Towards Modeling False Memory with Computational Knowledge Bases

Justin Li (justinnhli@oxy.edu), Emma Kohanyi (emmajk2012@gmail.com)

Occidental College

Towards Modeling False Memory with Computational Knowledge Bases

One challenge to creating realistic cognitive models of memory is the inability to account for the vast common sense knowledge of human participants. Large computational knowledge bases such as WordNet and DBpedia may offer a solution to this problem, but may pose other challenges. This paper explores some of these difficulties through a semantic network spreading activation model of the Deese-Roediger-McDermott false memory task. In three experiments, we show that these knowledge bases only capture a subset of human associations, while irrelevant information introduces noise and makes efficient modeling difficult. We conclude that the contents of these knowledge bases must be augmented and, more importantly, that the algorithms must be refined and optimized, before large knowledge bases can be widely used for cognitive modeling.

Keywords: False Memory; Spreading Activation; WordNet; DBpedia; Knowledge Base; Cognitive Architecture.

Introduction

The modeling of human memory phenomena has a long history, from equations describing the strength of individual memory elements over time, to the embedded memory subsystems in modern cognitive architectures. One limitation of memory models, however, is their failure to account for how experimental subjects do not come into the laboratory as a blank slate, but with a large set of common-sense knowledge and facts about the world, as well as associations built up from individual experience. This background knowledge is impossible to fully elicit from subjects and often omitted from computational models. As a result, these models are over-simplified and may fail to account for phenomena in which the contents of memory play a role.

At the same time, the increasing number of artificially intelligent agents that operate in knowledge-rich environments has led to the development of large computational knowledge bases. Knowledge bases such as WordNet (Miller, 1995) and DBpedia (Bizer et

al., 2009) endow artificial agents with lexical and conceptual knowledge, allowing them to perform human-like reasoning. These collections of semantic knowledge, in a form that can be incorporated into the long-term memory of cognitive architectures, present an opportunity to build models that match real human memory in scope and scale. Recent work has adapted DBpedia for factual question-answering in the ACT-R architecture (Salvucci, 2015), a task for which the knowledge base is well suited, as it mirrors the use of DBpedia in artificial intelligence research. Whether knowledge bases can be used to model cognitive phenomena outside of reasoning and inference, however, remains an open question.

In this paper, we explore some of the challenges that researchers may face when incorporating large computational knowledge bases into a cognitive model. Specifically, we use WordNet and DBpedia to model the formation of false memories through human associations in the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995). We selected the false memory task specifically because it involves a broad range of knowledge that large knowledge bases could provide, while requiring associations for which WordNet and DBpedia may not be particularly well suited. The partial success of our model suggests that while large knowledge bases hold promise for general cognitive modeling, they present representational and algorithmic challenges that have yet to be overcome.

Background

The DRM task is a well-known procedure for inducing false memory in humans. First proposed in Deese (1959), it did not receive widespread attention until replicated and extended by Roediger and McDermott (1995). In the paradigm, participants are told they are part of a memory experiment and presented with a list of fifteen *stimuli words* at a moderate pace. After the presentation, participants are occupied with a filler task, before being given two minutes to recall as many words from the list as possible. Crucially, the list of words are not random, but are all associated with a *lure*, which itself does not appear on

the list. For example, for the lure *"needle"*, the list of words presented to the participants includes *"pin"*, *"sharp"*, *"prick"*, *"haystack"*, *"thorn"*, *"cloth"*. (For the remainder of this paper, stimuli words from a DRM list will be in quotes and italicized; lure words will be in quotes, italicized, and underlined; and all other words will be in quotes but unitalicized.) The result is that experiment participants will recall the lure at roughly the same rate as the stimuli words, and will further report that the lure was presented – a false memory. After a break, another list built around a different lure is presented, for 36 published false memory word lists (Stadler, Roediger, & McDermott, 1999). In the original study, participants recalled 62% of the stimuli words, and falsely recalled the lure 55% of the time.

In a different publication (Roediger, McDermott, & Robinson, 1998), the authors suggested that this phenomenon could be explained through a *spreading activation* mechanism. They hypothesized that the semantic concepts represented by the stimuli words are connected in a *semantic network*; nodes in the network represent concepts, while edges between nodes represent an association of some kind. Thus, every word on a DRM list would be connected to the lure, possibly with additional connections between stimuli words. Each word would also have an *activation value* that represents its salience at any particular time; the higher the activation, the more likely that concept will be recalled at that time. When a stimulus word is presented, it is hypothesized that not only is the activation of that concept *boosted*, but so is the activation of associated concepts, including the activation of the lure. The presentation of multiple stimuli words would boost the activation of the lure multiple times, causing its activation at the end of the presentation phase to be indistinguishable from the activation values of the stimuli words. Then, during the recall phase, words with the highest activation are recalled. Since participants could not determine whether the high activation of a word is due to its presentation or due to spreading activation (a *source monitoring failure*), they report the lure as having been presented.

