An Architecture Approach to Modeling the Remote Associates Test

Jule Schatz (schatzju@umich.edu)
Steven J. Jones (scijones@umich.edu)
John E. Laird (laird@umich.edu)
University of Michigan, 2260 Hayward Street
Ann Arbor, MI 48109-2121 USA

Abstract

The remote associates test (RAT) depends heavily on memory retrieval and is difficult for humans. A previous model of difficulty on this task accounted for difficulty with a measure incorporating fan and association strength. This paper investigates how the choice of knowledge base and agent strategy impact difficulty on the task while providing a more comprehensive account for human difficulty on this task in terms of cognitive architecture components. The models we created, using the cognitive architecture Soar, vary by using two distinct methods of retrieval from semantic memory. The knowledge bases used in our models vary in that one uses only collocations and compound words to form word associations while the other is from a crowd-sourced dataset with unrestricted types of word association. The model which best matches human difficulty relies on spreading activation to drive retrieval and uses the crowd-sourced dataset for its knowledge base.

Keywords: Semantic Memory; Remote Associates Test; Soar; Association-based Retrieval.

Introduction

This paper investigates computational models for the remote associates test (RAT) (Mednick, 1962). A single RAT problem consists of presenting three words and then asking the test taker to respond with a forth word that is associated with the three given words. For example, if "Swiss," "cake," and "cottage" are the given words, then "cheese" would be the correct response. Bowden and Jung-Beeman developed 144 RAT problems and tested human performance on those problems. To minimize any variance from confounding factors, they used only compound word or phrase associations. For example, the association between "deep" and "sleep" is valid because those two words are often found next to each other. The association between "deep" and "complex" is not valid because "deep" and "complex" do not form a compound word or common phrase even though they are associated through similar meanings. The problems use common words, to avoid vocabulary difficulties. To avoid priming effects, solution words are never repeated or used as problem words. Three additional example problems are shown in Table 1. The human study included four time limits: 2 seconds, 7 seconds, 15 seconds, and 30 seconds. The paper provides the mean time to solution, the standard deviation for time to solution, and the percent of participants that correctly answered each question in each time limit.

Word 1	Word 2	Word 3	Answer
man	glue	star	super
dew	comb	bee	honey
rain	test	stomach	acid

Table 1: Example RAT items. The answer is associated with the words through collocation or as a compound word.

As opposed to generally characterizing word association memory (Griffiths & Steyvers, 2002), Olteţeanu and Falomir (2015) created a model intended to provide an account for human performance on those 144 RAT problems. Their model, comRAT-C, was created within the CreaCogs architecture (Olteţeanu, 2014) and was inspired by their account of creative problem solving which posits two extremes of behavior: "creative search" and "productive representation construction processes". The "creative search" extreme is embodied in comRAT-C, which uses associational links to search a knowledge base for a representation that affords a solution to a problem. Their knowledge base, called RAT-KB, builds off of the most frequent 2-grams from the Corpus of Contemporary American English (COCA) (Davies, 2008). Associational links in RAT-KB are bidirectional.

In their analysis (Olteţeanu & Schultheis, 2017), they state that the difficulty of this task depends on "(i) the frequency of a query-answer association, as a form of associative strength and (ii) the ratio between such an associative strength and the number of answer associations." We interpret these factors as being analogous to how association strength (Anderson & Pirolli, 1984) and fan (Anderson, 1974) govern retrieval difficulty. Their results were not in terms of providing a model with matching timing and correctness. Instead, they show the correlation between their difficulty estimates for RAT items and human data, i.e., both human timing data and human correctness data on the RAT items. We interpret this as characterizing the relative difficulty for humans by the ordering of human solution time and human correctness.

