Project Report

Predictive Models for Individual Human Reasoning

Leon Kaltenbrunn

Examiner: Prof. Dr. Marco Ragni
Adviser: Nicolas Riesterer

Albert-Ludwigs-University Freiburg
Faculty of Engineering
Department of Computer Science
Cognitive Computation Lab

November 9, 2020

Abstract

This report presents an evaluation of the cognitive model "mSentential" for human propositional reasoning in terms of predictive accuracy on individual responses. By contrasting the model with statistical baselines such as random guessing or the most frequently selected response and a pure logic-based model as well as a machine learning model, we gain understanding of the extent to which the model performs in a propositional reasoning task. To further investigate the model's potential, we propose some optimization options and also discuss what can be done to further improve performance.

Contents

1 Introduction

Humans reason about many things. Especially about facts, possibilities, and probabilities (Khemlani et al., 2018, p. 1). Among other things the field of cognitive sciences is trying to understand the reasoning process. One theory is that human behavior is a set of mental simulations of situations in the real world (Johnson-Laird, 1975). From that, cognitive sciences inferred systematic patterns in the reasoning process, creating multiple cognitive models with different approaches of explaining human reasoning. It has been found that using a pure logical model that always gives the logic answer does not represent the human answering behavior (Johnson-Laird et al., 2015). This is due to the fact that human reasoning is not based on pure logic (Leighton, 2004) and even tasks that seem straightforward at first glance may turn out to be beyond the reach of some subjects. Thus, a theory about the reasoning process must explain the difficulties encountered by the unsuccessful subjects as well as the correct procedures of the successful ones (Rips, 1983). To test their theories, cognitive scientists create simple factual inferences that can look like this:

If the card is an ace, then it is a heart.
The card is an ace. (1)
Therefore, the card is a heart.

Words such as “if”, “or” and “and” connect the sentence together. Reasoning with such connected sentences is called sentential or propositional reasoning (Khemlani et al., 2018). One of the most important cognitive reasoning theories based on the so-called mental model theory.

The goal of this project is to analyze the cognitive model called "mSentential" (Johnson-Laird & Khemlani, 2017) which is a concrete implementation of sentential reasoning, against a statistical and a machine learning based baseline. It has been shown, in other domains, that cognitive theories are often not capable of generating accurate predictions of individual human participants (Riesterer et al., 2020). This is why we want to evaluate the predictive performance of mSentential and its potential in predicting for propositional reasoning while measured in concrete benchmarks and create a context with our baselines to accurately assess the generated results. Recent work has also shown that mSentential in particular is well suited for modal reasoning tasks (Guerth, 2019). Because modal reasoning augments propositional reasoning by adding operations about necessity and possibility (Lewis, 1912) the model is expected to perform similar on a propositional reasoning task.

First, we are going to take a look at the state of research and some theoretical background to understand why and how cognitive models work and how the model mSentential in particular tries to emulate sentential reasoning. Then we are going to test the model against our baselines and present an interpretation of the empirical result. In the final section we then discuss the problems that arise and the conclusion of the project.

2 State of Research and Theoretical Background

2.1 Mental Models

In cognitive science, a hypothesis for human reasoning is that humans create mental simulations of situations that can be described with a so-called mental model (Johnson-Laird, 1975). Those mental models are based on a small set of fundamental assumptions about the given situation and so, each of these models represents one possibility of outcome, capturing what is common to all the different ways in which the possibility may occur (Johnson-Laird & Byrne, 2002). Furthermore, mental models are based on the principle of truth. That means, they typically only represent what is possible in the given context. Although it is possible for a mental model to represent what is false, temporarily assumed to be true for example in counterfactual thinking (Byrne, 2005).

Mental models in reasoning establish validity by ensuring that the conclusion holds true over all of the model's premises (Jeffrey, 1981, p. 1). To refute invalid interferences the model needs to rely on valid counter examples to exist. So, a human that reasons tend to reject a conclusion if they find a valid counterexample. Or in other words, if they find a possibility in which the premises hold, but the conclusion does not (Schroyens et al., 2003; Verschueren et al., 2005).