Although spreading activation is an intuitive and appealing explanation for how false memories are induced in the DRM paradigm, creating a cognitive model of the task requires

capturing human associations between words. The breadth of the stimuli and lure words – which range from everyday objects such as *"window"* and *"pen"* to relatively obscure words such as *"sash"* (a type of window) and *"Cross"* (a pen company) – makes the creation of a comprehensive model challenging. Traditional word-association paradigms cannot cover a sufficiently large range of words, even when converted into a "game with a purpose" and crowd-sourced to players on the Internet (Hees, Khamis, Biedert, Abdennadher, & Dengel, 2013).

A previous model of the DRM task estimated word associations from co-occurrence information in a text corpus, using the latent semantic structure to "recall" words that are semantically similar to the stimuli words (Johns & Jones, 2009). As the authors themselves noted, these lexical-semantics techniques only capture the structure of memory at best, but do not shed light on the recall processes. While the resulting model leads to good fits for the stimuli and lure recall rates from the original study, the computational linguistic techniques used were not designed to model recall tasks, using instead a custom procedure as a generate-and-test model. Furthermore, these models cannot accommodate complex reasoning with the encoded concepts, meaning that the knowledge captured by these associations is unusable for modeling human inference.

This paper is instead an exploration of the spreading activation mechanism that was originally hypothesized, using large computational knowledge bases as the semantic network and taking advantage of the existing memory mechanisms of cognitive architectures. The assumption is that the organization of these knowledge bases naturally encode association information, with more strongly associated concepts represented by nodes separated by a shorter network distance. Gleaning association information from computational knowledge bases would be a step towards the ideal of a single source of semantic knowledge that can be broadly used to model both human associations and inference.

Model Description

This section first describes the relevant components of the Soar cognitive architecture, before describing the model built using Soar.

Soar's working memory contains knowledge that is available for immediate reasoning. Working memory is represented as an edge-labeled directed graph, which is matched on and modified by procedural rules. In addition to knowledge in working memory, Soar has a long-term semantic memory, which contains general knowledge about the world. Each piece of knowledge (a node) in either memory is known as a memory element. Knowledge in semantic memory must be retrieved into working memory before it can be used. To do so, a Soar agent must create a cue that describes features of the desired piece of knowledge. Each element in semantic memory is associated with a base-level activation value, which reflects the recency and frequency of the retrieval of the element. The more recently and frequently an element is retrieved, the higher its activation value; however, the activation automatically decays over time. Unlike ACT-R, there is no retrieval threshold in Soar, and activation only serves as a bias mechanism but does not filter the retrieval. When the agent creates a cue, semantic memory returns the most-activated memory element that matches the cue, and places it in working memory to be matched on by procedural rules.

Spreading activation, as the hypothesized mechanism that leads to false memories, operates on the knowledge in semantic memory. Unfortunately, there is no standardized spreading activation algorithm, nor is there consensus on the meaning of spreading activation. In Soar, every retrieval of a memory element not only boosts the activation of that element, but also boosts the activation of neighboring elements in semantic memory, hence "spreading" the activation (Li & Laird, 2015). The number of elements that receive a boost is implicitly defined by a maximum spreading depth parameter, with a spreading depth of zero meaning that only the retrieved element receives a boost. All neighboring elements (regardless of edge direction) receive the same boost – the effect is not attenuated by distance, nor are there differential effects due to the strength of the connection between elements. In fact, the boost

due to spreading is indistinguishable from the boost received by the element retrieved; both changes are to the base-level activations of the elements and will therefore affect future retrievals. This is notably different from the spreading activation in ACT-R, which comes from elements in working memory, is considered separately from the base-level activation of memory elements, and only affects the current retrieval. Since the sources of activation (the stimuli words) are not present (not in working memory) at the time of recall in the DRM task, our model uses Soar's spreading activation mechanism in order to take advantage of its temporal extent. We note that this is not to say a model cannot be built with ACT-R's declarative memory, only that Soar's semantic memory more directly reflects of the hypothesized mechanism by which spreading activation occurs (ie. that the effects of spreading activation persists from the presentation phase to the recall phase).