We use their work as inspiration for our research but deviate in hopes of providing a more comprehensive analysis. First, we note that RAT-KB includes only associational links for collocations and compound words, and that all links are bidirectional. Humans know many words and associations beyond this constraint, and possibly do not have bidirectional links between these words. Thus, our first step is to use a larger, more comprehensive knowledge base, where links are not necessarily bidirectional. We then determine how such a knowledge base influences task performance, and more specifically, relative problem difficulty. Second, we wish to determine whether existing architectural long-term declarative memory retrieval theories, as developed in ACT-R (Anderson, 2009), are sufficient to accurately model RAT problem difficulty. In these declarative memory models, retrieval is determined by base-level activation, association strength

(Anderson & Pirolli, 1984), and fan (Anderson, 1974). To explore these questions, we develop models in Soar (Laird, 2012) whose long-term declarative memory retrieval mechanisms mimic those in ACT-R (Jones, Wandzel, & Laird, 2016). Third, there does not exist a simple, deliberate model of retrieval that does not primarily rely on association strength or fan that can be used as a baseline for comparison with the association-based models. Thus, our third step is to develop such a model in Soar which uses queries that are not influenced by associate strength or fan except in the case of ties.

In the remainder of the paper, we proceed through these steps, one by one. First, we introduce a new crowd-sourced knowledge base. Second, we describe the two models we developed in Soar. Third, we evaluate these models on the new knowledge base as well as on a replica of the original RAT-KB knowledge base, focusing on how well these models (and knowledge bases) model human difficulty. The primary result is that the more comprehensive knowledge base combined with associational retrieval allow us to model human difficulty with high correlation ($R^2 = 0.89$).

Knowledge Bases

To allow us to compare our new knowledge base to prior work, we reconstructed RAT-KB, by creating a knowledge base called COCA-TG based on the steps described in the original paper. The final number of words and associations for COCA-TG are shown in Table 2. RAT-KB includes bidirectional associations for all words, as does COCA-TG.

Because COCA-TG leaves out other types of associations that can indirectly influence retrieval (e.g. through competition and the fan effect), we created a larger knowledge base using the Human Brain Cloud (HBC) database. HBC was crowd-sourced through an online game of word associations (Gabler, 2013), where the player is presented with a word and asked to type in any other word that they believe to be closely related to the given word (if the given word is "bird" the player might type "feather" or "fly" or "nest"). The website records the human responses, and creates a dataset that consists of triples in the form of "word1," "word2," and weight, where weight is the number of times "word1" was associated with "word2". HBC only includes links entered by a player, so not all word pairs have bidirectional links (as in COCA-TG and RAT-KB). As shown in Table 2, HBC contains close to twice as many words, and over three times as many associations as contained in COCA-TG.

	HBC	**COCA-TG**
words	40,652	20,809
associations	1,298,831	349,196

Table 2: For each knowledge base, the number of unique words, and associations between words. Bidirectional associations count as two associations.

Models

Our models are developed in Soar, which features a long-term semantic memory that can be queried to retrieve information into working memory (Derbinsky & Laird, 2010). To run a model, semantic memory is initialized with the contents of a knowledge base, where nodes in the memory consist of words and the links are associations. The weights of the associations are those in the knowledge bases.

Retrieval in Soar returns the most highly-activated element which satisfies the provided cue. The activation of an item is the sum of base-level activation and spreading activation. Base-level activation represents the frequency and recency of prior retrievals, but for these models we had no prior values. Instead, we initialized all words in the knowledge base with a single base-level activation. However, we assume that words should have some usage history and we return to this issue in the discussion.

To solve a RAT problem, a model uses the three presented words to find the associated answer. There are potentially many strategies for doing this; however, we focused on two strategies that are directly supported in Soar. In Soar, an agent can retrieve information from semantic memory, either by providing a specific cue that is matched against elements in long-term memory (Cued Retrieval model), or an agent can use a general cue and leverage spreading activation to retrieve words based on context as defined by the contents of working memory (Free Recall model). The Cued Retrieval model uses queries that include the original words, whereas the Free Recall model does not include the original words and relies on spreading activation, which incorporates both association strength and fan.