2.2 mSentential

2.2.1 Building of Mental Models

Now that *mental models* have been introduced, let's look how they are built and how they can be fleshed out into *fully explicit models*. As stated in the introduction factual inferences like this:

If the card is an ace, then it is a heart.
The card is an ace. (2)
Therefore, the card is a heart.

consist of some compound assertions and are connected via sentential connectives. For simplification we can rewrite the inferences by substituting the problems facts/content with abstract symbols as follows:

If A, then B.
A (3)
Therefore, B

From this we can extract our first *single mental model*. The compound assertion "If A, then B" represent the set of the models "A" and "B". Then the assertion "A" obviously has the model "A" in it. Compound assertions with only a single model refer to facts, whereas compounds with multiple models refer to conjunctions of default possibilities (Khemlani et al., 2018, p. 9). Reasoning now depends on conjoining the sets of models for the different assertions as seen in Table 1. The process is then simple in principle (Johnson-Laird, 2006, Ch. 8). Two sets of models are conjoined pairwise to form their product. So, when models are consistent in both sets, the result is a model of all propositions represented in both models (Khemlani et al., 2018, p. 9). However, if they are not a null model is generated. The null model represents a contradiction and therefore no longer has an impact on the result.

From our example a conjoined product for the *single mental models* will look as follows:

Table 1: Example of how single mental models are created.

If A then B			**A**		**Product**	
A	B	&	A	→	A	B

As previously mentioned, models represent only possibilities, not impossibilities (Johnson-Lair & Savary, 1999). But in our example one of the premises is an "if-clause", so our model does not make explicit the possibilities in which the "if-clause" is false. This is where *fully explicit models* come in and use negation to represent clauses in the premises that are false as seen in Table 2.

Table 2: Example of how fully explicit models are created.

If A then B			**A**		**Product**
A	B	&	A	→	A B
¬ A	¬ B	&	A	→	null model
¬ A	B	&	A	→	null model

The result in both cases is the product "A, B", but obviously that doesn't always have to be the case. See Table 1 from "Facts and Possibilities: A Model-Based Theory of Sentential Reasoning" by Khemlani et al. (2018) for further information on possible cases. The so called "model theory" now assumes two Systems for reasoning. The first one called System 1 uses *single mental models* and the other called System 2 uses *fully explicit models*.

The so called "new model theory" builds on top of the model theory creating the principles of sentential reasoning as seen in the next section. For a in depth look at the "new model theory" we refer the reader to the paper "Facts and Possibilities: A Model-Based Theory of Sentential Reasoning" by Khemlani et al. (2018).

2.2.2 Principles of Sentential Reasoning

The cognitive model mSentential implements the principles of sentential reasoning in common Lisp. The principles of sentential reasoning (Khemlani et al., 2018, Table 3) are that

i) Representation: reasoners interpret compound assertions as conjunctive sets of possibilities so they should draw modal conclusions from non-modal premises e.g., A or B or both. Therefore, possibly A.

ii) Inference: Necessary inferences are those in which the models of the premises support all and only the models of the conclusion. So, reasoners should reject inferences in which the premises do not support one of the models of the conclusion e.g., A or B but not both. Therefore, A or B or both.

iii) Dual Systems: Intuitive inferences depend on mental models and deliberative inferences depend on fully explicit models. So that mental models should lead to fallacies in certain cases, e.g., One of these assertions is true and one of them is false: A and B. B or else C. Therefore, it is possible that A and B.

iv) Modulation: Background knowledge blocks the construction of possibilities and can add relations. So that reasoners should interpret ambiguous disjunctive constructions, e.g., A or B, as exclusive disjunctions when the contents block the model of A and B

v) Verification: It depends on relations between the evidence and models of assertions. So that intuitions should evaluate some evidence as irrelevant to the truth or falsify of conditionals.

2.2.3 Program Architecture

Implementing these principles from the new model theory results in a dual system architecture, as seen in Figure 1 below.

