The CCOBRA Framework for Benchmarking Cognitive Models

Demonstrated for the Domain of Syllogistic Reasoning

Nicolas Riesterer

April 5th, 2019

Cognitive Computation Lab,
Department of Computer Science,
University of Freiburg

Motivation

Cognitive Modeling of Reasoning

- Humans do not follow classical mathematical logics
- Psychology:
  On which principles does human cognition work?
- Artificial Intelligence:
  Benefit from versatility of human cognition (object detection, classification, reasoning)
- Modeling allows the formulation of hypotheses about latent unobservable cognitive processes
- Modeling makes assumptions testable

Goals of Modeling

Descriptive Modeling

- Psychological effects
- Statistic data models
- Algorithmic process models

→ Plausibility Arguments

⇒

Predictive Modeling

- Testing the assumptions
- Putting the models to use

→ Model Falsification

The goal of cognitive modeling is to develop accurate models with high explanatory power.

Example:
Modeling Syllogistic Reasoning

State of the Art

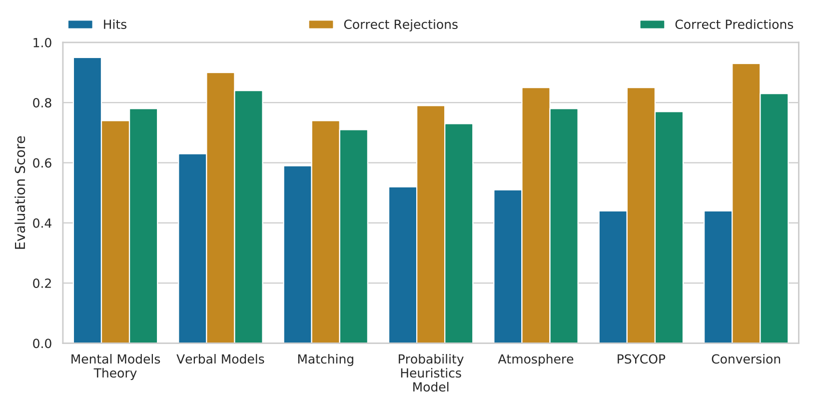
Table 1: Twelve main theories of syllogistic and monadic reasoning. Taken from Khemlani & Johnson-Laird (2012).

Heuristic Theories	Formal Rule Theories	Theories based on Models
Atmosphere	PSYCOP	Euler Circles
Matching	Verbal Substitutions	Venn Diagrams
Illicit Conversion	Source-Founding	Verbal models
Probability Heuristics	Monotonicity	Mental Models

Research Questions:
How good are they? Which account is to be preferred? How can we evaluate them?

Analysis by Khemlani & Johnson-Laird (2012)

- Data aggregation by pooling and dichotomizing conclusions (16%)
- Comparison based on hits, correct rejections, and correct predictions



Shortcomings of the Traditional Model Analysis

1. Mismatch between evaluative and interpretatory levels:
    - Aggregate evaluation used to generate insight about individuals
    - Reason: Reduce impact of noise in the data
    - Problems:
        - Lacking group-to-individual generalizability (Molenaar 2004, Fisher 2018)
2. Evaluation metrics tied to model formalisms (e.g., Bayes Factors):
    - Reason: Leverages model capabilities to their fullest potential
    - Problems:
        - Excludes incompatible model formalisms
3. Lack of general benchmark for evaluating and ranking models

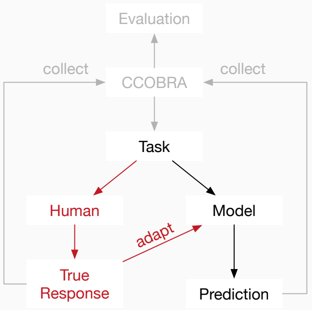
The CCOBRA Framework

Background

- Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA)
- General benchmarking framework for behavioral research
- Based on models generating predictions to individual problems
- Close connection to the underlying experimental paradigm (models simulate an experimental participant)
- No restrictions imposed on modeling methodology (probabilistics vs. logics vs. machine learning)
- Strong focus on leveraging inter-individual differences by incorporating multiple phases of learning/fitting

