How can Cognitive Modeling Benefit from Ontologies? Evidence from the HCI Domain

Marc Halbrügge[1], Michael Quade[2], and Klaus-Peter Engelbrecht[1]

[1] Quality & Usability Lab, Telekom Innovation Laboratories,
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin
marc.halbruegge@tu-berlin.de, klaus-peter.engelbrecht@telekom.de
[2] DAI-Labor, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin
michael.quade@dai-labor.de

Abstract. Cognitive modeling as a method has proven successful at reproducing and explaining human intelligent behavior in specific laboratory situations, but still struggles to produce more general intelligent capabilities. A promising strategy to address this weakness is the addition of large semantic resources to cognitive architectures. We are investigating the usefulness of this approach in the context of human behavior during software use. By adding world knowledge from a Wikipedia-based ontology to a model of human sequential behavior, we achieve quantitatively and qualitatively better fits to human data.The combination of model and ontology yields additional insights that cannot be explained by the model or the ontology alone.

Keywords: cognitive modeling, ontology, human performance, human error, memory for goals

1 Introduction

Cognitive architectures like Soar [13] and ACT-R [2] have enabled researchers to create sophisticated cognitive models of intelligent human behavior in laboratory situations. One major drawback of cognitive modeling, especially from the artificial general intelligence perspective, is that those models tend to be very problem-specific. While a cognitive model of air traffic control may show human-like intelligence in exactly that task, it is completely unable to perform anything else, like solving a basic algebra problem. One major cause of the thematic narrowness of cognitive models is the restricted amount of knowledge that those models have access to. In most cases, every single piece of information has to be coded into the model by a researcher. This has been critized before, as a human cognitive architecture should be able to maintain and integrate large amounts of knowledge [3].

One recent approach to overcome this issue is the combination of existing cognitive architectures with large knowledge databases like WordNet [8, 7, 6, 16] or DBpedia [14], a Wikipedia-based ontology [19]. Common to all those approaches is that they focus on feasibility and the technical implementation of

their knowledge system, while the validity of the resulting architectures is still an open question.

This is the starting point for the research project presented here. Instead of describing how vast knowledge bases can be added to a cognitive architecture, we combine an existing solution with an existing cognitive model of sequential behavior and analyze how the predictions of the model change and whether this adds to our unterstanding of the model, its task, and the underlying knowledge base.

Our research is situated in the human-computer interaction (HCI) domain. We are analyzing how long human users need to perform simple tasks with a home assistance application, how often they make errors, and which user interface (UI) elements are tied to these errors. Our cognitive model receives knowledge about the world based on Wikipedia content, following Salvucci's work on the integration of DBpedia into ACT-R [19]. The modeling effort presented in this paper relies mainly on the general relevance of different Wikipedia articles. The higher the number of links inside Wikipedia that point towards an article, the higher the relevance of the article and the entity or concept that it explains. Our data suggests that UI elements that correspond to highly relevant concepts are handled differently than elements that correspond to less relevant concepts.

1.1 Human Action Control and Error

The link from human error research to artificial intelligence is not an obvious one. We think of error as "window to the mind" [15]. Understanding why and when humans err helps identifying the building blocks of intelligent human behavior. Of special interest are errors of trained users. Using software systems after having received some training is characterized by rule-based behavior [17]. Goals are reached by using stored rules and procedures that have been learned during training or earlier encounters with similar systems. While errors are not very frequent on this level of action control, they are also pervasive and cannot be eliminated through training [18].

Our focus on rule-based behavior allows a straightforward definition of error: Procedural error means that the (optimal) path to the current goal is violated by a non-optimal action. This can either be the addition of an unnecessary or even hindering action, which is called an *intrusion*. Or a necessary step can be left out, constituting an *omission*.

1.2 Memory for Goals

A promising theory of rule-based sequential action is the Memory for Goals (MFG) model [1]. The MFG proposes that subgoals, i.e., atomic steps towards a goal, are underlying memory effects, namely time-dependent *activation*, *interference*, and associative *priming*. Higher activation leads to faster recall and thereby shorter execution times. If the activation is too low, the retrieval of the subgoal may fail, resulting in an omission. Interference with other subgoals may

lead to intrusions. Priming is the most important concept in the context of this paper as it provides the link to the ontology in the background.

