Natural Language Generation for Embodied Conversational Agents

Day 5

Kristina Striegnitz

ESSLLI 2008
Hamburg, Germany

Today

- Social ECAs

- Evaluation

- DIY pointers

The media equation   Cliff Nass and colleagues

- People treat computers as social actors. I.e. they apply social rules of interaction.

- E.g., politeness.
  A computer program is judged more positively when the judging is done on the same computer where the program was tried than on another computer.

- Stereotypes.
  A virtual tutor with a female voice is rated more knowledgeable about love and relationships than about engineering; vice versa for male voices.
- Flattery.
  People like computers better that flatter them.

Things people are looking at

- rapport (e.g., Bickmore, ICT)

- politeness (e.g., Andre & Rehm)

- emotion (e.g., Pelachaud, Poggi, work in the HUMAINE project)

- personality (e.g., Isbister)

REA: building trust through small-talk — Bickmore & Cassell

- A virtual real-estate agent.

- Uses social-language (small talk) with appropriate non-verbal behaviors to to build trust.

- Small talk is used to avoid face-threats (Brown & Levinson), build common-ground, and increase coordination.

REA: building trust through small-talk — Bickmore & Cassell

Move
1. How about this weather?
2. I think winters in Boston are awful.
3. How do you like Boston?
4. I have lived in Boston all my life. Come to think of it, I have lived inside this room all of my life. It is so depressing.
5. Boston is certainly more expensive than it used to be.
6. So, Where would you like to live?
7. How many bedrooms do you need?
8. Do you need access to the subway?
9. Is one bath enough?
10. You know, I keep showing the researchers here the same houses, over and over again. Maybe one day I will get lucky.
11. Have you been in the Media Lab before?
12. Do you know that the Media Lab is going to expand into another building. Things are really going well for the researchers here.
13. It is pretty cool do you think?
14. They are doing some crazy things in here.
15. I have shown houses to lots of students and faculty from M I T. But I always enjoy talking to them.
16. Anyway, What can you afford?
17. What kind of down payment can you make?
18. Let me see what I have available.

GRETA: expressing emotions — Catherine Pelachaud

- How to express emotions (in face and gesture)?

- Do people recognize them?

- How to overlay them?

Today

- Social ECAs

- Evaluation

- DIY pointers

Why Evaluation?

- To test whether we have achieved our goals.

- Goals: to improve human-computer interaction

    to better understand human-human communication

- To get insights into what went wrong and how to fix it.

Evaluation questions

- Is our ECA better than another system? (text based, speech only, some alternative ECA, a human)

    → What does it mean to be better?

- When people interact with our ECA, do they behave like they behave when they are interacting with other humans?

- Does our ECA behave like humans?

Do we want our system to behave like humans?

- Often assumption: if the system behaves like a human speaker, that's good for the human listener as well.

- For example: ASGRE challenge 2007 (attribute selection for referring expressions)

- Systems were evaluated wrt. how similar their REs were to human produced Res (humanlikeness measures).

- Gatt & Belz (2008) also evaluated the systems wrt. to task-performance measures: How reliable and how fast can readers/listeners pick out the right object based on these descriptions.

Do we want our system to behave like humans?

Gatt & Belz (2008) results:

- While there is high correlation within the two classes of metrics, there is no correlation between them. I.e., humanlikeness measures are not correlated to task-performance measures in the case of ASGRE.

Evaluation questions

- Is our ECA better than another system? (text based, speech only, some alternative ECA, a human)

    → What does it mean to be better?

- When people interact with our ECA, do they behave like they behave when they are interacting with other humans?

- Does our ECA behave like humans?

What does it mean to be better?

[Ruttkay et al. 2004, Christoph 2004]

- usability

    ease of use, efficiency, task completion

- user perception

    satisfaction, engagement, helpfulness, naturalness/believability, trust, perceived task difficulty, likeability, entertainment

- user behavior

    use of certain linguistic expressions, non-verbal behaviors, longterm behavior change

- biological measures

    heart rate, skin conductance, ...

Other dimensions of evaluation

- efficiency

- robustness

What to test: design parameters

- the looks

- communication modalities

- communication capabilities and strategies

- social issues: personality, social role, emotions

What to test: user parameters

- scenario (tutoring, information kiosk, sales, personal companion, ...)

- demographic factors

    gender, age, ethnicity, education, ...

- psychological factors

    personality, cognitive abilities, ...

Data collection framework

- survey

- experiment

- case study

What data to collect

- quantitative data (task performance, user behavior, biological data)

- questionnaire (subjective perception)

- detailed observation/analysis of interaction

- ratings by judges

Potential difficulties and problems

- hard to separate the different design parameters (looks, communicative

    modalities and capabilities, and the social behaviors)

Example Evaluations

Evaluating perception:

• Krahmer & Swerts 2004

Evaluating recall and perception:

• Buisine & Martin 2007

⎫ goal: refine knowledge about language

Evaluating perception and learning performance:

• Lester et al. 1997

Evaluating behavior change:

• Tartaro & Cassell 2008

• Bickmore

⎫ goal: prove that the systems work

Krahmer & Swerts: Effect of eye brow raises on the perception on focus

Q: What *is* the role of eyebrow movements for the perception of focus?

Material:

speech: *blauwe vierkant*

(contrasting 1) rode driehoek, 2) blauwe driehoek, 3) rode vierkant)

animation: cartoon face with eye brow movement on the stressed syllable of either the first or the second word

blauwe vierkant	*blauwe* VIERkant	*BLAUwe* vierkant
blauwe *vierkant*	blauwe *VIERkant*	BLAUwe *vierkant*

Krahmer & Swerts: Effect of eye brow raises on the perception on focus

Experiment 1: Subjective Preference

• subjects prefer the eye brows to be aligned with the pitch accent.

