Integrating Distributed Semantic Models with an Instance Memory Model to Explain False Recognition

Minyu Chang (minyu.chang@mcgill.ca)
Department of Psychology, McGill University
Montreal, QC, H3A 1G1

Brendan T. Johns (brendan.johns@mcgill.ca)
Department of Psychology, McGill University
Montreal, QC, H3A 1G1

Abstract

In this paper, we simulated true and false recognition in the Deese/Roediger/McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) paradigm by incorporating word embeddings derived from distributed semantic models (word2vec) into an instance memory model (MINERVA2). Previously, Arndt and Hirshman (1998) demonstrated that MINERVA2 (Hintzman, 1984) could capture multiple classic false recognition findings with randomly generated word representations. However, as randomized representations deviate systematically from semantic representations learned from the natural language environment, there remains uncertainty about whether MINERVA2 can capture the false memory illusion when scaling up to real-life complexity in word representations. To address this uncertainty, we used word2vec embeddings that are derived from large corpora of natural language instead of randomized representations in MINERVA2. Our results showed that MINERVA2 can still capture the standard true and false recognition, and it can also accommodate the true and false recognition effects of various classic manipulations (e.g., associative strength, number of associates, divided attention, retention interval).

Keywords: DRM illusions; false recognition; MINERVA2; distributed semantics; memory models

Introduction

False memory refers to remembering events that have never been experienced or events attributed to wrong sources. In settings like eyewitness memory and psychotherapy, false memory can lead to severe downstream consequences (Thompson-Cannino & Cotton, 2009; Wilkomirski, 1997). Thus, it is critical to understand the underlying mechanisms of false memory formation. In false memory research, one of the most widely used paradigms is the Deese/Roediger/McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) paradigm. In this paradigm, participants encode lists in which all the words are forward associates (e.g., bed, pillow, snore, nap, etc.) of an unpresented critical lure (e.g., sleep). As a result, the critical lures are usually falsely recognized at a high rate, and such memory illusions are robust across different populations and under various conditions (Gallo, 2006; Chang & Brainerd, 2021).

Theoretical Explanations for the DRM illusion

Two of the most influential theories of the DRM illusion are the fuzzy-trace theory (FTT; Brainerd & Reyna, 1998) and activation/monitoring framework (AMF; Roediger et al., 2001). FTT assumes that people separately store and retrieve two distinct episodic memory representations: verbatim and gist traces. Verbatim traces refer to surface details that are diagnostic of the prior occurrence of a specific item. Gist traces refer to the semantic and elaborative content of the items. Verbatim and gist traces are stored and retrieved in parallel. In terms of storage, gist traces are assumed to decay at a slower rate than verbatim traces. In terms of retrieval, retrieval of verbatim traces leads to a vivid recollection of an item's prior presentation, which allows people to recall or recognize the item perfectly. Thus, verbatim retrieval supports true memory and suppresses false memory. Retrieval of gist traces, however, supports a non-specific feeling of familiarity that can lead to both true memory and false memory (Brainerd & Reyna, 2005). To illustrate, the critical lure *chair* and list words *table*, *desk*, *couch*, etc. will both feel familiar as they fit into the gist of "furniture". Thus, gist retrieval will lead to both true memory for list words like *couch* and false memory for the critical lure *chair*.

Alternatively, according to AMF, encoded list words are assumed to be represented as interconnected nodes in an associative network (Collins & Loftus, 1975). Words that are more strongly associated with each other appear closer in the network. When a word is encoded, its node will be activated, and the activation will spread from the activated node to other nodes that are connected to it, with the strength of the spreading activation being proportional to the distance between the nodes. Consequently, when a DRM list was studied, the critical lure of the list receives repeated activation that is spread from the nearby nodes of encoded list words, and thus it is likely to be falsely remembered. For the sake of memory accuracy, people also implement a monitoring operation to discriminate between words that are actually encoded and words that are merely activated via spreading activation.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Computational Models of the DRM Illusion