Agent Description

A Soar agent plays the role of an experimental participant in our approach. Before a list of words is presented, the agent's semantic memory is pre-loaded with the knowledge base for the experiment. The base-level activation of each element is left at the default value zero, as there is no consistent method of doing so for all three databases. Once the database is loaded, the agent is sequentially presented with the stimuli words as strings. The agent must then retrieve the element that represents the associated concept from semantic memory, causing activation to spread to neighboring elements. Only after this retrieval is the next stimulus word presented, at which point the agent removes all previous elements from working memory. After all fifteen words from a list have been presented, the agent enters the recall phase. It retrieves the fifteen most activated words (without repetition) from semantic memory, from which the recall statistics are calculated. The semantic memory of the agent, including the activation of the elements, is then reset for the presentation of the next list.

Our model therefore has two free parameters: the base-level activation decay rate, and the maximum depth to which activation spreads. Since there are no probabilistic elements in

this model, the results are deterministic; all follow results are from a single presentation of the DRM lists.

We note two caveats to this model. First, the base-level activation of each element in the knowledge base is not initialized, and instead left at the default value of zero. Selecting the initial activation is a non-trivial problem, and in itself presents a challenge to any spreading activation model of the DRM paradigm. Activation is most commonly interpreted as encoding frequency information, which could be either calculated from a new corpus (Johns & Jones, 2009) or using existing datasets such as Google Ngrams. Applying this information, however, requires mapping the word to a node in the semantic network, which is problematic for two reasons. First, word frequency is inherently linguistic, when a better match is the frequency of the *concepts* that the words represent. Second, the purpose of this paper is not to create a single semantic network that perfectly models false memory, but consider the robustness of spreading activation across different existing computational knowledge bases. There is no automatic method for mapping words to their concepts, and manually doing so for all words is infeasible. Other any non-frequency-based heuristics, such as using the degree of each node (Salvucci, 2015), not only means that activation levels are not consistent between knowledge bases, but would also mean that the results would vary wildly between knowledge bases, detracting from the comparison between knowledge bases Instead, we opted to keep all activation at their initial levels, to focus on the effects of spreading activation.

The second caveat to our agent is the design of the recall phase. In the human experiments, the participants were given 2 to 2.5 minutes to recall as many words as possible; no information was provided on the number of words recalled, nor on the percentage of confabulations. In contrast, the agent in this model only retrieves the first 15 words, equivalent to a retrieval every eight seconds – a slow but not unreasonable rate. Using ACT-R's simulated retrieval times to approximate the procedural constraints would likely lead to the opposite problem of too many recalled words, since retrievals take less than a

second by default (even with additional time for rule firings). This is calculated using the retrieval time equation $Fe^{-(fA)}$ (in seconds), where F is the latency factor parameter, f the latency exponent parameter, and A the activation of the memory element (which must be above a retrieval threshold τ to be retrieved). By default, ACT-R sets $F = 1.0$, $f = 1.0$, and $\tau = 0.0$; this means for a memory element to be retrievable at all, the retrieval will take at most $e^{-0} = 1$ seconds, with faster retrievals for more activated elements. While τ could be adjusted, it remains that the original study did not provide sufficient information to fit this parameter. Additional data and more advanced memory mechanisms – perhaps rules for determining whether a retrieved word should be reported as a stimuli – may be needed to model the DRM task with higher fidelity.

While we acknowledge that both initializing activation and recall protocol are important components of a memory model that is not captured in this work, we also believe that they are not the main reason for model mismatch. We justify this belief in the General Discussion and Conclusions section.

Metrics

We are interested in two key measures that were obtained from the original false memory study:

- The stimuli recall rate, which is the proportion of stimuli words recalled after the presentation of a list, averaged over all 36 lists. The original study reports a stimuli recall rate of 62%, meaning that on average participants recalled 62% of the fifteen words in a list.

- The lure recall rate, which is the proportion of the 36 lists in which the lure was (falsely) recalled. Note that this metric is about a proportion of *lists*, and not about a proportion of the *stimuli words* in a list, and thus has no direct relationship to the stimuli recall rate. The original study reports a lure recall rate of 55%, meaning that on average participants had a false memory of the lure on 55% of the lists.

