Dialog System Evaluation

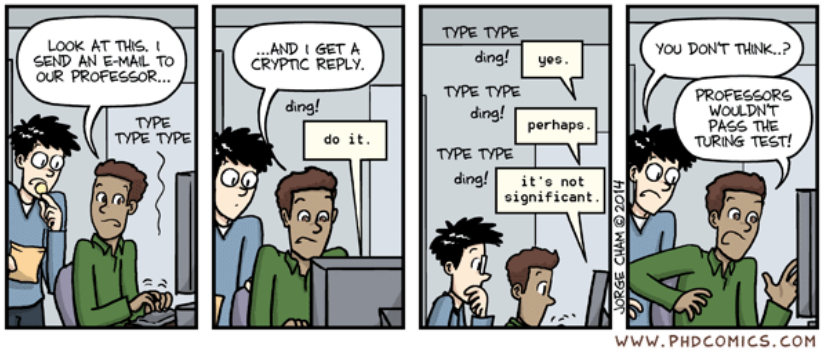
Camilo Thorne

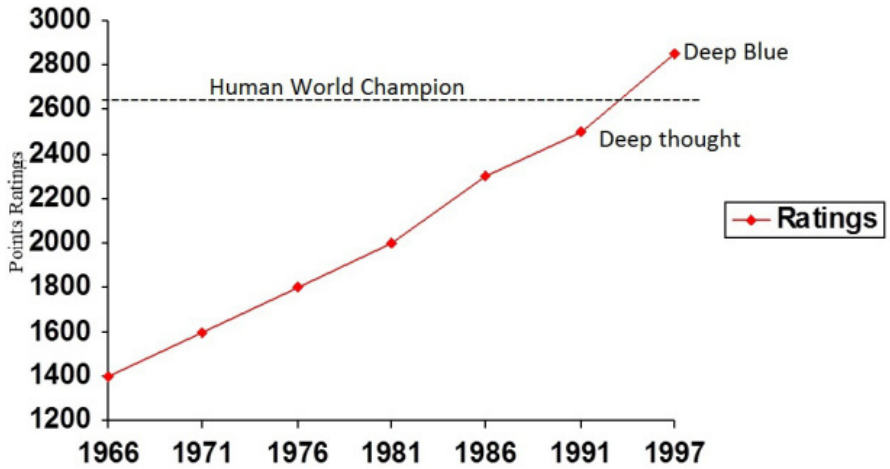
DWS Group, Universität Mannheim, Germany
camilo@informatik.uni-mannheim.de

ESSLLI 2015
Barcelona, 10.8.2015-14.8.2015

Outline

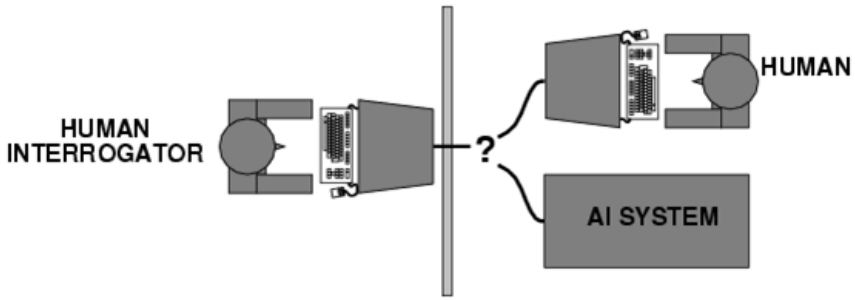
Evaluation [SA07, Gla99, WLKA97]

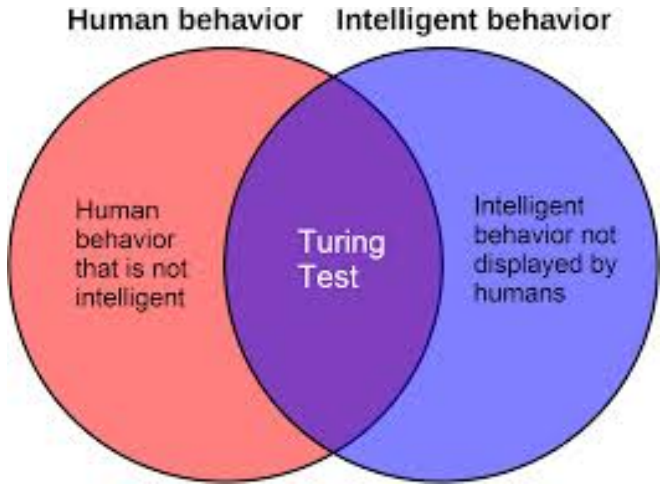




- It is common knowledge that dialog systems have been performing better and better over the years
- But: how can we actually measure such performance meaningfully?
  1. adapt tests used in AI for robots and artificial agents broadly understood
  2. adapt known performance metrics common to NLP and information systems