Our basic assumption is that subgoals receive priming from the general concepts that they represent. Hitting a button labeled "Search" is connected to the concept of search; choosing an option called "Landscape" in a printing dialog is related to the concept of landscape. If the general concept that is semantically linked to a subgoal is highly activated in the knowledge base, the respective subgoal should receive more priming, resulting in a higher overall activation of the subgoal. Taken together with the MFG, this results in three high-level predictions for subgoals, corresponding UI elements, and their respective concepts:

1. Execution time should decrease with concept activation.
2. Omission rate should decrease with concept activation.
3. Intrusion rate should increase with concept activation.

2 Experiment

The empirical basis for our model is provided by a usability study targeting a kitchen assistant from an ambient assisted living context. The kitchen assistant provides basic help during the preparation of meals by proposing recipes, calculating ingredients quantities, and by presenting interactive cooking instructions.

In order to assess the three ontology-based predictions stated above, we performed a reanalysis of previously published data [12]. We are concentrating on a single screen of the kitchen assistant that allows searching for recipes based on predefined attributes. A screenshot of the search attribute form translated to English is given in Fig. 1. The search attributes are grouped into nationality (French, German, Italian, Chinese) and type-of-dish (Main Course, Pastry, Dessert, Appetizer). We excluded three health-related search options as they were neither well represented in the experimental design, nor in the ontology. For the eight remaining buttons, we identified the best matching concept from the DBpedia ontology and use the number of links to it as measure of relevance of the concept. As can be seen in Table 2, the buttons in the nationality group are two to three magnitudes more relevant than the buttons in the type-of-dish group. Our empirical analysis therefore unfolds around the differences between those two groups.

2.1 Method

Twenty participants recruited on and off campus (15 women, 5 men, M_{age}=32.3, SD_{age}=11.9) took part in the experiment. Amongst other things, each participant completed 34 recipe search tasks using the attribute selection screen (see Fig. 1). One half of the tasks was done using a tablet computer, a large touch screen was used for the other half. Instructions were given verbally by the experimenter. All user actions were logged and videotaped for subsequent task execution time and error analysis.

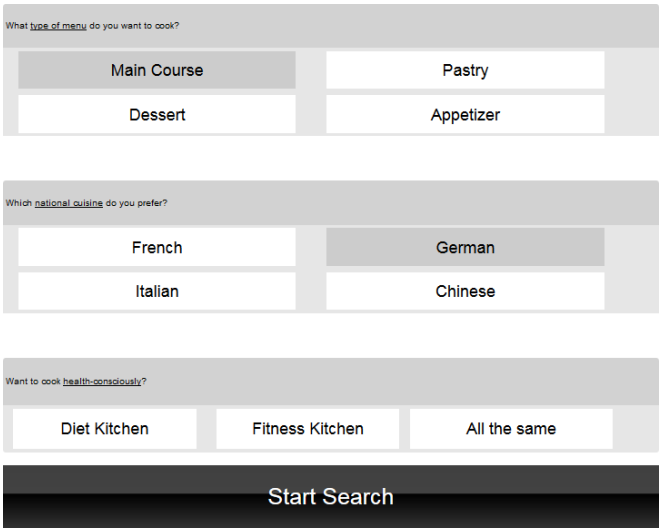

Fig. 1. Screenshot of the English version of the recipe search screen

2.2 Results

We observed a total of 1607 clicks on the eight search attribute buttons under investigation. The results for our three ontology-based predictions are as follows.