Before we describe the three experiments with different knowledge bases and their results, we reiterate that the goal of this work is not necessarily to perfectly model the stimuli and lure recall rates. We are not looking for the exact depth limit to spreading activation that should be used in future false memory models. Rather, the experiments below should be seen as an exploration of some of the challenges that cognitive modelers may face when attempting to leverage large knowledge bases, especially on tasks for which the knowledge bases are not designed. Towards this goal, while the metrics above provide a rough sense of the goodness of fit, the discussion for each experiment is more focused on properties of the knowledge base that led to those results.

Experiment 1: Hand-crafted Network

The goal of this experiment is to validate spreading activation as a viable explanation for false memory in the DRM task. The semantic network used in this experiment was created manually from the words in the *"needle"* and *"doctor"* lists. For each list, the fifteen stimuli words are all connected to the lure, with additional connections created based on whether the words are intuitively and informally associated. For example, *"pin"*, *"thimble"*, and *"prick"* are all connected, while none of the three are connected to *"haystack"*. Finally, four connections were added between the stimuli words of the two lists, such as *"injection"* (from the *"needle"* list) and *"medicine"* (from the *"doctor"* list) and *"hurt"* and *"sick"*, for a total of 109 edges between 32 nodes. It is important to note that the resulting network is representative of how semantic networks are depicted in non-computational literature.

Only the *"needle"* and *"doctor"* lists were presented using this network, with an activation decay rate of 0.5 and a spreading depth limit of 1. The results for both lists are similar. The lure is the first word to be retrieved (as it has the highest activation), with the stimuli words for the list retrieved afterwards. As would be expected, since activation spreads only to the immediate neighbors of the stimuli words, the four words that bridge the two lists are also activated, but not the lure of the not-presented list.

Although only two lists are used for this experiment, there is no reason to believe that the results would not generalize to similar hand-crafted semantic networks for the other lists. The quantitative results cannot be meaningfully compared to the stimuli and lure recall rates of the original study; however, the qualitative results are in line with the description that the lure is more highly activated than some stimuli words. While there is a tendency for words towards the end of a list to be retrieved first – as would be consistent with the decay of activation over time – the actual order of words retrieved is also affected by the structure of the semantic network due to spreading activation. Since the retrieval order would once again be different if the activation was initialized with other information, the rest of this paper does not consider the order in which words are retrieved. Regardless, this experiment suggests that spreading activation on a naive semantic network could cause the retrieval of the lure, which in this model indicates the formation of a false memory.

Experiment 2: WordNet

WordNet (Miller, 1995) is a database containing lexical knowledge, and is widely used both independently (for tasks such as parsing and word sense disambiguation) as well as in conjunction with other knowledge bases and ontologies. Nodes in WordNet represent not only words and phrases (for example, "sewing needle"), but also additional information about the meaning of those words, including word meanings (*senses*), synonym sets (*synsets*), antonyms, and certain types of entailments (for example, buying entails paying, so "buy" is connected to "pay"). WordNet nodes that represent words can be identified by an outgoing edge labeled `string`, which links to a string representation of the word; these edges do not exist for other concepts (such as synsets). The version of WordNet imported into Soar's semantic memory contains over 474,000 nodes and 1.7 million edges. Although WordNet has been previously used in cognitive architectures (Derbinsky & Laird, 2011; Douglass, Ball, & Rodgers, 2009), neither of those tasks used spreading activation, which requires activation updates on a large number of nodes.

The Soar agent used in this experiment is roughly the same as the one used in the first experiment. The only difference is in the recall phase, when the agent restricts the retrievals to words by specifying the `string` edge in the cue. All words from the DRM lists are used as is, with the exception of *"Bic"* and *"Cross"* from the *"pen"* list. These pen companies do not exist in WordNet and were excluded from the experiment; the *"pen"* list therefore only contains thirteen stimuli words.

For this experiment, separate trials were run for different spreading depths (1 through 6) and different decay rates (0.25, 0.5, 0.75, 0.9).

Results

The overall results are shown in Figure 1. For each parameter setting, we plot both the stimuli recall rate and the lure recall rate, as well as average proportion of recalled words (out of 15) that are neither stimuli words nor the lure (which we shall call *external* words). The human data from the original DRM study is shown for comparison; the results for depths 1 and 2 are left out for reasons explained below. Across all parameter settings shown, the stimuli recall rate ranges from 9% to 41%, well below the reported rate of 62% in humans, while the lure recall rate ranges from 0% to 72%, compared to the reported rate of 55% in humans. In particular, using the ACT-R and Soar default decay rate of 0.5, a spreading depth of 5 results in a lure recall rate of 56%. In general, however, no parameter setting accurately matches human data on both stimuli and lure recall rates.