Despite the sophisticated theoretical explanations of the DRM effects, there have been limited attempts at formal computational modeling of the DRM illusion. One of the first attempts was provided by Arndt and Hirshman (1998), who stimulated DRM false recognition with the MINERVA2 model (Hintzman, 1984). MINERVA2 is an instance model of memory (see Jamieson, et al., 2022 for a recent review), which posits that each encoded item is a unique trace and recognition is driven by the similarity between the test probe and the encoded items' traces. Arndt and Hirshman chose MINERVA2 because the model had successfully accounted for schema abstraction (Hintzman, 1984) and categorical false recognition (Hintzman, 1988), which is similar to DRM false recognition. Moreover, MINERVA2 is flexible in terms of accommodating different factors that affect false recognition. For instance, Arndt and Hirshman showed that by simply manipulating stimuli characteristics and learning rates, the model can provide a satisfactory approximation to behavioral data under the manipulations of presentation rate, number of associates, and associative strength.

Although Arndt and Hirshman demonstrated that MINERVA2 is a promising process model for explaining DRM false recognition, there is a critical limitation in their study. Namely, they used randomized representations instead of realistic semantic representations of words. Specifically, they represented a critical lure as a randomly generated vector containing either -1, 0, or 1 and represented the respective list words by inserting random perturbations into the vector of the critical lure. Thus, list words shared systematic similarities in vector composition with the critical lures. By doing so, it is assumed that the similarity between those randomly generated vectors provides a good approximation to the semantic relatedness among real words.

However, Johns and Jones (2010) showed that this assumption is not true. They demonstrated that the probability distribution of semantic similarity between two random words is positively skewed with semantic representations but symmetrical with randomized representations. This means that two randomly selected words tend to be less similar in natural language than when represented by randomized vectors. Moreover, Johns and Jones rerun Arndt and Hirshman's model simulations both with randomized representations and with semantic representations. For the latter, they derived word vectors from Steyvers, Shiffrin, and Nelson's (2004) word association space (WAS) model. Using the randomized representations, they successfully replicated Arndt and Hirshman's finding. However, with the WAS vectors, MINERVA2 provided a worse fit to the actual data. Therefore, the use of randomized presentation can threaten the validity of conclusions made about a process model. Namely, as randomized representations can fail to capture important features of natural language, even though a process model performs well with randomized representations, it may not do so with realistic representations.

Semantic Representations of DRM Lists

Recently, there have been attempts at integrating realistic vector representations of words that are derived from distributed semantic models (i.e., word embeddings) with computational process models in simulating DRM false recognition (e.g., Johns, Jones, & Mewhort, 2012, 2021; Reid & Jamieson, 2023). Here, distributed semantic models, refer to a cluster of models that represent word meanings as high-dimensional vectors, which are extracted from large corpora of natural language (Günther, Rinaldi, & Marelli, 2019). For example, Johns et al. (2012) used word-by-document co-occurrence vectors that are derived from large text corpora as word representations (similar to the latent semantic analysis [LSA]; Landauer & Dumais, 1997) and imported them into the Recognition through Semantic Synchronization (RSS) model.

In distributed semantic modeling, it has been found that prediction-based models such as word2vec (Mikolov et al., 2013) tend to outperform the more traditional count-based models such as LSA in predicting human performance across various tasks (e.g., semantic priming and semantic relatedness ratings; Baroni, Dinu, & Kruszewski, 2014; Mandera, Keuleers, & Brysbaert, 2017). The difference between these model types is that count-based models rely on counts of word-context occurrence (e.g., how frequently a word appears in a document) to determine the words' semantic representation. In contrast, prediction-based models utilize error-driven learning where a target word is predicted from words that co-occur with that word in a context or vice versa. Given that prediction-based models such as word2vec seem to provide a better approximation to human data in various semantic tasks, it is reasonable to expect that such models may do so for the DRM paradigm as well.

