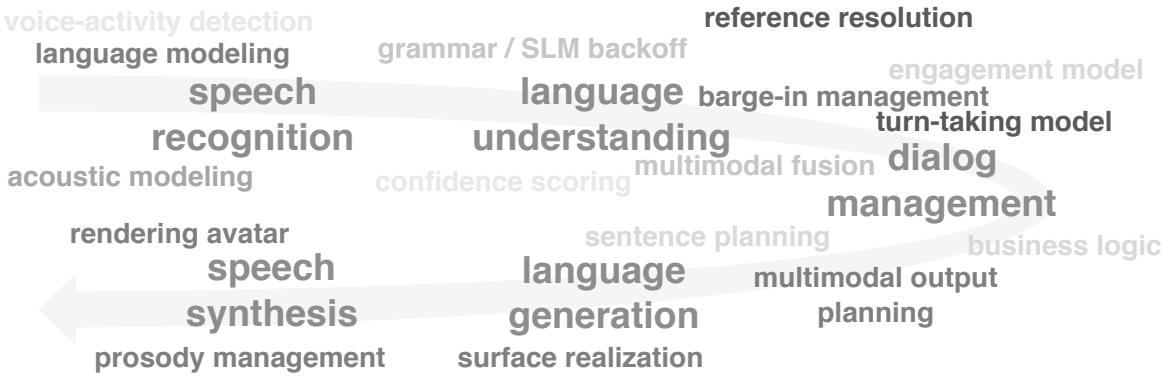
evaluation

in spoken dialog systems

evaluation

- is challenging and costly
  - other fields seem to have it better…
- why?
  - complex
  - interactive

complexity



latency reduction
F-measure for confidence annotation
sentence-level accuracy
concept-level accuracy
reduction in data transcription requirements
dialog duration
# of turns
task completion
sentence distance score
dialog failure rate

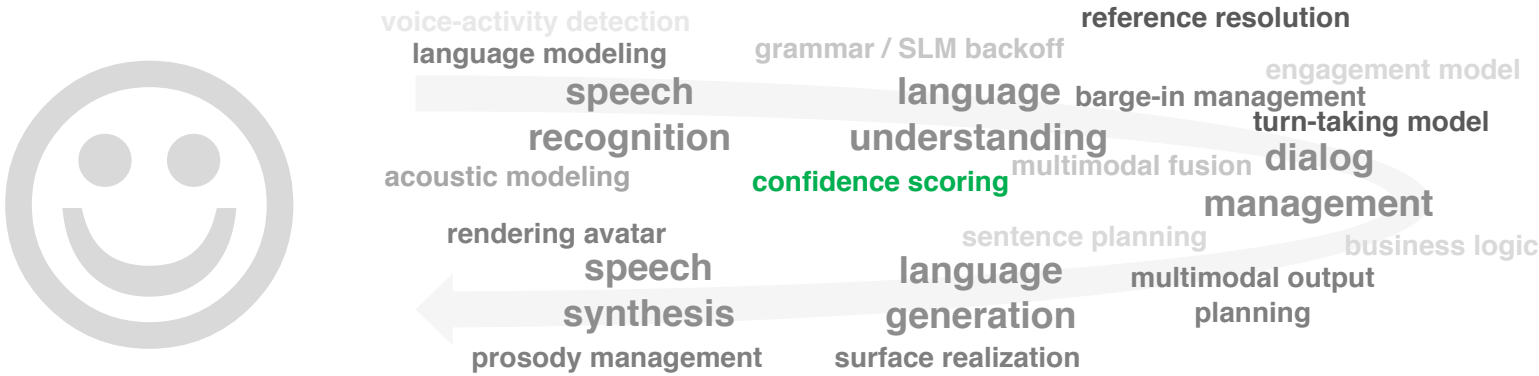
overall reward
dialog coherence
user satisfaction
ease of communication
perceived humanness
usability
clarity
persuasiveness of dialog (via Kendall τ distances)
F-score for definite exophoric reference resolution

complexity



- large design space
- some, but relatively little reuse
- lack of common basis
  - frameworks, domains, users, datasets

local evaluations



- which metrics are most appropriate?
- statistically significant?
- what does it mean?
  - is it *really* significant?
  - interactive aspect challenges soundness?
  - need global evaluation

global evaluations



- global evaluations are costly
- questionable without real users
- attempts to mitigate costs by:
  - local → global performance modeling
  - simulation

evaluation issues

- large design space: lack of common basis
  - grand challenges (common task/platform)
- local evaluations
  - is it *really* significant?
- end-to-end evaluations: costly
  - mapping from local to global
  - simulation