Execution Time We exluded all clicks with substantial wait time (due to task instruction or system response) from the analysis. The remaining 822 clicks still differ in the necessary accuracy of the finger movement which is strongly related to the time needed to perform the movement as formulated in Fitts' law [9]. Individual differences in motor performance were large, and the device used also had an effect on the click time. We therefore added subjects as random factor with device and Fitts-slope within subject to the analysis. The click time was analyzed using a linear mixed model [4], fixed effects were tested for significance using the Satterthwaite approximation for degrees of freedom. Results are given in Table 1. Besides the expected effects of Fitts' law and device, we observed a significant difference between the buttons for type-of-dish and nationality, with type-of-dish needing approximately 100 ms longer.

Omissions and Intrusions If those 100 ms are caused by lack of activation (as predicted by the MFG), then this lack of activation should cause more omissions for the type-of-dish group and more intrusions for the nationality group. We observed 14 intrusions and 19 omissions during the handling of the search attribute page (error rate 2.0%). Mixed logit models with subject as random factor showed no significant influence of the attribute group, but at least for omissions, the effect points into the expected direction (omissions: $z = 1.50, p = .133$; intrusions: $z = -.05, p = .964$). The omission rates for nationality and type-of-dish are 0.8% and 1.6%, respectively.

Table 1. Linear mixed model results for the click time analysis

Factor	Estimate	t	df	p
Fitts' Index of Difficulty in bit	173 $\frac{\text{ms}}{\text{bit}}$	4.95	22.4	$< .001$
Device (Tablet vs. Screen)	213 ms	4.38	24.3	$< .001$
Attr. Group (Dish vs. Nationality)	112 ms	2.47	611.3	.014

Discussion We investigated the difference between frequently vs. less frequently used concepts (nationality vs. type-of-dish) on a home assistance UI with regards to three dependent variables. The MFG predicts faster execution, less omission errors, and more intrusion errors for the higher used concept.

The empirical results are mixed. Buttons in the nationality group are clicked faster and weakly tend to be less prone to omissions. We did not find an intrusion effect, but this does not necessarily contradict the theory. The MFG explains intrusions by interference with earlier subgoals that are still present in memory. In the context of the experiment presented here, those intruding subgoals are memory clutter from already completed trials. In experimental design terms, this is called a carry-over effect. Due to the order of trials being randomized between subjects, intrusions should not happen on a general, but a subject-specific level.

3 Cognitive Model

The cognitive model presented here has been created using ACT-R 6 [2]. It has been shown to reproduce omission and intrusion errors for task-oriented vs. device-oriented UI elements well [12]. A comparison of the model's predictions for the different search attribute buttons has not been done before.

Following the MFG, the model creates and memorizes a chain of subgoal chunks when it receives task instructions through ACT-R's auditory system. It follows this chain of subgoals until either the goal is reached or memory gets weak. In case of retrieval failure, the model reverts to a knowledge-in-the-world strategy and randomly searches the UI for suitable elements. If it can retrieve a subgoal chunk that corresponds to the currently attended UI element, this subgoal is carried out and the cycle begins again.

The only declarative knowledge that is hard-coded into the model is that some UI elements need to be toggled, while others need to be pushed. The model interacts directly with the HTML interface of the kitchen assistant by the means of ACT-CV [11].[3]

3.1 Adding World Knowledge to the Model

In order to assess how cognitive modeling can benefit from ontologies, we took the barely knowledgeable model and added applicable pieces of information from

[3] See [12] for a more detailed description. The source code of the model is available for download at http://www.tu-berlin.de/?id=135088

Table 2. Semantic mapping between UI and ontology. Inlink count obtained from DBpedia 3.9 [14]. Subtitle-based word frequency (per 10^6 words) from [5]

Concept	UI label	DBpedia entry	Inlink count	per 10^6 links	Word freq.
German	Deutsch	Deutschland	113621	2474.3	10.2
Italian	Italienisch	Italien	56105	1221.8	6.2
Chinese	Chinesisch	China	10115	220.3	8.2
French	Französisch	Frankreich	79488	1731.0	17.4
Main Course	Hauptgericht	Hauptgericht	35	0.8	0.8
Appetizer	Vorspeise	Vorspeise	72	1.6	1.5
Dessert	Nachtisch	Dessert	193	4.2	6.5
Pastry	Backwaren	Gebäck	165	3.6	0.3

Wikipedia to its declarative memory. We propose semantic priming from long-living general concepts to the short-lived subgoal chunks that are created by the model when it pursues a goal.