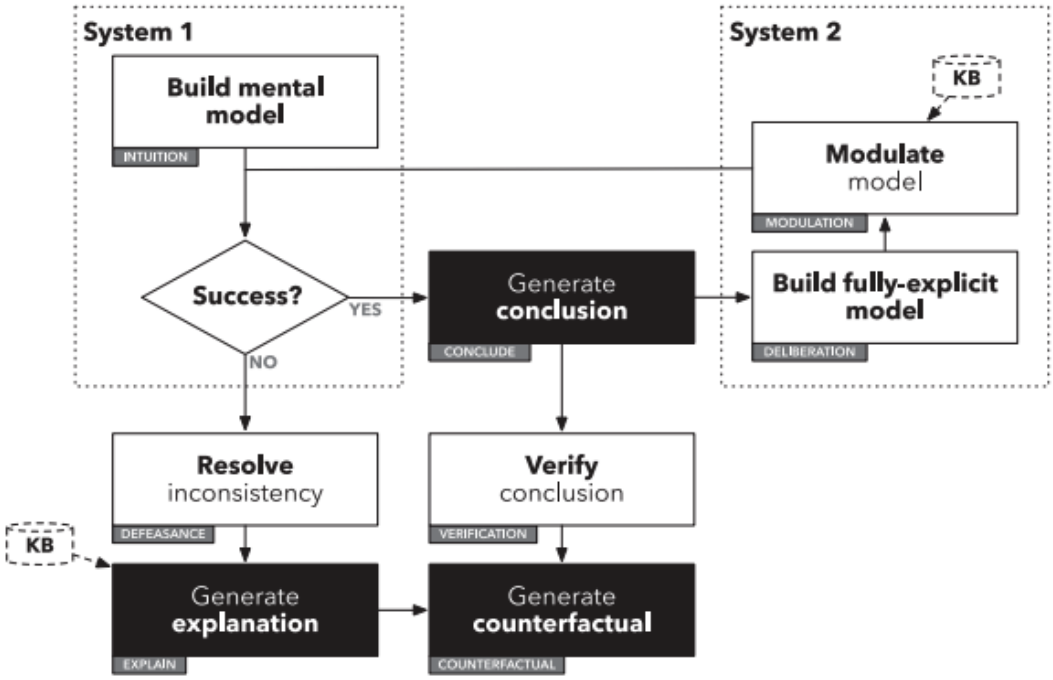

Figure 1: A diagram of the reasoning program implementing the new model theory taken from (Khemlani et al., 2018, Fig. 1.)

In Figure 1, the terminology System 1 and System 2 represents the intuitive and deliberate thinking parts of the brain, respectively (Kahneman, 2011). System 1 bases its conclusions on a *single mental models* . Whereas the deliberative component is the heart of System 2 for reasoning with *fully explicit models*. The white boxes show the main components of the program with the black boxes representing the output the program generates. The box with the "Success" label denotes whether the System constructed a non-null model, indicating if a conclusion can be generated. The "KB" labeled cylinders represent a knowledge base wherefrom the program looks up information. This represents a kind of background knowledge the model can have.

3 Approach

To approach the subject of evaluating the performance of our cognitive model we first need to take a look at our evaluation process and our data. The data consists of different participants finding conclusions for inferences in a random sequence. Because every participant is giving about 30 inferences and there are 16 different inferences in total, every participant has to find a conclusion to a specific inference more than ones. To find a conclusion to an inference a participant is given four out of five possible conclusions to choose from. To evaluate our models, we use the Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) (Riesterer & Brand, 2018) Framework. It provides the possibility to generate a predictive performance measure for individual participants by evaluating the model's prediction against the response given by the participant in a function called "predict". Models evaluated in the CCOBRA Framework, are also able to gain information about the inferences and responses given by other participants in a function called "pre training". The "adapt" function can be used to adapt to the response patterns of the current participant.

3.1 Baseline

To analyze the effectiveness of the model mSentential, we first need to assess a baseline of what can be a lower and an upper bound, to create a context for our results.

3.1.1 Random Model

To create a lower bound the "RandomModel" assumes a uniform distribution between the possible conclusions and then randomly samples one to create a baseline that should be outperformed by every other model. In our data for propositional reasoning, we always have four different possibilities for the model to choose from, which results in an average 1/4 probability for the "RandomModel" to predict a correct conclusion.

3.1.2 Logic Model

Because of the nature of the task, it is useful to provide another baseline to compare against. A logic solver should give us useful insights in how accurate our reasoners compare against a pure logical approach. This model was created as part of the Bachelor project by Giessel (2019a).

3.1.3 MFA Model

On the upper end we have the "MFAModel". MFA stands for Most Frequent Answer and it creates a response distribution from the data for each possible answer and then returns the one with the highest probability. This approach is optimal if no background information about the individual participants is added (Riesterer et al., 2020). So, to beat the accuracy of the "MFAModel" we can assume that it is necessary to add some amount of knowledge about the individual. This model was also created as part of the Bachelor project by Giessel (2019b).