For spreading depths of 1 and 2, the stimuli words were consistently retrieved, while the lure was never retrieved. Upon examination, this is because WordNet is structured with most words only being connected through word senses and synsets. The node representing *"thorn"*, for example, is connected to three word senses, each of which is connected to a synset – which means that, within a network distance of two, *"thorn"* is not connected to any words at all, never mind the lure *"needle"*. Since the retrieval cue used by the agent limits results to words, the retrieval fails after the stimuli words are retrieved. This explains both the high

stimuli recall rate and why the lure is never retrieved.

The data shows additional trends regarding the stimuli and lure recall rates. In general, the spreading depth is *proportional* to the lure recall rate but *inversely proportional* to the stimuli recall rate. That is, the stimuli recall rate decreases as spreading activation extends deeper from the stimulus word, while the lure recall rate increases from the same manipulation.

These results can be explained by the same WordNet structure mentioned previously. When spreading activation is limited to nearby nodes, only a small number of words (as opposed to word senses, synsets, etc.) are boosted, hence the majority of words retrieved are the stimuli. When the depth limit is increased, however, spreading activation now reaches other words in the synsets. These words – which may include the lure – may in fact receive activation boosts spread from multiple stimuli words. The word "shot" falls into this category, as it means both *"injection"* and *"hurt"* (as in *a solid shot to the chin*). Other external words may simply be boosted by stimuli words later in the list, and therefore have higher activation during the recall phase than stimuli words earlier in the list. Together, this leads to a decrease in the stimuli recall rate as well as an increase in the lure recall rate.

Discussion

Although the lure recall rate from WordNet spans a range that includes the human lure recall rate of 55%, the structure and content of WordNet does not directly match human associations. The nodes representing the stimuli words in WordNet are not structured such that activation will spread to the lure. We discuss two categories of such failure here: cases where additional edges lead to model errors, and cases where edges are missing.

First, as we noted, WordNet is structured with individual words arranged in "spokes" around lexical constructs such as synsets. While synsets do represent some of the relationships between stimuli words and the lure – as in *"syringe"* and *"needle"* – they are not the only relationships around which words are organized. Since WordNet is a dictionary

in knowledge base form, it also contains information about the derived form of words, such as the relationship between *"inject"* and the words "injectable", "injecting", "injection", and "injector". With the exception of *"sit"* and *"sitting"* in the *"chair"* list, derived words do not appear in the DRM lists, and more importantly, are unlikely to be produced during human recall. This mismatch may be due to the *lexical* relationships encoded in WordNet, as opposed to the *conceptual* relationships on which spreading activation is hypothesized to occur. Human participants would only produce one word for each concept, but spreading activation (at least over WordNet) leads to the retrieval of multiple derived words. Algorithmic changes may be necessary before spreading activation can correctly model the generation of false memory; we propose one such change in the general discussion.

Although WordNet contains connections that extend beyond human associations, it fails to capture other relationships that the DRM lists exploit. A careful examination of the word lists reveals that they contain multiple types of associations. Some, such as antonyms (*"high"* and *"low"*), are encoded in WordNet despite being more conceptual. Others, however, are not captured despite being lexical in nature. For example, the *"high"* list contains the word *"noon"*, clearly intending to invoke the phrase "high noon". Crucially, while "high noon" does exist as a phrase in WordNet, it is not connected to its component words *"high"* and *"noon"*. At the same time, other idiomatic phrases, such as "*needle* in a *haystack*" and "making a *mountain* out of a *molehill*", are not represented in WordNet. Also missing are cultural references; the inclusion of *"tiger"* and *"bear"* in the *"lion"* list appears peculiar, but may be explained by the lyric *lions and tigers and bears, oh my!* from *The Wizard of Oz*. Unlike the first type of failure due to an over-abundance of connections, there is no algorithmic solution to missing data, at least not without expanding the database using a text corpus, which presents challenges of its own.

Mismatched and missing data is not unexpected in large knowledge bases, although in this case some of them seem to arise from WordNet's specialization in lexical knowledge. Our third experiment looks at whether a different knowledge base may lead to a better model of

human associations in false memory.