Cued Retrieval

As a baseline, we created the Cued Retrieval model, with the goal of correctly answering as many RAT items as possible given the knowledge available in long-term memory, while keeping agent design simple and in accordance with architectural constraints. This model first retrieves all three given words into working memory. It then creates a cue that specifies semantic memory should only return a word that has an outgoing link to all three of the given words. Semantic memory then returns either a word that matches the cue (that is associated to all given words), or it reports a failure if no such word exists. If semantic memory has multiple possible solutions, spreading activation acts as a tie breaker. If the initial query failed, the model changes the cue to only require semantic memory to return a word that is associated to two of the given words. The model will try all combinations of two words, and report an answer for the first one it finds. If all of those fail, it tries each given word individually and reports that answer. Because the model deliberately queries semantic memory for a word with all associations first, it will find a correct solution if one exists in the database, which is not guaranteed in the Free Recall model.

Free Recall

The Free Recall model incorporates association strength and fan via spreading activation (Jones et al., 2016). The agent first retrieves the three given words into its working memory from semantic memory. Having these words in working memory causes activation to spread to words linked to those words in semantic memory. Each given word acts as a source for an equal amount of activation, which is then divided proportionally among the outgoing links based on association strength. Association strengths from a given source are normalized to sum to one. Activation decays with the distance of spread, but in this model, for simplicity, spreading is limited to a depth of 1.

Consider an example where there is a source word s and recipient word r with a pre-normalized association strength from s to r of $a_{s\to r}$. Assume a set, R, of all recipients. The contribution of spread from the source to the recipient in this case is $\frac{a_{s\to r}}{\sum_{k\in R} a_{s\to k}}$. Therefore, an item with a stronger association from the source word will get more activation than one with a weaker association. In addition, the more links or fan the source has, the less activation will spread to its recipients.

To retrieve a word, the model initiates a retrieval from semantic memory with the only constraint being that the word is not one of the three given words. Semantic memory then returns the word with the highest activation. A high activation is no guarantee that the retrieved word is associated with all three words because words can be retrieved that have strong associations to only one or two of the original words, especially if they have low fan. Once a word is retrieved from semantic memory, the model tests how many of the three initial words relate to it by testing if there are links between it and those initial words. If it is related to all three words, the model uses the word as its solution. If the retrieved word is related to two or fewer of the given words, the model queries again and retrieves a new word from semantic memory, inhibiting any it has previously retrieved. The number of attempts it will make is a parameter, which we vary in the evaluation. If the model runs out of attempts, it chooses one of the retrieved words that has the most relations with the given words.

This model incorporates the findings from Olteţeanu and Schultheis's research in terms of the two factors (association strength and fan) that influence the difficulty of RAT items for humans. Their findings indicated that those factors influence whether humans can solve a RAT problem and the time it takes for them to find a solution. Words that have stronger associations are more likely to be retrieved by this model, as well as words from low fan sources.

Evaluation

We tested both the Cued Retrieval and the Free Recall models using both the HBC and the COCA-TG databases, giving four model configurations. Our results compare the models' timing and correctness to human timing and human correctness on the task, focusing on correctness. ComRAT-C provided a probability value that they consider an estimate of the probability that a word is an answer. Their results were that for a RAT item with a given correct answer, comRAT-C's estimate of the probability that the correct answer was correct correlates positively with the number of humans who answered correctly and correlated negatively with human mean time to solution. However, Soar models retrieval as competitive, so only a single element is selected. The significance of this is that the activation of the correct answer does not completely determine if it will be the model's answer. An activation can be high, but if it is not the *highest* with respect to words that compete for retrieval, then it will not be retrieved. For this reason, our results are not directly comparable with those of comRAT-C. We instead adopt an approach where we compare the correctness of the answers produced by our models to the correctness of the human answers.