3.1.4 UBCF Model

Because cognitive models need to be plausible in explaining the reasoning process, they are limited in terms of predictive performance. To show an example for what can be achieved in terms of predictive performance we created a UBCF Model to compare our mSentential against. UBCF stands for "**U**ser **B**ased **C**ollaborative **F**iltering" and is a technique used by recommender Systems to automatically predict or filter the preferences of users (Ricci et al., 2011). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. So, the UBCF model, in a more general sense tries to filter information or patterns involving collaboration of multiple participants. Note that participants aren't actually collaborating together, the model just assumes a similarity between participants who's answering patterns look similar.

As already mentioned, the UBCF model generates a similarity between participants by looking for patterns in answering behavior. This is done by creating a database of all other participants and their answering behavior in pre training. Also, all previous predictions of the current participant are held to then compare against the database. Then, every time a new prediction has to be done, the UBCF computes the collaborative prediction vector by comparing the top k similar participants to the current one and then making a prediction by choosing the most likely answer between those top k participants. This prediction can be modified by another parameter called *exp*. With this parameter the similarities of the top k participants are modified so that the impact of the not so similar ones is reduced. So, increasing this parameter reduces the impact of the not so similar participants. After some testing, we found out that the best values for k and *exp* are $k = 12$ and $exp = 0.1$ for this particular dataset.

3.2 mSentential

With our baseline established we then want to modify our cognitive model mSentential for propositional reasoning. This implies analyzing the already implemented version of the model by Guerth (2019) for modal reasoning and modifying it for our use case. The model when given a inference and a indicator if System 1 or 2 should be used, returns a tuple of possible and necessary conclusions by using the "What follows" function. So, we need to apply some basic filtering to get a single useable conclusion. If the used system returns just one conclusion, for example there is one possible and no necessary conclusion that follows the given inference, then we just return that conclusion. But if the system has multiple conclusions, we need to apply some method to figure out which of the conclusions is the correct one to return. After trying multiple methods prioritizing necessary over possible conclusions or picking only possible or necessary conclusions, always picking a random conclusion out of the ones given by the model was the most effective. Finally, if the model returns an empty tuple then "nothing" is predicted. With the basic modifications done, we result in two variants of the model outlined in the next sections.

3.2.1 Paper Variant

As stated by Khemlani et al. (2018) we implemented a parameter σ to control the probability that System 2 is engaged in making inferences. As we can see from the predictive accuracies from Figure 2 there is a clear correlation between an increased σ parameter and a higher accuracy. Ranging from $\sigma = 0$ with 35.10% to $\sigma = 1$ with 68.24%. Hence, we can make the conclusion that System 2 and with that, using *fully explicit models* is inherently better for our predictive use case and the data we are working with.

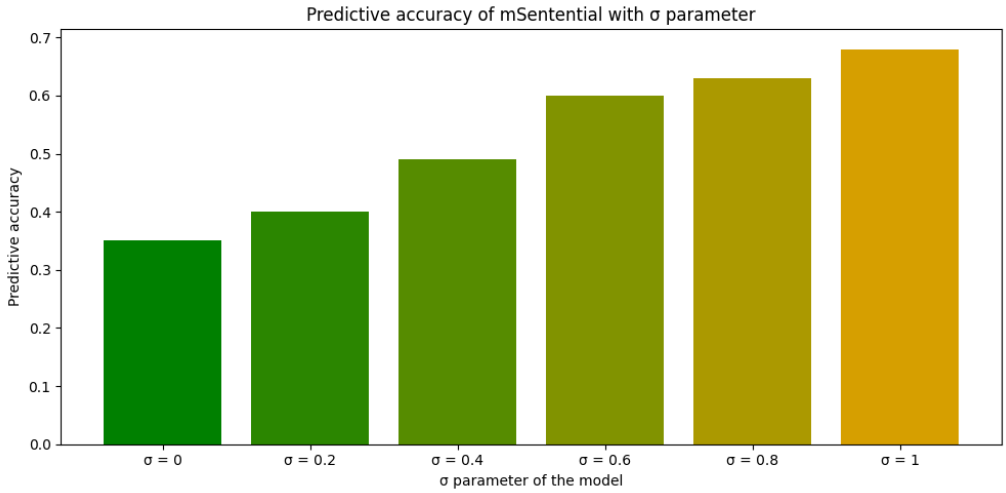

Figure 2: Predictive accuracy of mSentential paper variant with σ parameter value in 0.2 increments.

