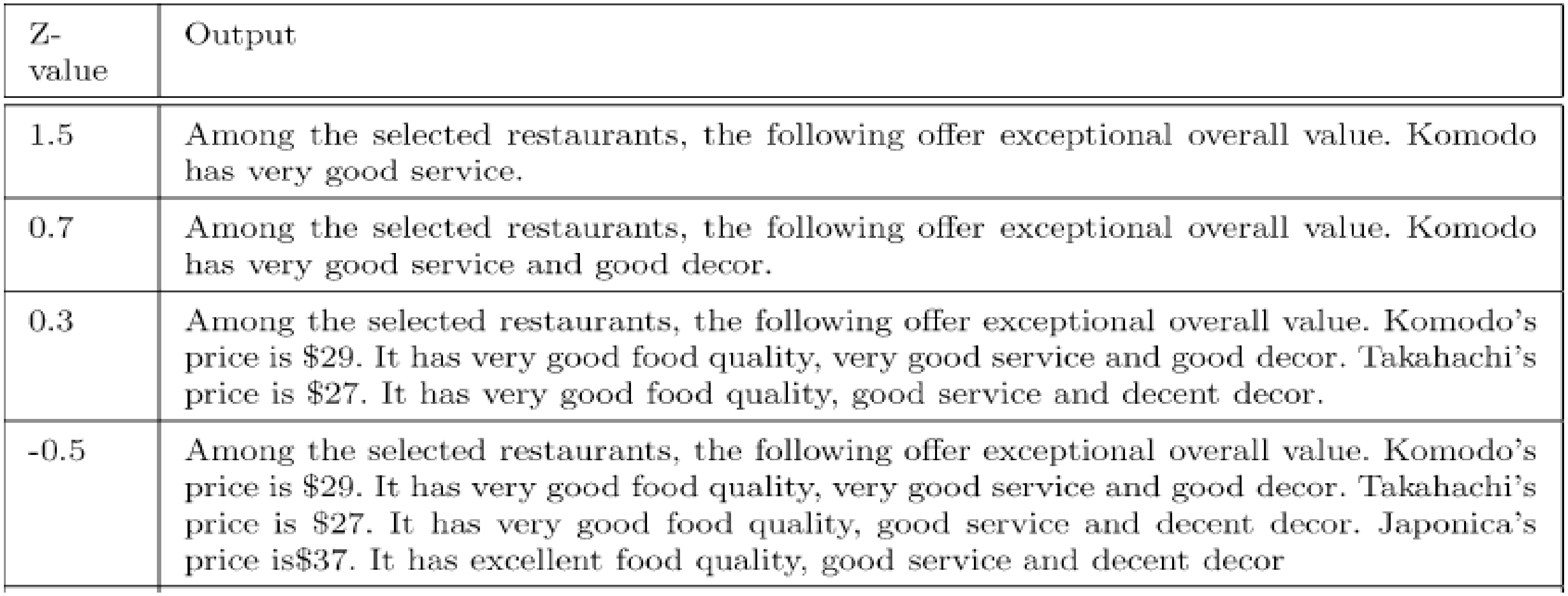
SPUR Comparison Strategy

1. Select several potential restaurants
   - Based on highest overall utilities
   - Use *z-score* to select restaurants that are **outliers** (positive outliers only)
2. Provide same facts about each restaurant
   - If a weighted attribute value is an outlier for **any** of the selected restaurants, mention that attribute for **all** restaurants
3. Build content plan
   - To be realized into natural language text

Varying Conciseness for Comparisons

Concise

Verbose



Z-value	Output
1.5	Among the selected restaurants, the following offer exceptional overall value. Komodo has very good service.
0.7	Among the selected restaurants, the following offer exceptional overall value. Komodo has very good service and good decor.
0.3	Among the selected restaurants, the following offer exceptional overall value. Komodo's price is $29. It has very good food quality, very good service and good decor. Takahachi's price is $27. It has very good food quality, good service and decent decor.
-0.5	Among the selected restaurants, the following offer exceptional overall value. Komodo's price is $29. It has very good food quality, very good service and good decor. Takahachi's price is $27. It has very good food quality, good service and decent decor. Japonica's price is$37. It has excellent food quality, good service and decent decor

Lower *z-score* indicates lower threshold for choosing restaurants (and attributes)

Note: Attributes are same across all restaurants per row

Realization of Content Plans

- Realizer takes recommendation or comparison content plans as input
- Uses templates to generate text to be passed to text-to-speech system
  - *Cost* remained in $
  - Other attributes: excellent, very good, etc.
- Main point followed by supporting/opposing evidence (follows GEA guidelines)

Experimental Procedure

- Each subject **overhears** a series of dialogues between a user and the MATCH system
  - 1 dialog for each restaurant-selection task (e.g., *"Find Italian restaurants in West Village"*)
  - Dialogues consist of multiple exchanges
- Each dialogue exchange presented on separate webpage
  - User always asked for **comparison** first, **recommendation** second in all tasks
- **Subject** user models collected separately

Experiment 1: Tailoring and Mode

- Hypotheses:
  - Users will prefer tailored over untailored responses
  - Users will prefer text over speech responses
  - Tailoring will have greater effect on judgments of speech over text responses
    - Speech has greater cognitive load