How much priming can we expect, based on the information that is available within DBpedia? We are using the inlink count as measure of the relevance of a concept. In ACT-R, this needs to be translated into an activation value of the chunk that represents the concept (i.e., Wikipedia article). Temporal decay of activation is modeled in ACT-R using the power law of forgetting [2]. Salvucci [19] has applied this law to the concepts within DBpedia, assuming that they have been created long ago and the number of inlinks represents the number of presentations of the corresponding chunk. The base activation B can be determined from inlink count n as follows

$$B = \ln(2n) \tag{1}$$

While we agree with Salvucci's rationale, deriving the activation from raw inlink counts is a little too straightforward in our eyes. Numerically, it creates very high activation values. And as the total number of entries varies between the language variations of DBpedia, switching language (or ontology) would mean changing the general activation level.[4] In the special case of our model, the use of (1) caused erratic behavior because the high amount of ontology-based activation overrode all other activation processes (i.e., activation noise and mismatch penalties for partial matching of chunks). We therefore introduced a small factor c that scales the inlink count down to usable values. Together with ACT-R's minimum activation constant blc, this results in the following equation

$$B = \max(\ln(c \cdot n), blc) \tag{2}$$

How is the semantic priming to subgoal chunks finally achieved? The declarative memory module of ACT-R 6 only allows priming from buffers ("working

[4] The English DBpedia is 2.5 to 3 times larger than the German one. "Intelligence" has 1022 inlinks in the English DBpedia, but "Intelligenz" has only 445 inlinks in the German one.

memory") to declarative ("long term") memory. We therefore introduced a hook function that modifies the activation of every subgoal chunk whenever it enters long term memory according to the general concept that is related to the goal chunk.

3.2 Goodness of Fit

The model was run 300 times with concept priming disabled ($c = 0$), and 300 times with priming enabled ($c = .005$, resulting average base activation of the eight concepts $M_B = 2.4$, $SD_B = 3.4$). For both conditions, we computed time and error predictions for each button and compared these to the empirical observations. The effect of the needed accuracy of the finger move was eliminated based on Fitts' law, using a linear mixed model with subject as random factor [4] for the empirical data and a linear regression for the model data, as the Fitts' parameters were not varied during the simulation runs. Correlations between the respective residuals are given in Table 3. Omission and intrusion rates per button were correlated without further preprocessing.

The results are given alongside R^2 and RMSE in Table 3. While the goodness-of-fit with R^2 constantly below .5 and substantial RMSE is not overwhelming, the difference to the baseline is worth discussion. The model without concept priming displays no or negative correlations between its predictions and the empirical values, meaning that the baseline model is even worse than chance. The corresponding regression lines are displayed on the upper part of Fig. 2. When concept priming is added, all three dependent variables show substantial positive correlations between observed and predicted values. The difference between the correlations is very large, i.e., always above .75.

The positive correlation for intrusions is noteworthy as we could not establish an empirical relationship between concept relevance and the observerd intrusion rates in the first place (see above). If our hypothesis of intrusions being caused by leftovers from previous trials with additional priming from ontology-based concepts holds, then this result underlines the benefits of adding ontologies to cognitive architectures. A closer look at Fig. 2 reveals that the correlation for intrusions is highly dependent of two outliers, the results should therefore be interpreted with care.

Table 3. Correlations between the empirical data and the model predictions.