So, when we put these results in comparison with our baseline (Figure 3), even with a σ value of 1 where not able to achieve a high enough accuracy to surpass the logic or MFA model. But as we previously discussed, the decision process of choosing a likely conclusion is in part a random one. This and the fact that the *fully explicit models* generated by System 2 have a higher likelihood of predicting multiple conclusions led us to the believe that optimizations can be done to further improve performance.

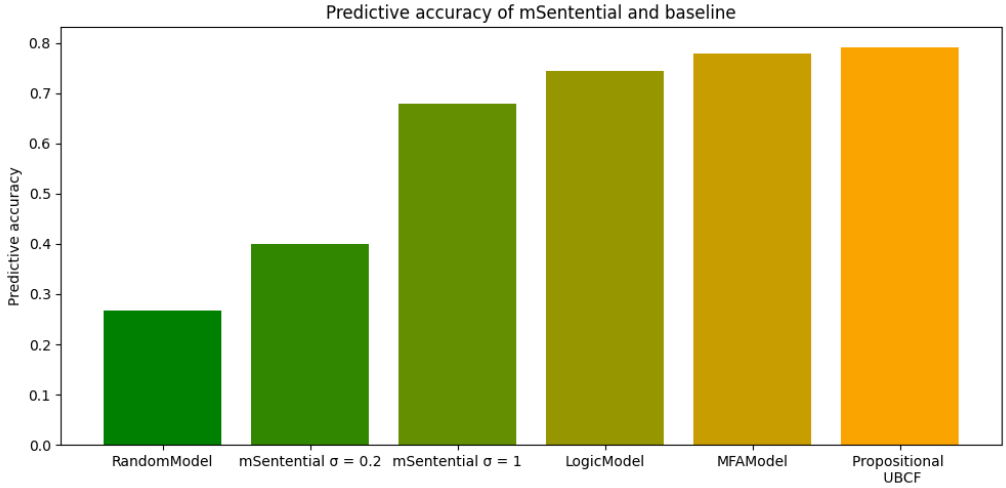

Figure 3: mSentential paper variant in comparison to the baseline.

3.2.2 Optimized Variant

Because of the partially random decision process and the higher likelihood of predicting multiple conclusions of System 2, we can optimize the model to improve performance. Generally, System 2 performs much better than System 1, so System 2 always provides the prediction, for this model. Further, a list of answered questions and responses is kept. With that said, a truth value of answered "nothing" is saved as well. This works as an indicator if a participant is not able to comprehend a given task and with that, might not be able to comprehend similar tasks. With those basic improvements always applying we then created some optimization options:

1. Consistence: If System 2 predicts multiple possible answers, this makes sure if System 1 has predicted an answer that it chooses the answer that is predicted by both systems. This provides prediction consistence between both systems.

2. Necessary: mSentential provides both necessary and possible conclusions for a given premise. Some participants may only provide an answer if it follows necessarily. So, with that option it is possible to return "nothing" if nothing follows necessarily.

3. Size limit: Possible working memory size limit. This option, when enabled checks the task for more or equal to 3 sentential connectives because this may lead to working memory overload of the participant and nothing is predicted.

4. Memory: If this option is enabled the list for answered questions and responses is used to predict answers if a question was already answered before.

Now to evaluate the optimization options we generated a model for each permutation of those options to see what works best and how those options might affect each other. To average out the leftover randomness in some permutations we created ten

different runs of all permutated models and took the mean of all those runs as seen in Figure 4 and 5.