Overall Difficulty

First, we consider overall task difficulty as it relates to modeling difficulty on the RAT. Figure 1 shows two diagrams, one for each knowledge base. These diagrams include the number of RAT items that were answered correctly for different model configurations, as well as the average number of correct responses made by humans. These averages are for when humans have only 15 seconds and 7 seconds to generate an answer, and as is obvious, this is a difficult task for humans. The x-axis is the number of attempts (1-20) for the Free Recall model. The models for humans (light and dark green) and Cued Retrieval (dots) have only a single attempt. We show them as straight lines for ease of comparison with the Free Recall model. We also include the number of items where all the given words and the answer exist in the database.

The top figure shows results from using COCA-TG for both models. The Cued Retrieval model with COCA-TG gets 65 RAT items correct. The Free Recall model initially improves as more attempts are made, and achieves better performance than the Cued Retrieval model from 3 attempts on. The best it achieves is 78 correct at ten, eleven, and fourteen attempts. This improvement is possible because this model makes guesses for problems where it cannot find an exact answer, and sometimes those guesses are correct. As the figure shows, the COCA-TG models outperform humans except in the case where the Free Recall model makes a single attempt.

The bottom figure shows results using HBC. HBC contains more correct answers than COCA-TG (105 vs. 55), invariably because of its larger size (see Table 2). Once again, through guessing, the Free Recall model achieves performance better than one might expect. With HBC, Free Recall achieves the 7 seconds human performance when it uses two attempts and the 15 seconds human performance with three attempts. We hypothesize that more attempts are required to achieve the same performance in the HBC database, because, on average, HBC words have higher fan (32 vs. 17).

Relative Difficulty

Next, we compare the results to human data provided by Bowden and Jung-Beeman. We are interested in which

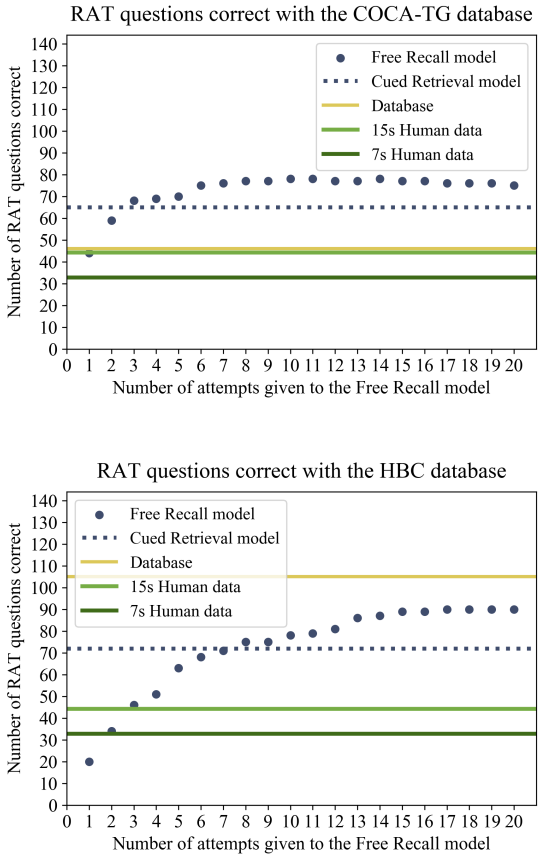

Figure 1: The number of RAT items each model with each database got correct out of the 144 possible items. Note the Free Recall model results are shown as 20 separate points.

configuration best matches human behavior. While our models generally perform better on the RAT than humans, we can separately characterize behavior by relative difficulty.