Gatti et al.'s (2022) recent finding supports such speculation. In this study, they derived word vectors for DRM lists from a word2vec semantic space (Mandera et al., 2017), and used frequency-weighted mean cosine similarity between the studied words' vectors and the critical lure's vector to predict false recognition. Their results demonstrated that both local similarity (i.e., average similarity between a critical lure and the list words on the given DRM list) and global similarity (i.e., average similarity between a critical lure and the list words on all encoded DRM lists) significantly predicts false recognition. Moreover, they also found a positive correlation between local similarity and mean backward associative strength (MBAS; i.e., the average probability of eliciting the critical lure given a list word), suggesting that the word2vec semantic space successfully captures the semantic structure of DRM lists.

The Current Study

On the one hand, Arndt and Hirshman (1998) demonstrated the potential of MINERVA2 to explain false recognition in the DRM paradigm. However, the traditional MINERVA2 model is devoid of realistic semantic representation of words. On the other hand, it has recently been demonstrated that distributed semantic models are a promising candidate for capturing the semantic representations of the DRM paradigm. The current study is aimed at integrating embeddings from a distributed semantic model (word2vec) with MINERVA2 to simulate various classic findings in the DRM paradigm. We choose the word2vec model because it has been verified to provide a realistic semantic representation of the DRM paradigm (Gatti et al., 2022), and it is also one of the most widely used distributed semantic models in the field. We chose the MINERVA2 model because of its prior success in modeling the DRM illusion and its parsimony and flexibility (Arndt & Hirshman, 1998). Below, we first explain how we derive word embeddings from the word2vec model and then discuss the machinery of the MINERVA2 model.

Representations: Wod2vec embeddings The word2vec model is one of the most influential distributed semantic models. This model uses a neural network architecture that includes an input layer, a hidden layer, and an output layer. In the current study, we used the continuous bag of words (CBOW) implementation, in which context words surrounding a target word (input layer) are used to predict the target word. The network is trained via backpropagation, in which the activation weights in the hidden layer are adjusted to minimize the error between the network's output layer and the target words. The vector representation of the target words is thus extracted using the activation weights in the hidden layer (Günther et al., 2019; Kumar, 2021).

We obtained the vector representations of words from a pre-trained word2vec model (Mikolov et al., 2013), which is available at https://code.google.com/archive/p/word2vec/. This model includes three million words and phrases, which was trained on a subset of Google News datasets that contains about 100 billion words. All the word vectors derived from the model have a dimensionality of 300.

Process Model: MINERVA2 In the MINERVA2 model, each encoded word is regarded as a unique memory trace, which is represented as a vector of features. Traditionally, MINERVA2 uses randomly generated vectors that contain feature values of either -1, 0, or 1, with 1 and -1 indicating whether the word has the feature or not, and 0 indicating encoding failure of the feature (Arndt & Hirshman, 1998; Hintzman, 1984, 1988). However, the current study used word vectors that were retrieved from the word2vec model; hence, the features are represented with continuous rather than categorical values.

During encoding, the probability of successfully encoding a feature is controlled by a learning parameter *L*. That is, each feature has a probability *L* of being properly encoded and stored in memory, and a probability 1 – *L* of being replaced by random noise. In the current study, we used noise values that are randomly sampled from a uniform distribution between -.05 and .05 (this range was chosen arbitrarily as the level of noise had a limited impact on model performance).

During retrieval, the vector of each test probe is compared to all the word vectors stored in memory, yielding a similarity value that indicates how similar the probe is to all the words stored in memory. This process is expressed in the following equation:

$$S_i = \cos(P, T_i) \qquad (1)$$

Where S_i indicates the cosine similarity (i.e., normalized dot product) between a test probe P and the i^{th} memory trace T_i . Here, each memory trace is a word on the study list. If the test probe and the memory trace are identical, their cosine similarity will be 1, and if they are completely irrelevant (i.e., orthogonal), their cosine similarity will be 0.