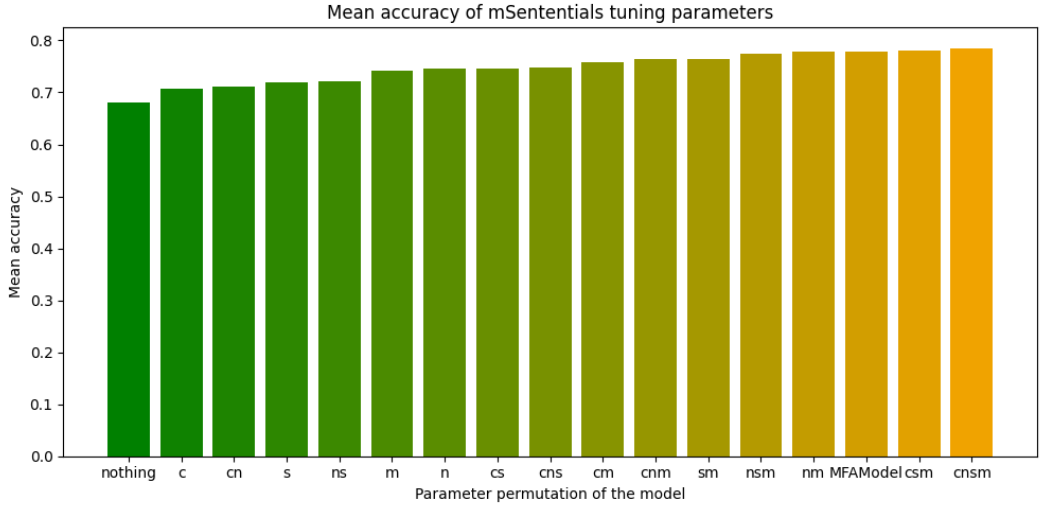

Figure 4: All parameter permutations of the optimized mSentential model also, MFA and "nothing" meaning no options enabled for reference **c** = consistence, **n** = necessary, **s** = size limit, **m** = memory.

If a parameter is active the parameter is listed as the model name. This means, for example in the model **cnsm** all parameters are active, whereas in model **cs** only the options "consistence" and "size limit" are active.

So first we can see that if we enable all options the model has the highest accuracy. This implies that the model was able to learn something about the individual, which is indicated by a higher accuracy then the MFA Model. Second, our "nothing" or $\sigma = 1$ paper model is lowest performing model. This means that every optimization has indeed improved performance in some form. But as we can see in Figure 5 some optimizations are affected by randomness and, or by other optimizations reducing performance in specific instances. Nevertheless, if we enable more and more optimizations randomness starts to disappear, indicated by reduction in size of the interquartile range of the individual models, seen in the boxplot.

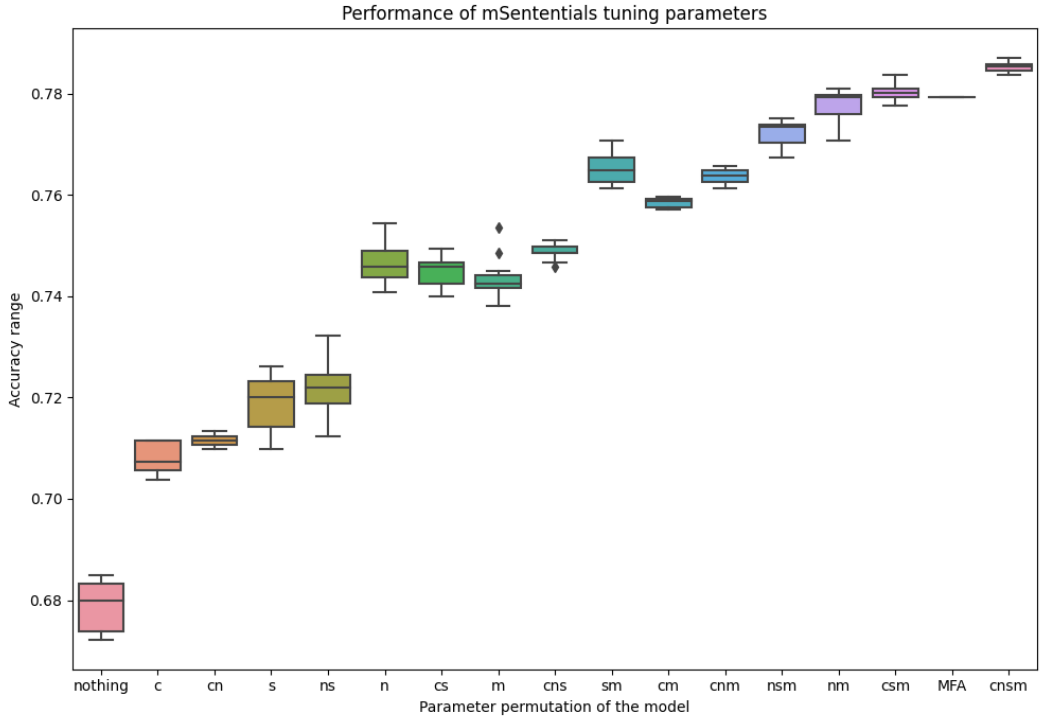

Figure 5: Boxplot of ten runs showing all parameter permutations of the optimized mSentential model.