Timing Comparison In measuring timing, our goal is to see if our models are in the ballpark of human response times. A Soar model's timing can be roughly compared to human times. For Soar models, we consider a single decision cycle as corresponding to 50ms of human behavior. However, we consider retrieval as requiring roughly 300ms. Using those parameters, the Free Recall model using HBC and given 2 and 3 attempts took an average time of 2.02 and 2.20 seconds respectively to find a solution and the Cued Retrieval model took an average time of 2.27 seconds. The subjects in the study took on average 4.87 seconds when given 7 seconds to solve the problem and 7.26 seconds when given 15 seconds. Thus, the time taken for our models to solve a RAT item is similar in magnitude to how long it took the subjects. This rough similarity in timing suggests that our models are using approximately the same number of steps and retrievals as is found in human behavior.

Correctness Comparison In this section, we evaluate whether the RAT questions that are difficult for humans are also difficult for the models. For human difficulty, we focus on correctness and we use the percentage of people who got the correct answer as the metric of difficulty. For our models, we use whether the model produces the correct answer for a given item.

In order to compare these two metrics, we binned the 144 RAT items into 12 groups of 12 based on correctness in humans. The first group being the most difficult for humans (the lowest percentage of people got them correct), the last being the easiest (the highest percentage of people got them correct). We did this for both the 7 seconds and 15 seconds human results, as they had times closest to those predicted by our model. From the 12 questions in each bin, we calculated the mean percentage of people who got the questions correct. We then compare the average RAT items correct for humans to the number of items our models got correct for each 12 question bin. We did this comparison for the Cued Retrieval and the 1-20 guesses Free Recall models for each knowledge base. The correlations between the number of items correctly answered by humans and the number of items correctly answered by our models are shown in Figure 2.

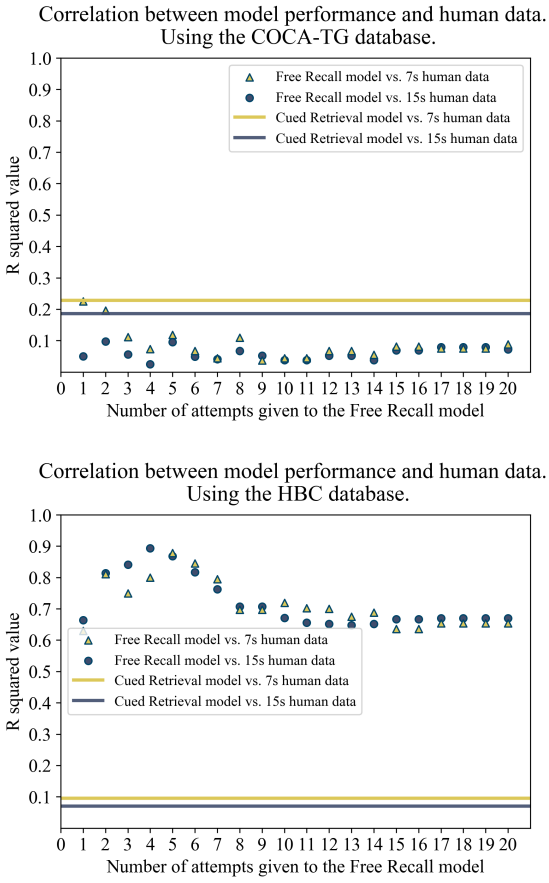

Figure 2: The correlations of model difficulty with human difficulty is displayed. The Free Recall model has a varying number of attempts displayed on the x axis.

The models using COCA-TG are shown in the top diagram and they have low correlations with human difficulty. Using COCA-TG, the highest correlated model is Free Recall with 1 attempt: 0.23. The models using HBC are shown in the bottom diagram, and all Free Recall model correlations are better using HBC. The highest correlated model is Free Recall with 4 attempts for the 15 second human data: 0.89. The Cued Retrieval model has low correlation for both databases. This suggests that HBC is a better model for the knowledge humans use to perform this task, and that the Free Recall model with 4 attempts is an excellent model of human difficulty.