Dependent Variable	r_{baseline}	r_{priming}	Δr	R^2_{priming}	$\text{RMSE}_{\text{prim.}}$
Execution time (residual)	-.218	.684	.758	.468	78 ms
Omission rate	-.390	.640	.824	.410	.027
Intrusion rate	-.654	.511	.873	.261	.011

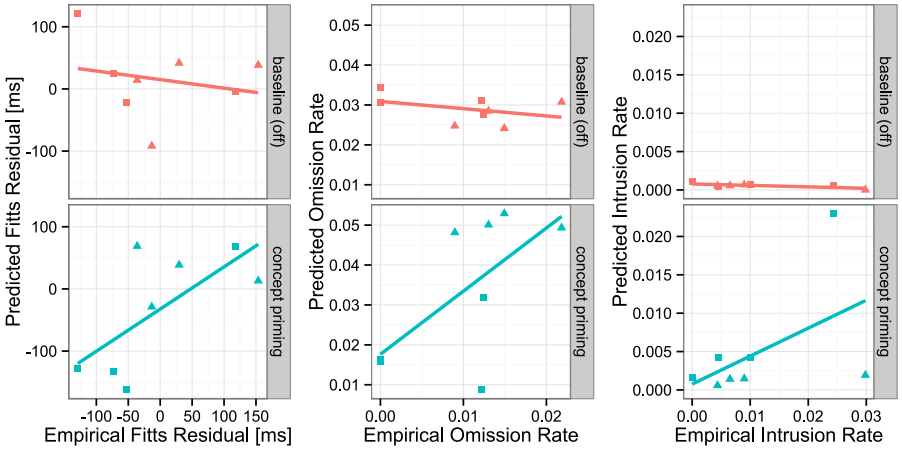

Fig. 2. Click time residuals after Fitts' law regression, intrusion and omission rates of the cognitive model with and without priming from the DBpedia concepts. Negative slopes of the regression line mean worse than chance predictions. Positive slopes mean better than chance predictions. Squares denote buttons of group "nationality", triangles denote "type of dish".

4 Discussion and Conclusions

We presented a cognitive model of sequential action that has been developed for the prediction of human error during the use of a home assistance system [12]. The original model did not have any world knowledge and accordingly was unable to reproduce effects of concept relevance on task execution time and omission rate that we found in a reanalysis of our empirical data. Adding concepts from DBpedia [14] to the declarative knowledge of the model and modulating the activation of these concepts based on the number of links inside DBpedia that point to them allowed not only to reproduce the time and omission rate differences, but to some extent also the rates of intrusions. While the prediction of execution time and omissions mainly lies within the ontology, intrusions can only be explained by the combination of cognitive model and ontology, highlighting the synergy between both.

To our knowledge, this is the first time that Salvucci's approach for adding world knowledge to a cognitive architecture [19] is empirically validated. The practical development of the model showed that the activation equation proposed by Salvucci, while being theoretically sound, creates hurdles for the combination of world knowledge with existing cognitive models. Therefore, we introduced a constant scaling factor to the ontology-based activation computation. This goes in line with the common practice in psycholinguistics to use standardized values that are independent of the corpus in use. The factor chosen here helped to keep the influence of the ontology on subgoal activation at par with the other activation sources applied (i.e., activation noise and partial matching).

It is also informative to compare our approach to research on information foraging, namely SNIF-ACT [10]. This system uses activation values that are estimated from word frequencies in online text corpora, which would lead to general hypotheses similar to the ones given above. But beyond this, a closer look unveils interesting differences to the DBpedia approach. While word frequency and inlink count are highly correlated (r=.73 in our case, see Table 2), the word frequency operationalization yields much smaller differences between the nationality vs. type-of-dish groups. Frequency based-approaches also need to remove highly frequent, but otherwise irrelevant words beforehand (e.g., "the", "and"). In Wikipedia, this relevance filter is already built into the system and no such kind of preprocessing is necessary. Empirically, we obtained inconclusive results when using word frequency in a large subtitle corpus [5] instead of Wikipedia inlink count as concept activation estimate.

While the combination of cognitive model and ontology provides some stimulating results, it also has some downsides and limitations. First of all, the small number of observed errors leads to much uncertainty regarding the computed intrusion and omission rates. Especially in case of intrusions, the empirical basis is rather weak. The goodness-of-fit is highly dependent on two outliers. While one of these matches the high-level predictions given in the introduction ("German" being more prone to intrusions), the other one points towards a conceptual weakness of the model ("Pastry" showing many intrusions in the empirical data although having just a few inlinks). The "Pastry" intrusions happened during trials with the target recipes baked apples ("Bratäpfel") and baked bananas ("Gebackene Bananen"). One could speculate that those recipes have primed the type-of-dish attribute that is linked to baking. This kind of semantic priming is currently not covered by our system. We are planning to integrate more sophisticated models of long-term memory [20] to allow dynamic priming between concepts as well.