3.3 Results

As we can see in Figure 6 our final results look as expected. With every other model outperforming the random model at 25.23%, mSentential in its unoptimized paper variant with $\sigma = 0.2$ at 42.74% and $\sigma = 1$ at 68.84% coming in below the Logic and MFA model with 74.42% and 77.93% respectively. Then with all optimization parameters active the mSentential tuned model with 78.71% comes right behind the UBCF Model for our propositional task at 79.22%.

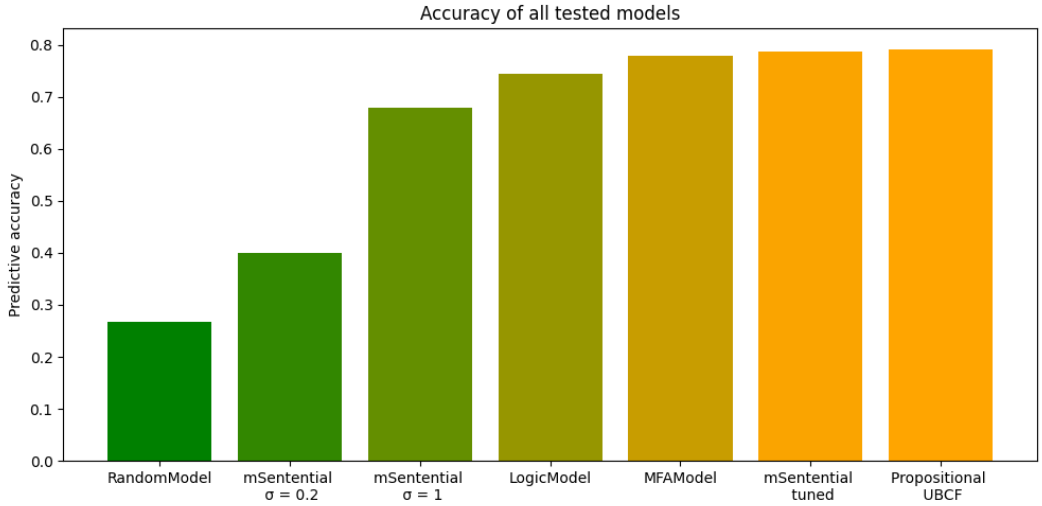

Figure 6: Final predictive accuracy results of all models

In summary, these final results show that a current upper bound in performance might be located at a predictive accuracy of about 80%. The fact that the cognitive model mSentential in its paper variant arrived at about 69% highlights that improvements to the underlying cognitive model can still be made. Some improvement possibilities where tested and implemented, even to surpass the MFA's performance at 78%. However, with only around a 1% gap the mSentential optimized model is only slightly better. With that said even a machine learning based model stagnated shortly before 80%. While this could be due to the simplicity of the machine learning based model, it could also indicate that the purely response-focused data are approaching an upper bound.

4 Discussion

In this project, we compared several different models to predict individual human reasoning while performing a propositional or sentential reasoning task. mSentential with its two-systems architecture ended up performing reasonably well in this particular use case, only after we created some optimizations and mostly ignored the intuitive System 1 part of the model. This can be due to many reasons but should be an indicator that more optimizations to the cognitive model can be done. Looking at the optimizations we implemented, we were able to surpass the performance of the most frequent answer model. This means, as previously stated, that we were able to learn some information about an individual's reasoning processes. Finding optimal ways to integrate this information into our models is key for achieving even higher accuracies. With that said, we are limited in terms of our data regarding number of inferences and data quality about those crucial interindividual differences. With only sixteen inferences and no additional information's about the individual participant in our dataset, further testing with multiple improved datasets needs to be done to mitigate those problems. Also, as seen in the plot comparing the different optimization options (Figure 4), some implemented optimizations counteract each other, which maybe an indicator that improvements to those can be made as well. However, with all those optimizations we also have to ask ourselves if the resulting model is a pure cognitive model anymore. One could argue that the model with all the decisions done after the initial predictions by the cognitive model have been made, it is now a hybrid between a statistical and a cognitive one. Knowing that,

there might be a better cognitive model for the propositional reasoning task. The model "mReasoner" (Johnson-Laird & Khemlani, 2016) might be a good start for some future work.