Next, an activation value is generated by raising the similarity values to the exponent of 2. In order to preserve the signs of the similarity values, we reversed the sign after squaring if the similarity value is originally below zero. The traditional MINERVA2 model typically uses cubing to preserve the sign of similarity, we used squaring instead due to preliminary analyses showing that cubing led to an excessive overestimation of false alarms for critical lures. The calculation of activation values is shown below:

$$A_i = \begin{cases} S_i^2, if\ S_i > 0 \\ -S_i^2, if\ S_i < 0 \end{cases} \qquad (2)$$

Here, S_i is the similarity value between a test probe P and the memory trace T_i, and A_i is the activation value after squaring the S_i, with the signs being preserved. Via the squaring process, the model amplifies the influence of items that are highly similar to the test probe relative to those that are dissimilar to the probe. Last, the echo intensity value is calculated by summing all the activation values for the given probe.

$$I = \sum_{i=1}^{M} A_i \qquad (3)$$

where I is the echo intensity value, A_i is the activation value, and *M* is the total number of memory traces. Finally, a recognition decision was made by comparing the echo intensity to the decision criterion *C*. Thus, if *I* is above *C*, then the probe will be recognized as "old," and if *I* is below *C*, the probe will be recognized as "new."

Simulations

In this section, we first simulate standard true and false recognition data in the DRM paradigm. Then, we proceed to simulate the true and false recognition results under various classic manipulations, including manipulations of associative strength, number of associates, attention levels, and retention interval. In all simulations, we first retrieved the word vectors for the given DRM lists from word2vec and then input the vectors into the MINERVA2 model.

Simulation 1: Levels of DRM True and False Recognition

The first set of simulations was meant to test whether the MINERVA2 model can account for the standard true and false recognition patterns in the DRM paradigm. Specifically, we compared the levels of recognition predicted by the MINERVA2 model to behavioral data from three classic DRM studies (Gallo & Roediger, 2002; Roediger & McDermott, 1995; Stadler, Roediger, & McDermott, 1999).

Method For the simulation of Roediger and McDermott (1995; Experiment 2), we used all 24 DRM lists in their experiment. For the other two simulations, we used 34 of 36 lists in Stadler et al. (1999) and 25 of 28 lists in Gallo and Roediger (2002; Experiment 1). Four lists were excluded because a few words on those lists were not available in the word2vec model we used.[1]

We simulated a total of 1000 trials for each experiment. In each simulated trial, 10 DRM lists were randomly selected from the respective list pool as study materials, and thus 150 words (15 words/list) were entered into the study list. The test probes include 10 critical lures, 30 targets (the words at the 1th, 8th, and 10th position on the 10 DRM lists), and 20 unrelated lures (the critical lure and the list words at the 1th, 8th, and 10th position of five lists randomly selected from the unused DRM lists). The mean hit rates for the targets and the mean false alarm rates for the critical lures and unrelated lures were recorded across the 1000 stimulated trials. We used a learning rate of .6 for all simulations, and slightly different decision criteria for the three studies: 2.35 for Roediger and McDermott (1995), 2.3 for Stadler et al. (1999), and 2.15 for Gallo and Roediger 2002).

Results Fig. 1 displays the data simulated by the MINERVA2 model compared to the actual data. As can be seen there, the model prediction closely approximates the qualitative trends in the actual data: Both true recognition for targets and false recognition for critical lures are well above the recognition for unrelated lures. There was only one systematic variation that the model tended to overestimate the false alarm rates for critical lures and underestimate the hit rates for targets. This may be due to the word2vec representations encoding a stronger semantic relationship among words on the DRM lists than human participants.

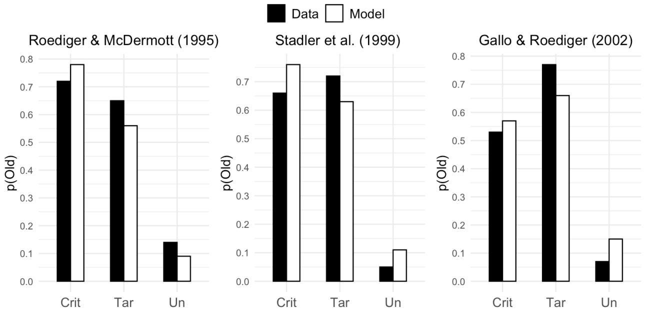

Figure 1: The simulated and actual recognition results. Crit = critical lure. Tar = target. Un = unrelated lure.