Experiment 2: Perceived Prominence

• pairwise comparison: same speech, eyebrow on first or second word
• In which version is the first/second word more prominent?
• Eye brows add to perceived prominence of words they are on; and downplay prominence of other word.

Experiment 3: Functional Analysis

• given a stimulus select previous utterance from among 1) rode driehoek, 2) blauwe driehoek, 3) rode vierkant
• stimulus may have matching pitch and eye brow or not
• both contribute to decision but pitch much more

Buisine & Martin 2007: The effect of gesture redundancy & complementarity on recall and perception

• Agent explains a gadget using redundant information in gesture in speech, complementary information in gesture or no content in gesture at all.

• Both deictic and iconic gestures were used.

• To test recall participants have to sketch the device (graphic recall) and write down the explanations of the agent (verbal recall).

• Redundant information improved verbal recall (not graphical) and judgments of the quality of the presentation, likeability, expressiveness, and personality of the agent.
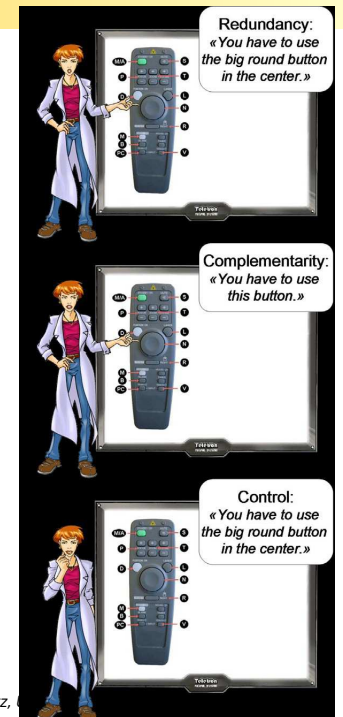
Lester et al. 1997: Effectiveness of pedagogical agents for learning and motivation

- A pedagogical agent teaching middle school children about plants.
- Q: How do different classes of explanatory behavior affect perception and learning?
- Material: 5 different versions of the system:
    1) mute
    2) task-specific suggestions (verbal)
    3) principle-based advice (verbal)
    4) principle-based (verbal+animation)
    5) full: direct suggestion + principle-based + task-specific (verbal + animation)
- Who: 100 middle school students
- Method: questionnaire about perception
    pre-test and post-test to test for knowledge increase

Lester et al. 1997: Effectiveness of pedagogical agents for learning and motivation

- test scores increased in all cases but most with the full agent and the versions that gave principle-based advice

- perception: the fully expressive version was rated highest in terms of helpfulness, believability, wanting to use it again,

Cassell & Tartaro 2008: Effectiveness of a virtual peer to improve social skills in children with autism

Q: Does interaction with virtual peers help children with autism to develop better social skills?

Method: Children with autism tell stories with a) another child and b) a virtual peer.

Who: 6 children with autism (7-11)

Analysis: contingency in the children's contributions (do their contribution follow from/relate to what's been said before
    1) Were their contributions contingent?
    2) Through which means? (repetition, elaboration, opposition, ...)
    3) topic management ((re-)introducing a topic, maintaining, shifting)

Cassell & Tartaro 2008: Effectiveness of a virtual peer to improve social skills in children with autism

Contingency:
- Children become more and more likely to make contingent contributions over the course of the interaction with the virtual peer.
- No such development over the course of the interaction with the real peer.

Topic management:
- Children use more appropriate topic management with the virtual peer than with the real peer.
- Their topic management becomes better over the course of the interaction with the virtual peer.
- No such development over the course of the interaction with the real peer.

Cassell & Tartaro 2008: Effectiveness of a virtual peer to improve social skills in children with autism

example interactions

Bickmore: Effectiveness of a virtual agent to motivate regular exercise

Daily interaction with an ECA to motivate more physical exercise.

Q: Does daily interaction with an ECA using relational behaviors increase people's motivation to exercise?

Method: Three groups interacting with the relational agent, the non-relational agent, no agent for one month.

Relational Behavior	Study Condition	
	RELATIONAL	NON-RELATIONAL
Social dialogue (Cassell and Bickmore, 2003)	Daily	None
Meta-relational dialogue (Stafford and Canary, 1991)	Frequent	Minimal
Form of address (Laver, 1981)	Friend	Stranger
Politeness (Brown and Levinson, 1987)	Indicate of small social distance	Indicative of large social distance
Empathy exchanges (Klein et al, 2002)	Daily	None
Humor (Morkes, et al, 1998)	Frequent	None
Continuity behaviors (Gilbertson, et al, 1998)	Daily	None
Nonverbal immediacy (Richmond et al, 1995)	High	Low

Bickmore: Effectiveness of a virtual agent to motivate regular exercise

- Relational agent was perceived more positively than the non-relational one (trust, likeability, respect).

- Participants who had interacted with the relational agent had a higher desire to continue interacting with the system.

- But: Exercise behavior in all three groups increased during the intervention. No difference between groups.

- Drop in exercise behavior immediately after the intervention in all groups. Especially in the relational agent group.

Today

- Social ECAs

- Evaluation

- DIY pointers

Building your own

- ECAs have a lot of (complex) parts – find somebody to cooperate with who has the pieces that you are missing.
- Some tools to download:
- full bodies, ready to be animated: GRETA, SmartBody
- talking head: RUTH
- dialogue toolkits: CSLU toolkit, Ravenclaw/Olympus, Midiki, TrindiKit, Dipper
- I will put links up on the web site.

The End

Feel free to send me emails with questions, comments, suggestions, ...

striegnk@union.edu

or

kristina.striegnitz@gmail.com