As for the UBCF, reaching an upper boundary at around 80% can be argued as well. User based collaborative filtering because of its approach seems to work for our particular use case and data but might not be the optimal approach for figuring out a true upper bound. Training a neural network or choosing a different machine learning model might result in a higher accuracy then 80%. But UBCF is also a promising model, so it might be possibility to improve performance by generating a new similarity function to determine a higher similarity between participants. With this or optimizations to the data collection process, to feature new mechanism, to then determine a better similarity, it might be possible to improve past the 80% mark.

5 Conclusions

In conclusion the cognitive model mSentential performs reasonably well for a proposition reasoning task. Nevertheless, it can be improved by different optimization methods. With some proposed optimizations the model was able to surpass the most frequent response model, which means that it was able to learn something about the individual. Comparing it with different statistical and machine learning based models yielded insights, such as that optimizations in the data collection process and increasing the number of inferences is necessary to further improve the task of predicting individual human behavior in propositional reasoning.

References

Byrne, R. (2005). The rational imagination: How people create counterfactual alternatives to reality.

Giessel. (2019a). *Logic model.* (`https://github.com/CognitiveComputationLab/cogmods/blob/master/propositional/student_projects/giessl2019/models/logic_model.py`)

Giessel. (2019b). *Mfa model.* (`https://github.com/CognitiveComputationLab/cogmods/blob/master/propositional/student_projects/giessl2019/models/mfa_model.py`)

Guerth. (2019). *mmodalsentential.* (`https://github.com/CognitiveComputationLab/cogmods/tree/master/modal/student_projects/2019_guerth`)

Jeffrey, R. J. (1981). Formal logic: Its scope and limits (2nd ed.). *New York: McGraw-Hill.*.

Johnson-Lair, P. N., & Savary, F. (1999). llusory inferences: A novel class of erroneous deductions. *Cognition*, *71*, 191-229.

Johnson-Laird, P. (1975). Models of deduction. In Falmagne (Ed.), *Reasoning: Representation and process* (1st ed., p. 7-54). Springdale, NJ: Erlbaum.

Johnson-Laird, P. (2006). How we reason. *Oxford University Press*.

Johnson-Laird, P., & Byrne, R. (2002). Conditionals: a theory of meaning, inference, and pragmatics. *Psychol. Rev.*, *109*, 646–678.

Johnson-Laird, P., & Khemlani, S. S. (2016). *mreasoner: A unified computational implementation of the model theory.* (`https://www.modeltheory.org/models/mreasoner/`)

Johnson-Laird, P., & Khemlani, S. S. (2017). *msententialv1.1.* (`http://modeltheory.org/programs/mSentential-v1.1.lisp`)

Johnson-Laird, P., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*(4), 201-214.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Khemlani, S., Byrne, R., & Johnson-Laird, P. (2018, 8). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science*, *42*(6), 201-213.

Leighton, J. P. (2004). The nature of reasoning. *Cambridge University Press*, 3-5.

Lewis, C. I. (1912). Implication and the algebra of logic. *Mind*, 522–531.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. *Recommender Systems Handbook*, 1-35.

Riesterer, N., & Brand, D. (2018). *Cognitive computation for behavioral reasoning analysis (ccobra) framework.* (`https://github.com/CognitiveComputationLab/ccobra`)

Riesterer, N., Brand, D., & Ragni, M. (2020). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, 1–15.

Rips, L. J. (1983). Cognitive processes in prepositional reasoning. *Psychological Review*, *90*(01), 38-71.

Schroyens, W., W.Schaeken, & Handley, S. (2003). In search of counterexamples: Deductive rationality in human reasoning. *Quart. J. Exp. Psychol.*, *56(A)*, 1129–1145.

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). Everyday conditional reasoning: A working memory-dependent tradeoff between counterexample and likelihood use. *Mem. Cognit.*, *33*, 107-119.