Besides the conceptual findings, our ontology-backed cognitive model also provides benefits to applied domains. With its ability to interact with arbitrary HTML applications, the model could be used for automatic usability evaluation of user interfaces. Its ability to predict omissions and intrusions could be used to spot badly labeled UI elements during early development stages.

Acknowledgements We gratefully acknowledge financial support from the German Research Foundation (DFG) for the project "Automatische Usability-Evaluierung modellbasierter Interaktionssysteme für Ambient Assisted Living" (AL-561/13-1).

References

1. Altmann, E.M., Trafton, J.G.: Memory for goals: An activation-based model. Cognitive science 26(1), 39–83 (2002)
2. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological review 111(4), 1036–1060 (2004)

3. Anderson, J.R., Lebiere, C.: The Newell test for a theory of cognition. Behavioral and Brain Sciences 26(05), 587–601 (2003)
4. Bates, D., Maechler, M., Bolker, B., Walker, S.: lme4: Linear mixed-effects models using Eigen and S4 (2013), r package version 1.0-5
5. Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., Böhl, A.: The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology 58(5), 412 (2011)
6. Douglass, S., Ball, J., Rodgers, S.: Large declarative memories in ACT-R. Tech. rep., Manchester, UK (2009)
7. Emond, B.: WN-LEXICAL: An ACT-R module built from the WordNet lexical database. In: Proceedings of the Seventh International Conference on Cognitive Modeling (2006)
8. Fellbaum, C.: Wordnet. In: Poli, R., Healy, M., Kameas, A. (eds.) Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer, Dordrecht (2010)
9. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47(6), 381–391 (1954)
10. Fu, W.T., Pirolli, P.: SNIF-ACT: A cognitive model of user navigation on the world wide web. Human-Computer Interaction 22, 355–412 (2007)
11. Halbrügge, M.: ACT-CV: Bridging the gap between cognitive models and the outer world. In: Brandenburg, E., Doria, L., Gross, A., Günzlera, T., Smieszek, H. (eds.) Grundlagen und Anwendungen der Mensch-Maschine-Interaktion. pp. 205–210. Universitätsverlag der TU Berlin, Berlin (2013)
12. Halbrügge, M., Quade, M., Engelbrecht, K.P.: A predictive model of human error based on user interface development models and a cognitive architecture. In: Taatgen, N.A., van Vugt, M.K., Borst, J.P., Mehlhorn, K. (eds.) Proceedings of the 13th International Conference on Cognitive Modeling. pp. 238–243. University of Groningen, Groningen, the Netherlands (2015)
13. Laird, J.: The Soar cognitive architecture. MIT Press, Cambridge, MA (2012)
14. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2014)
15. Norman, D.A.: Slips of the mind and an outline for a theory of action. Tech. rep., Center for Human Information Processing, San Diego, CA (1979)
16. Oltramari, A., Lebiere, C.: Extending cognitive architectures with semantic resources. In: Artificial General Intelligence, pp. 222–231. Springer (2011)
17. Rasmussen, J.: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. Systems, Man and Cybernetics, IEEE Transactions on 13, 257–266 (1983)
18. Reason, J.: Human Error. Cambridge University Press, New York, NY (1990)
19. Salvucci, D.D.: Endowing a cognitive architecture with world knowledge. In: Bello, P., Guarini, M., McShane, M., Scassellati, B. (eds.) Proc. CogSci 2014, pp. 1353–1358 (2014)
20. Schultheis, H., Barkowsky, T., Bertel, S.: LTM C – an improved long-term memory for cognitive architectures. In: Proceedings of the Seventh International Conference on Cognitive Modeling. pp. 274–279 (2006)