AI Metrics: Turing Test





- The Turing test was proposed by Alan Turing in the 1950s
- Idea: make a human tester interact (unknowingly) with a dialog system
  1. measure how long it takes to identify it as a dialog system
  2. measure how many system turns match the answers of an (arbitrary) human agent

HCI Metrics: User Satisfaction

- The user satisfaction (US) metric estimates (in usually a scale of 1 to 5) the impression the system made on end-users

judgment	points
liked it	5
rather liked it	4
neutral	3
rather disliked it	2
disliked it	1

Build table:

1. obtain numbers via user surveys
2. consider statistically representative sample of users

- Compute Cohen's/Fleischer's κ coefficient to measure agreement on scores

$$\kappa = \frac{\overline{P}_a - \overline{P}_e}{1 - \overline{P}_e} \qquad (\kappa \text{ close to } 1 \Rightarrow \text{high agreement})$$

- Run a Student t-test to understand which system shows an statistically significant gain in US

IR Metrics: Precision, Recall

- Dialog precision measures the rate of correct vs. incorrect turns within a conversation

$$P = \frac{\#(\text{correct turns})}{\#(\text{true answer turns}) + \#(\text{false answer turns})}$$

- Dialog recall measures the rate of correct vs. expected correct turns (as per the systems' knowledge) within a conversation

$$R = \frac{\#(\text{correct turns})}{\#(\text{true answer turns}) + \#(\text{false turns})}$$

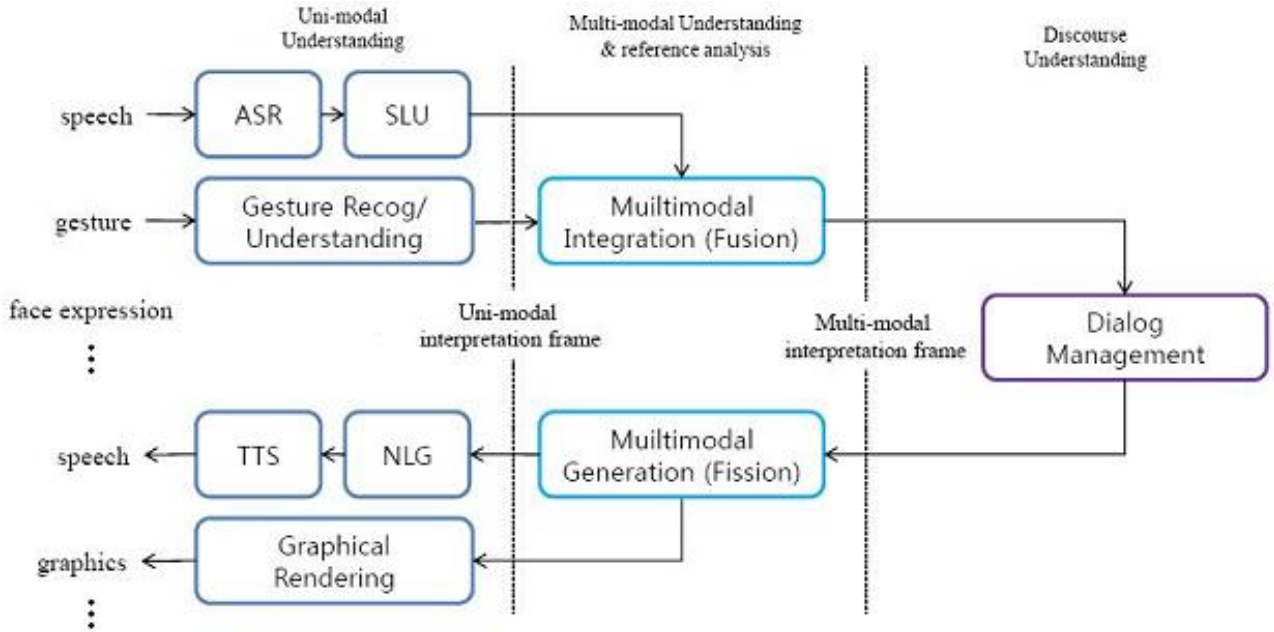
But: if the dialog is task-directed, measure task completion $\Rightarrow$ % of DS completion!