Simulation 2: Effects of Associative Strength

In early research on DRM false recognition, Deese (1959) found that the associative strength between list words and critical lures is highly correlated with the cross-list variability in false recall. Later, Roediger et al. (2001) replicated the finding and extended it to false recognition, establishing associative strength as a strong predictor of false memory. Behavioral studies have experimentally manipulated the associative strength across lists and confirmed that strong lists (i.e., lists with higher associative strength) produced more false recognition than weak lists (i.e., lists with lower associative strength; e.g., Arndt & Hirshman, 1998; Brainerd, Reyna, & Forrest, 2002; Gallo & Roediger, 2002). Such a finding is consistent with AMF since critical lures should receive higher activation if list words are stronger associates (closer to the critical lure in the associative network). The finding is also consistent with FTT, as stronger associations between critical lure and list words are usually accompanied by strong semantic relatedness too, which in turn form a stronger global gist (Brainerd, Chang, & Bialer, 2020). Here, we tested whether the MINERVA2 model can capture the differences in false recognition between strong and weak DRM lists.

Method We used 22 of the 24 lists from Gallo and Roediger (2002; Experiment 2), including 11 strongly associated lists and 11 weakly associated lists.[2] In each simulated trial, we randomly sampled five strong lists and five weak lists. We used a constant learning rate (L = .6) and decision criterion (C = 2.3) across strong and weak lists. Thus, the locus of the effects of associative strength lies in the different words encoded in memory, rather than different processing across list types. The average hits for targets and false alarms for critical lures were recorded across 1000 simulated trials.

Results The results are displayed in Fig. 2, which demonstrates that the MINERVA2 model again simulates a similar qualitative pattern compared to the actual data from Gallo and Roediger (2002; Experiment 2). Specifically, false alarms for critical lures dropped sharply from strong lists to weak lists, whereas hits for targets only dropped slightly from strong to weak lists.

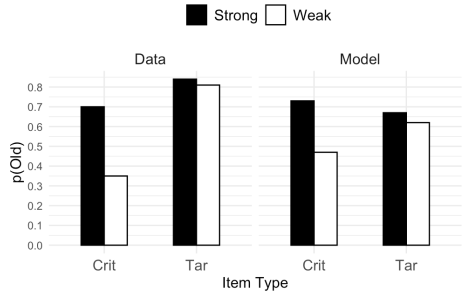

Figure 2: The simulated and actual recognition of Gallo and Roediger (2002). Crit = critical lure. Tar = target.

[1] For the simulation of Stadler et al. (1999), two lists whose critical lures are *army* and *city* were excluded. For the simulation of Gallo and Roediger (2002), three lists whose critical lures are *citizen*, *city*, and *swift* were excluded.

[2] Two lists whose critical lures are *city* and *citizen* were excluded because some list words are unavailable in word2vec.

Simulation 3: Effects of Number of Associates

A classic finding in the DRM literature is that false recognition increases with the number of associates studied (Arndt & Hirshman, 1998; Gallo & Roediger, 2003; Hutchison & Balota, 2005; Robinson & Roediger, 1997). Namely, lists that contain more forward associates to the critical lure tend to elicit higher false recognition for the critical lure. In addition, lists with more associates also elicited higher true recognition for targets, although such an effect was weaker than that for critical lures, and it was also less robust (Gallo, 2006). This is consistent with AMF such that the critical lure, which is strongly associated with all list words, will receive increased spreading activations as the number of list words increase. However, list words themselves may not be as strongly associated with each other as with the critical lures, and thus they may receive less increase in activation when other associates are added to the list. FTT predicts the same pattern with different reasons. It assumes that increased number of associates forms a more coherent global gist but simultaneously introduce more interference with verbatim traces. Thus, false recognition is elevated by the enhanced gist traces, whereas true recognition is under the contrasting influence of enhanced gist traces and impaired verbatim traces. In the current simulation, we test whether MINERVA2 can account for how true and false recognition varies as a function of the number of associates.