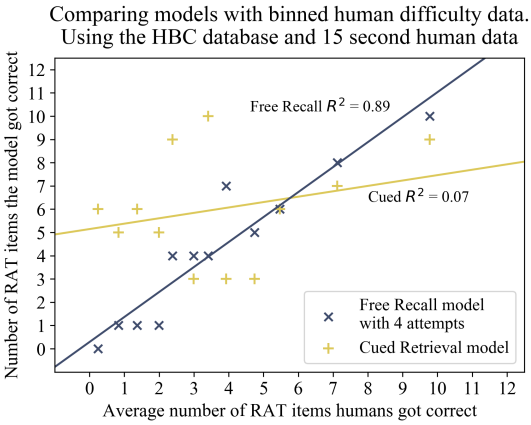

Figure 3: Agent performance is displayed with respect to human performance. The human data refers to the average number correct within a difficulty bin, for 12 bins. The best fit line is shown for both the Free Recall and the Cued Retrieval model data.

In Figure 3 we further investigate the most highly-correlating model configuration (Free Recall, 4 attempts) with a comparison to the Cued Retrieval model (both using HBC). This figure shows the number of RAT items each model got correct for each of the 12 groupings of items, ordered by difficulty, correlated with the expected number that humans got correct for each of those groups. The best fitting line has a slope of 1.074 and an y-intercept of 0.289 making it a close one-to-one relationship between the human's relative difficulty and the model's relative difficulty. To further verify our claim that the Free Recall model relates better to human data than the Cued Retrieval model, we ran a logistic regression test with the null hypothesis that the model's correct versus incorrect output for RAT items does not relate to the percentage of humans that got the RAT items correct. For the Free Recall model given 4 attempts, we reject the null hypothesis with a p-value of 2.86e-07. For the Cued Retrieval model we do not reject the null hypothesis, given a p-value of 0.184.

Fan and Association Strength Influence on Relative Difficulty Given a model and knowledge base which correlate with human relative difficulty, we next attempt to characterize the effects of association strength and fan on model difficulty. We selectively lesion the effects of fan and associ-

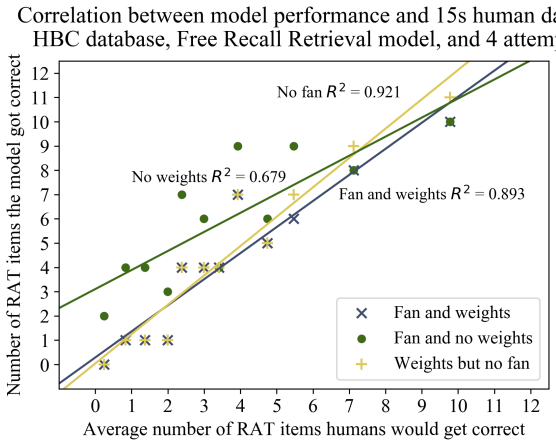

Figure 4: For our model configuration with the highest correlation to human relative difficulty, we also display models corresponding to the removal of association strength (weights) and fan.

ation strength on retrieval to show how the correlation with human difficulty changes as a result. Lesioning of fan leads to a model of spread where only association strength governs spread, and where there is no normalization with respect the number of outgoing links from the source. Lesioning of association strength leads to a model of spread where all association strengths from a given node are equal. These additional models are plotted with lines of best fit alongside our best matching model in Figure 4, where we again present model correctness compared to human correctness.

We expect that lacking association strength and only using fan should give better results in terms of absolute number correct because a single strong association can dominate during retrieval, whereas with equal strength for all associations, only those items that have associations with all given words will be retrieved. We confirm that the lesioned HBC with 4 attempts gets 72 RAT items correct versus the original HBC with 4 attempts which gets only 51 correct. Additionally, we expect that lesioning either fan or association weights should led to worse match to human difficulty. This is the case for removing weights, with the correlation dropping to 0.679, but removing fan improves the correlation to 0.921, suggesting a mismatch between the associations in our database and those in humans.