Going Beyond Rules and Chat

- There are many kinds of dialogs and dialog systems that we skipped in this course due to time constraints
- In general, they involve advanced NLP and machine learning techniques
- Dialog turns can be very hard to disambiguate due to their shortness and little explicit textual context $\Rightarrow$ rely on sensor data!
- Gestures partially overcome the sparseness of a turn's information
- Whereas in other use cases such as answering questions single or limited turn dialogs are necessary
- We will thus briefly consider two types

1. multimodal dialog systems
2. question answering (cognitive) systems

Multi-Modal Dialog Systems [GMM00]

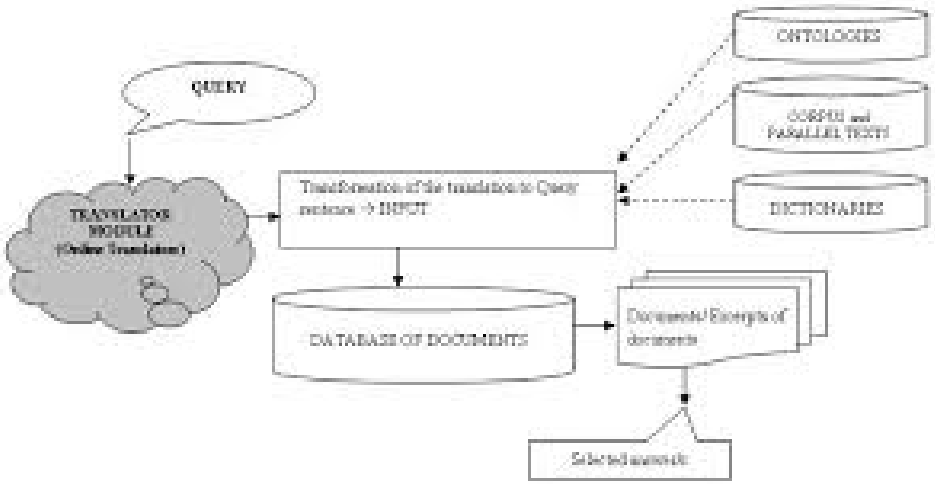


A multi-modal-dialog-system is a dialog system that in addition to speech input and output, can handle e.g.

- sensor data
- graphical data

recognizing gestures, emotions and the user's physical environment and/or location

Question Answering Systems [JM09]



View questions Q and answers A as k-dimensional word vectors

- $Q = (w_1, \ldots, w_k)^T \in W^k$
- $A = (w'_1, \ldots, w'_k)^T \in W^k$

The expected system answer to question Q maximizes similarity

$$A^* = \arg\max_A \mathit{sim}(Q, A)$$

$\Rightarrow$ where (e.g., tf-idf cosine similarity)

$$\mathit{sim}(Q, A) = \frac{\sum_i \mathit{tf\text{-}idf}(w_i) \times \mathit{tf\text{-}idf}(w'_i)}{\sqrt{\sum_i \mathit{tf\text{-}idf}(w_i)} \times \sqrt{\sum_i \mathit{tf\text{-}idf}(w'_i)}} \text{ and}$$

$$\mathit{tf\text{-}idf}(w) = \frac{\#(w \text{ occurs in corpus})}{\#(\text{corpus})} \times \left(\frac{\#(w \text{ occurs in question/answer})}{\#(\text{question/answer})} \right)^{-1}$$

Conclusions

1. We have seen techniques for measuring dialog system performance
2. We looked at three: IR-metrics, usability metrics and the Turing test
3. Each has its pros and its cons w.r.t. an specific use case $\Rightarrow$ use all ideally
4. IR-metrics and usability are more task-oriented whereas the Turing test is more domain-independent
5. We have also discussed briefly some common variants to standard dialog systems


Thank You!!!

References I

James R. Glass.

Challenges for spoken dialogue systems.

In *Proceedings of the 1999 IEEE ASRU Workshop*, 1999.

Dafydd Gibbon, Inge Mertins, and Roger K. Moore, editors.

Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology, and Product Evaluation.

Springer, 2000.

Daniel Jurafsky and James Martin.

Speech and Language Processing.

Prentice Hall, 2nd edition, 2009.

Pierre Lison.

A hybrid approach to dialogue management based on probabilistic rules.

Computer Speech & Language, 34(1):232–255, 2015.

References II

Bayan Abu Shawar and Eric Atwell.

Different measurement metrics to evaluate a chatbot system.

In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, 2007.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella.

PARADISE: A framework for evaluating spoken dialogue agents.

In *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics EACL 1997*, 1997.