CCOBRA Model Interaction

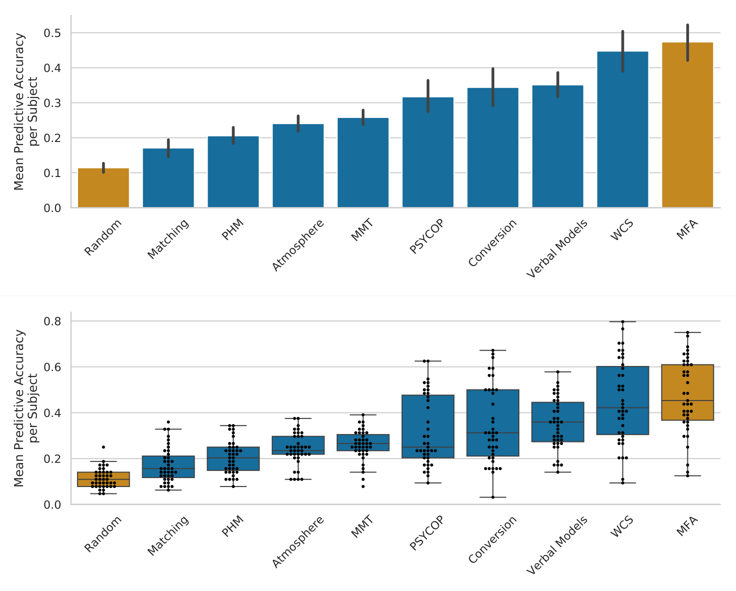
- Individual prediction scenario: Models generate a response to a given problem input
- Evaluation based on predictive accuracy (percentage of hits)
- Two model fitting phases:
  1. General fitting to training data
  2. Adaption to true response after each prediction step



Key Questions to be Tackled using CCOBRA

1. How well are models performing on an absolute scale (0-100%)?
2. Do performance-based upper bounds exist?
3. What is the impact of inter-individual differences on model performance?
4. Which cognitive properties (e.g., working memory capacity) are useful predictive features?
5. Do model formalisms generalize across domains?

Evaluation Results



Observations

- Even most frequent response (MFA) is not a useful predictor for individual behavior
- Cognitive models below MFA, because they do not integrate inter-individual differences yet
- Raw performance values are suboptimal with WCS scoring highest at 44%
- Questions:
  - Can we really claim we have grasped human syllogistic reasoning?
  - Is lacking accuracy result of noisy data (fatigue, lacking concentration) and as such cannot be modeled?
  - How much potential is left in the domain?

Empirical Upper Bounds

Machine Learning to Determine Upper Bounds

- Leverage general pattern recognition capabilities of data-driven ML methods
- Identify empirical upper bounds pointing to the levels of performance cognitive models should achieve

Recommender Systems

- User-based Collaborative Filtering
- Item-based Collaborative Filtering

Neural Networks

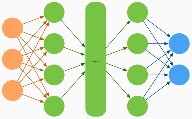
- Feed-Forward Multi-Layer Perceptron
- Autoencoder
- Recurrent Neural Network

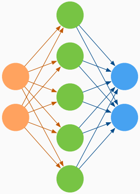
Recommender Systems (Collaborative Filtering)

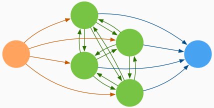
- Maintain database of users
- Prediction based on some form of similarity
- Related to nearest-neighbor approaches
- User-based Collaborative Filtering (UBCF):
  - Predictions based on similar users
  - "Users similar to you have answered..."
- Item-based Collaborative Filtering (IBCF):
  - Predictions based on item-response dependencies
  - "Users who respond with X to task A also respond with Y to task B"