Experiment 3: DBpedia

DBpedia (Bizer et al., 2009) is a knowledge base created using information from the online encyclopedia Wikipedia. The nodes in DBpedia represent articles on Wikipedia (or more accurately, they represent the concepts that the Wikipedia articles describe), while the edges come from the categories to which the articles belong, as well as the *infoboxes* that provide basic information. As a result, the type and amount of information varies between concepts. The version of DBpedia used in this experiment contains 6 million nodes and 27 million edges.

The size and scope of DBpedia led to two differences in this experiment from the previous ones. First, since DBpedia does not contain a comprehensive dictionary of English words, and not all words in the DRM lists have their own Wikipedia article, the stimuli words can no longer be presented as strings. Instead, we manually mapped each word to a concept in DBpedia, mostly following the redirections on Wikipedia. This led to some words being mapped onto the same concept (*"waste"* and *"refuse"* both mapped onto "waste"), while others mapped onto concepts that are overly specific (*"garbage"* mapped onto "municipal solid waste"). More problematic were words that differed in meaning from their Wikipedia articles. Words from the *"thief"* list are good examples: Wikipedia does not contain articles for *"thief"*, *"robber"*, *"burglar"*, *"bandit"*, or *"criminal"*, only articles for "thievery", "robbery", "burglary", "banditry", and "crime". These words were excluded from this experiment.

To accommodate the size of DBpedia, a custom Python script that simulated spreading activation was used instead of Soar, although the same algorithm as Soar's semantic memory is followed. For this experiment, the fifteen "retrieved" concepts are simply the fifteen most-activated nodes. The size of DBpedia and the density of its connections remains daunting; as an example, a fifth of the nodes in DBpedia are only two connections

away from the nodes selected for the *"army"* list. One million concepts would be activated even at a spreading depth limit of 2, making spreading activation not only cognitively implausible, but with the raw DBpedia file being 10.56GB, also computationally expensive to simulate. That these limits are only reached for some stimuli words but not on others highlight that the DRM lists are not equal and, in fact, may be used as a dimension to understand false memory. We leave suggestions for more realistic models in the General Discussion and Conclusions section. As a result of these two problems, only about half the lists (seventeen) were used in this experiment, with an average of 14.1 concepts.

Results

Due to the reduced dataset, the results in this section should be treated with some skepticism; however, we believe they are nonetheless representative of using DBpedia to model false memory and human associations.

The overall results are shown in Figure 2. For spreading depth 1 at the default decay rate of 0.5, spreading activation on DBpedia resulted in stimuli and lure recall rates of 15% and 0% respectively; for spreading depth 2, the stimuli recall rate decreases to 3%, while the lure recall rate increases to 12%. These numbers follow the trends found from the WordNet experiment. To understand the low lure recall rate, we found it instructive to look at the *"shirt"* list, one of two lists for which the lure was consistently retrieved. Unlike other DRM lists, the *"shirt"* list is unique in that the vast majority of items belong to the same category. This shared classification means that the lure is only a network distance of two away from the stimuli words, and is therefore sufficiently boosted in activation for it to be retrieved. In contrast, the stimuli words for other DRM lists do not conform as neatly to the taxonomic structure of DBpedia – the lure is not as directly connected to the stimuli, causing the lure to not be retrieved.

That the lure is not retrieved, however, does not mean that the stimuli words are retrieved; the highly connected network structure also led to the low stimuli recall rate. Page

links on the Internet are known to have a small-world structure, where the pairwise distance between all nodes are small and where there are many nodes with large degrees. For example, *"anger"* is connected to "red", which in turn is connected to over 600 concepts, mostly organizations whose representational colors include red. Because these "hub" nodes are often connected to multiple stimuli words, their activation is boosted above that of the stimuli words and are retrieved instead, resulting in a low stimuli recall rate.

Discussion

The failures in both WordNet and DBpedia are representational; we discuss these issues in the next section. For DBpedia alone, we faced the additional difficulty of mapping the stimuli and lure words to a concept. One concern not yet raised is that the choice of concepts used to represent nodes requires association and reasoning on the part of the modeler. A number of words in the DRM lists are polysemous; *"prick"* and *"hurt"*, for example, would fit just as well as "goad" and "heckle" into a different *"needle"* list (as a verb instead of as a noun). If DBpedia is to be used for modeling associations and false memory, a better protocol would be for unknowing coders to determine which concepts correspond to the lure and the stimuli words. This would remove confirmation bias that may be inherent in how words are currently mapped to concepts.