Conclusion

We created models that perform the remote associates test by employing two distinct methods. While a previous model for difficulty on this task did find association strength and fan to govern retrieval difficulty, our work provides a better account of how such influences impact difficulty by using a more realistic knowledge base and implementing our models as agents that complete the task, getting answers which can be directly compared to human answers. The Cued Retrieval model does a cued query to semantic memory to find the solution, if it exists. If a solution was not found, the model makes a plau-

sible guess. The Free Recall model iteratively uses spreading activation to retrieve a potential solution until it finds a solution or until it hits a threshold. The semantic memory knowledge bases only contained word associations. This is limited in comparison to a human's semantic memory. However, the use of spreading activation and the associations found in HBC's memory network give results surprisingly consistent in terms of relative difficulty for answering RAT problems with human performance. While we replicated the RAT-KB knowledge base associated with the previous work's model with our COCA-TG knowledge base, we found that despite it only consisting of the relevant type of associations for the 144 RAT problems, it performed worse than the HBC knowledge base in terms of modeling human difficulty. Our hypothesis is that a combination of inclusion of bidirectional links in COCA-TG leads the model astray by allowing it to find associations that are either missing or have very low association strength in humans.

From the results, we find that the Free Recall model with 4 attempts is an excellent match to relative problem difficulty in human behavior for when humans have 15 seconds for the task, although it is also highly correlated for a range of number of attempts. While the Cued Retrieval model can retrieve more answers (depending on the choice of attempt parameter for the Free Recall model), the Free Recall model using the HBC database has higher correlation with human results than the Cued Retrieval model with the same knowledge. This is seen with both the 7 and 15 seconds binned human data.

In attempting to characterize the role of fan and association strength in these results, we found that a better match to human difficulty is achieved when association strength but not fan influences spreading activation. One possible explanation is that there are artifacts in the HBC knowledge base in terms of missing items and their connectivity that do not reflect human semantic memory. We already know that they do not contain all the relevant knowledge for the RAT questions. Thus, we plan to expand the HBC database by adding other databases that include more of the relevant words and associations, while still being representative of human word associations, such as the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998).

Another shortcoming of our databases is that they have no information about the recency and frequency of the words they include, and thus there is no contribution of base-level activation to our model (Anderson, Bothell, Lebiere, & Matessa, 1998). A reasonable extension would be to initialize our databases with usage information derived from other databases, such as COCA (Davies, 2008).

Acknowledgments

The work described here was supported in part by the Office of Naval Research under Grant Number N00014-18-1-2010. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the ONR or the U.S. Government.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

Anderson, J. R., & Pirolli, P. L. (1984). Spread of activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 791.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 634–639.

Davies, M. (2008). *The corpus of contemporary american english*. BYE, Brigham Young University.

Derbinsky, N., & Laird, J. E. (2010). Extending soar with dissociated symbolic memories. In *Symposium on human memory for artificial agents, aisb* (pp. 31–37).

Gabler, K. (2013). *Human brain cloud.* Retrieved from `https://humanbraincloud.com/`

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 24).

Jones, S. J., Wandzel, A. R., & Laird, J. E. (2016). Efficient computation of spreading activation using lazy evaluation. In *Proceedings of the 14th international conference on cognitive modeling.*

Laird, J. E. (2012). *The Soar cognitive architecture*. MIT Press.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220-232.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of south florida word association, rhyme, and word fragment norms.* `http://www.usf.edu/FreeAssociation/`.

Olteţeanu, A.-M. (2014). Two general classes in creative problem-solving? an account based on the cognitive processes involved in the problem structure-representation structure relationship. In *Proceedings of the international conference on computational creativity. publications of the institute of cognitive science* (Vol. 1).

Olteţeanu, A.-M., & Falomir, Z. (2015). Comrat-c - a computational compound remote associates test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81–90.

Olteţeanu, A.-M., & Schultheis, H. (2017). What determines creative association? revealing two factors which separately influence the creative process when solving the remote associates test. *The Journal of Creative Behavior*.