Neural Networks

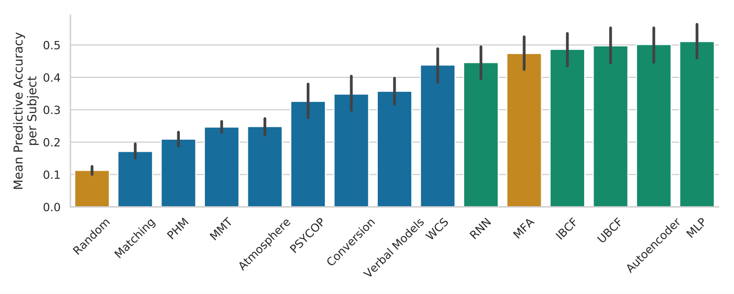
- Prediction based on patterns in the data
- Adaptive Multi-Layer Perceptron (MLP):
  Regular feed-forward network continuously fitted to true responses
- Denoising Autoencoder:
  Imputes missing input information by learning associations in the data
- Recurrent Neural Network (RNN):
  Trained on the experimental task sequence. Bases predictions on sequence effects in the data.

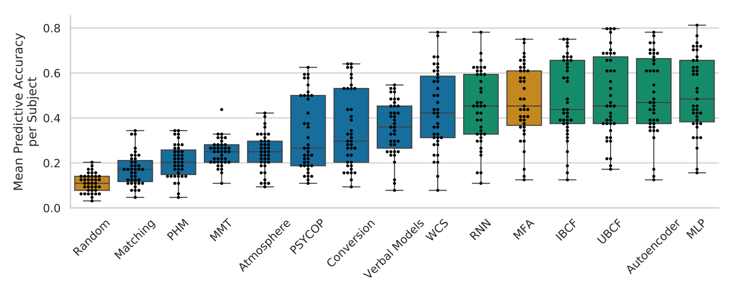






Upper Bound of Model Performance





Observations

- Adaptive ML models are able to exceed MFA because they incorporate different approaches to individualization:
  - Autoencoder: Maintains a reasoner profile consisting of all previously given responses
  - MLP: Fits to true responses after each prediction step
  - UBCF & IBCF: Maintain user-profile to identify similarly behaving reasoners and tasks
- Similar levels of accuracy reached by most ML methods (not substantially beyond MFA)
  - Explanation 1: Unobserved or unmodeled factors (e.g., working memory capacity, fatigue, experience)
  - Explanation 2: Noise, i.e., effects unrelated to (realistically) observable factors (e.g., influence of individual brain anatomy)

Conclusion

- Traditional cognitive modeling focusing on aggregated data is rapidly approaching the MFA border
- Future work should shift its perspective towards modeling individuals instead of groups
- Benchmarking based on directly interpretable results: "model is able to account for X% of given responses" (performance ranking)
- Data-driven methods reach an upper bound of performance slightly above MFA
  - Lacking information density of the data?
  - Inconsistent response behavior even within individuals?

Future Work

- Integration of individual differences
  - Demographic information (e.g., age, educational background)
  - Cognitive properties (e.g., working memory capacity)
- Extension of the domain to enrich the data
  - Cross-Domain Modeling: Models have to cope with multiple domains
  - Extended syllogisms: Higher number of premises requiring true generalizability to unseen problem instances from models
- Enhanced integration of high performing and highly explainable models

PRECORE Challenge

- Two competitions on developing models for predicting human syllogistic reasoning:
  - IJCAI 2019: Part of the *Bridging the Gap* workshop series
  - CogSci 2019
- More information on our website:

  `https://www.cc.uni-freiburg.de/modelingchallenge`

- Contact mail address:

  `precore2019@cs.uni-freiburg.de`

Thank you for your attention!

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27), E6106-E6115.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological bulletin*, 138(3), 427.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4), 201-218.