Experiment 1: Design

- Subject overhears 4 dialogues about different restaurant selection tasks
- Entire sequence presented twice
  - First using text, second using speech
- 16 fluent English-speaking subjects

Experiment 1: Evaluation

- Subject makes 6 judgments per dialogue (rating 1 to 5) about *Information Quality*
  - i.e., dialogue easy to understand, attributes are appropriate
  - 1 recommendation and 2 comparisons each for:
    - System using **subject-tailored** user model
    - System using **randomly selected** user model
- Subject makes 4 judgments per dialogue (rating 1 to 5) about *Ranking Confidence*
  - i.e., selected restaurant(s) are places I would visit

Experiment 1: Results

- Subjects **preferred** tailored over pseudo-random responses for *Information Quality* and *Ranking Confidence* (*P* < .05)
  - Filtered out random user models too close to user's own for experiment
- Subjects **preferred** text over speech responses for *Information Quality* (*P* < .05)
- No significant interaction between model type and mode

Experiment 2: Conciseness

- Hypotheses:
  - Users will be sensitive to amount of information given in system responses
    - *“Concise” responses* - judged as providing too little information
    - *“Sufficient” responses* - judged as providing right amount of information
    - *“Verbose” responses* - judged as providing too much information
  - Users will prefer concise over verbose responses

Experiment 2: Design

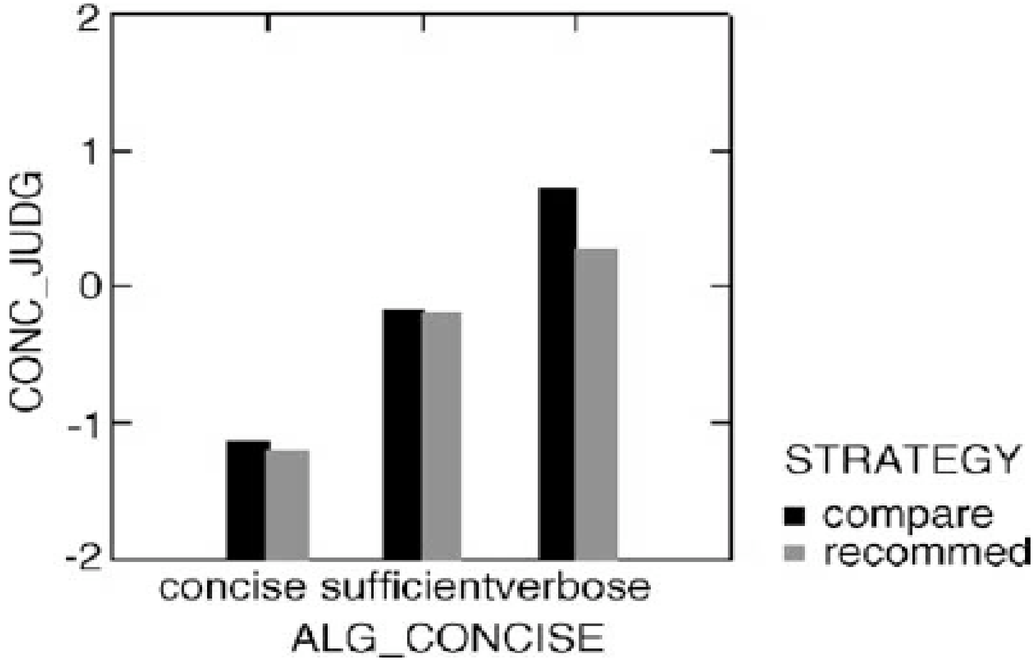
- Subject overhears 6 dialogues about different restaurant selection tasks
- 21 fluent English-speaking subjects

Experiment 2: Evaluation

- Subject makes 4 judgments per dialogue (rating 1 to 5) about *Conciseness*
  - i.e., amount of information far too little,…,far too much
- Subjects saw one webpage each for recommend and compare
  - Presented with 3 system responses (*concise, sufficient, verbose*)

Experiment 2: Results

- *"concise"* outputs judged as having too little information compared to *"sufficient"*



Conclusion

- Developed algorithms for information presentation with multi-attribute decision theory
  - Enabled option and attribute selection for recommendations and comparisons using SPUR
- User models based on multi-attribute decision theory **generalize across domains**
  - MATCH (restaurants), GEA (real estate), FLIGHTS (airline booking)
- Subjects preferred
  - Tailored over untailored & text over speech
- Presentation conciseness can be controlled

For Discussion

- Can a system like MATCH support **real-time** interactions?
- Will the same results hold in domains where user models can be built **implicitly**?
  - e.g., interacting with a music player
- What other domains would user-tailoring dialogue responses be appropriate?

References

- Carenini, Giuseppe and Johanna D. Moore, Generating and Evaluating Evaluative Arguments, Artificial Intelligence 170(11):925-952, 2006.
- Moore, Johanna D., Mary Ellen Foster, Oliver Lemon, and Michael White, Generating Tailored, Comparative Descriptions in Spoken Dialogue, in Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, AAAI Press, 2004.
- Walker, M. Lecture Slides for *An Introduction to AI*. University of Sheffield, 2006.
- Walker, M., S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, G. Vasireddy. Generation and Evaluation of User Tailored Responses in Multimodal Dialogue, Cognitive Science, 28: 811-840, 2004.