General Discussion and Conclusions

This paper attempted to use large computational knowledge bases to model the human associations that lead to false memory in the DRM paradigm. Our model was able to qualitatively recreate the DRM false memory phenomenon, but only on a hand-crafted semantic network that resembles their traditional depiction. When large computational knowledge bases such as WordNet and DBpedia are used, however, the naive spreading activation algorithm fails to simultaneously match the stimuli and lure recall rates. We believe that these results are indicative of three general problems with using large knowledge

bases in cognitive modeling: missing concepts from the knowledge base, missing connections between existing concepts, and finally, the sheer amount of existing knowledge.

Missing Concepts The type of common sense knowledge required to make associations in the DRM task is neither lexical nor conceptual – it exists neither in a dictionary nor in an encyclopedia. One example of such knowledge is the fact that *"rubber"* is *"elastic"*, *"springy"*, *"flexible"*, and *"resilient"*. It is infeasible to manually encode all descriptions for all objects, and it may be necessary to employ techniques from information retrieval and natural language processing to extract this knowledge from text.

Missing Connections between Concepts Even for concepts/words that exist in the knowledge base, neither WordNet nor DBpedia fully capture the relationships between their nodes. Some of these missing relationships, such as phrases from popular culture, can only be obtained through similar means as the missing concepts/words; others, by systematically adding edges to these knowledge bases, such as connecting phrases to their component words. Perhaps more relevant for cognitive modelers, however, is that there is no consensus on the cognitive plausibility of the content and structure of knowledge bases. In understanding the experimental results of this paper, we have tried to determine how the stimuli words relate to the lure, and whether these relationships generally apply to other concepts. A complete catalog of human associations would more clearly indicate the types of connections that knowledge bases currently lack.

Amount of Knowledge Available The final problem of the scale of the knowledge base is only made worse by the addition of missing knowledge. The solution here may be more algorithmic in nature, by modifying the spreading activation algorithm such that it remains valid as the size of the knowledge base grows. One possibility is for spreading to occur only on particular edges, perhaps informed by the context of the retrieval. This is similar to using theory to extract a smaller, more specialized network on which the network distance may be more meaningful (Tenenbaum, Griffiths, &

Kemp, 2006). Such an algorithm would reduce the computational requirements of spreading activation, while simultaneously filtering out connections that are irrelevant for fitting human data. The same mechanism may also allow lexical, conceptual, and other knowledge to exist in the same knowledge base, as a unified semantic memory to be used in cognitive modeling, without leading to the confusions demonstrated in the results of this paper.

These issues lead us to believe that, although the lack of activation initialization is unrealistic and the recall phase of the agent does not directly match that of human experiments, they are not the most significant obstacle to a spreading activation model of the DRM false memory task. Neither changes to the initialization, nor to the design of the model, would have led to the retrieval of concepts that are not in the knowledge base or are not connected to the stimulus. Similarly, neither changes would reduce the computational resources - as calculated by the number of concepts activated - that the model requires. Thus, while these components of the model may be simplistic, focusing on these aspects while ignoring the algorithmic structure of the model is akin to the proverbial drunk looking by the street lamp for their keys. Alternately, these computational knowledge bases represent not the knowledge of any individual person, but the collective knowledge of many. Modeling the performance of an individual may avoid some of the problems discussed here, but would present other challenges, such as determining which pieces of knowledge to lesion. Instead, the focus should be on developing more refined algorithms that can efficiently operate on millions of concepts and relations. Only with such algorithms can large computational knowledge bases become a valuable resource for modeling the wealth of background knowledge that participants bring into experiments.

References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia — a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(3), 154–165.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22.

Derbinsky, N., & Laird, J. E. (2011). A functional analysis of historical memory retrieval bias in the word sense disambiguation task. In W. Burgard & D. Roth (Eds.), *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)* (pp. 663–668). San Francisco, CA: AAAI Press.

Douglass, S. A., Ball, J., & Rodgers, S. (2009). Large declarative memories in ACT-R. In *Proceedings of the 9th International Conference on Cognitive Modeling (ICCM).* Manchester, UK.

Hees, J., Khamis, M., Biedert, R., Abdennadher, S., & Dengel, A. (2013). Collecting links between entities ranked by human association strengths. In *Proceedings of the 10th European Semantic Web Conference* (pp. 517–531). Montpellier, France.

Johns, B. T., & Jones, M. N. (2009). Simulating false recall as an integration of semantic search and recognition. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society)* (pp. 2511–2516). Amsterdam, The Netherlands.

Li, J., & Laird, J. E. (2015). Spontaneous retrieval from long-term memory for a cognitive architecture. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)* (pp. 544–550). Austin, TX.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Roediger, H. L., McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in producing false remembering. In M. A. Conway, S. E. Gathercole, & C. Cornoldi (Eds.), *Theories of Memory II* (pp. 187–245).

Salvucci, D. D. (2015). Endowing a cognitive architecture with world knowledge. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society)* (pp. 1353–1358). Pasadena, CA.

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory and Cognition*, *27*(3), 494–500.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

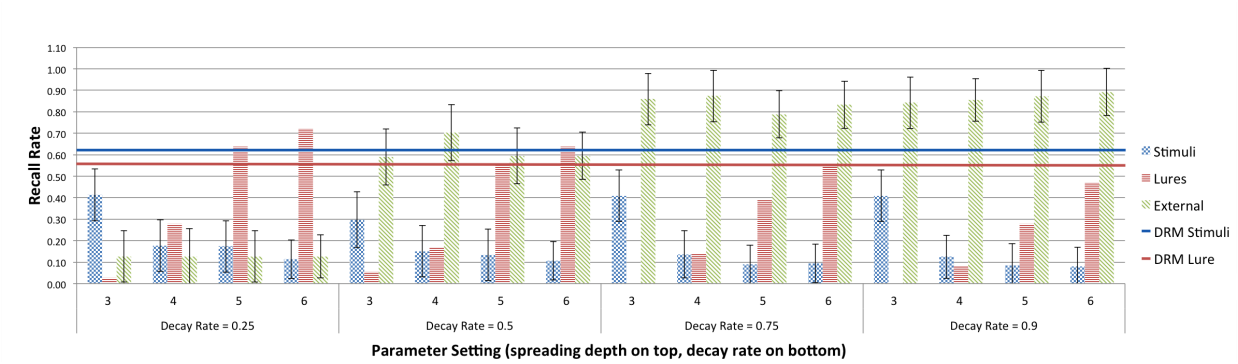

Figure 1. Lure, Stimuli, and External Recall Rates for WordNet. The error bars show one standard deviation from the mean over the 36 lists. The lure recall rate, as proportion of lists where the lure word is recalled, does not have a distribution.

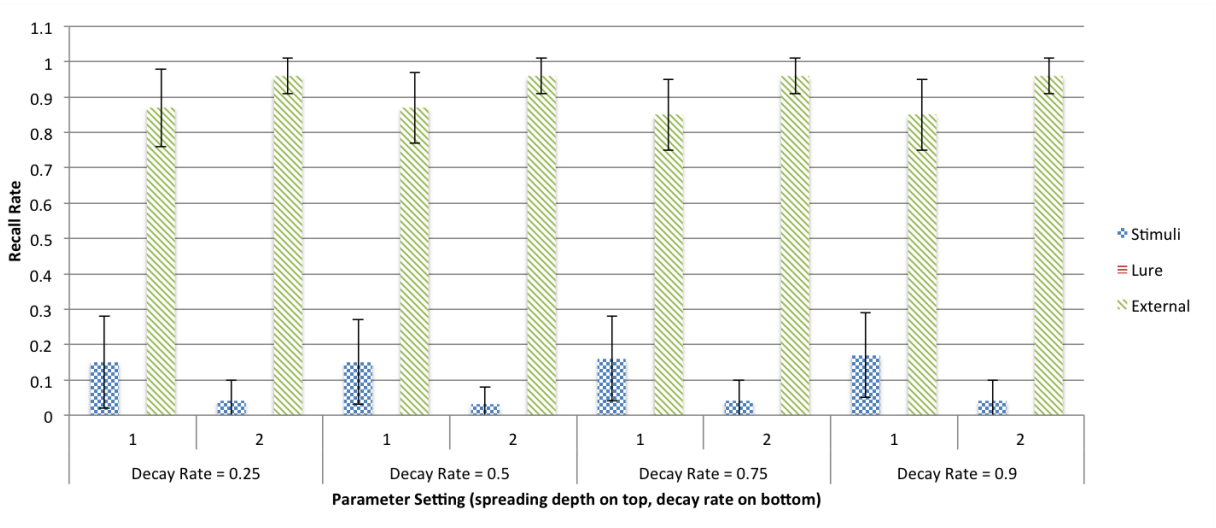

Figure 2. Lure, Stimuli, and External Recall Rates for DBpedia. The error bars show one standard deviation from the mean over 17 lists. The lure recall rate, as proportion of lists where the lure word is recalled, does not have a distribution.