Method We again simulated 1000 trials. In each simulated trial, we randomly selected five DRM lists from a combined list pool of Stadler et al. (1999) and Gallo and Roediger (2002). For each of the five lists, we added the first 3, 6, 9, 12, or 15 associates of the list to the study list. Again, we used constant learning rate (*L* = .5) and decision criterion (*C* = .85) across all five types of lists (3-, 6-, 9-, 12-, 15-word), so the recognition effects are completely dependent on different lists rather than different processing.

Results The false alarms for critical lures and hits for targets are plotted as a function of the number of associates in Fig. 3. There, the data simulated by MINERVA2 closely tracks the pattern of the actual data in Robinson and Roediger (1997; Experiment 2). Specifically, false recognition for critical lures increases constantly as the number of associates increases. True recognition for targets also increases with the number of associates, but to a less extent than false recognition, as the slope for true recognition is much flatter than that for false recognition. The model predicts a slightly steeper slope for both critical lures and targets than the actual data, which is again possibly due to word2vec model being too adept at grasping the semantic relationship among the list words. Thus, as the number of associates increases, word2vec captures a stronger semantic relation than humans.

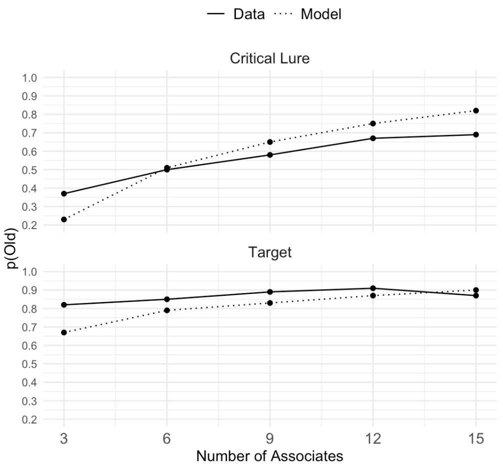

Figure 3: The simulated and actual recognition of Robinson and Roediger (1997).

Simulation 4: Effects of Divided Attention

In the full versus divided attention manipulation, participants in the full attention condition simply encode word lists without distractions whereas participants in the divided attention condition encode the word lists when simultaneously performing a secondary task. According to FTT, divided attention is mainly a verbatim manipulation, such that it impairs the encoding of verbatim traces more than that of gist traces (Brainerd et al., 2019). Consistent with FTT, Jacoby (1996) demonstrated that recollection was sharply reduced by divided attention, whereas familiarity remained relatively stable. As verbatim traces support true memory rather than false memory while gist traces support false memory more than true memory, it follows that divided attention should impair true recognition more than false recognition. Indeed, multiple studies showed that although true and false recognition both decline with divided attention, true recognition declines to a larger extent than false recognition (Dewhurst et al., 2007; Knott & Dewhurst, 2007; Seamon, Luo, & Gallo, 1998; Seamon et al., 2003).

Method In the MINERVA2 simulations for both the full and divided attention conditions, we randomly sampled five DRM lists from the combined list pool of Stadler et al. (1999) and Gallo and Roediger (2002) in each trial and run 1000 simulated trials. Because divided attention should result in poorer encoding than full attention, we used a lower learning rate in the divided attention condition (*L* = .5) than in the full attention condition (*L* = .7). The decision criteria were 1.3 and 1.5 for divided and full attention conditions, respectively.

Results The simulation results are displayed in Fig. 4, which is compared to the results in Knott and Dewhurst (2007; Experiment 1). As we can see, the simulated patterns are again highly comparable to the actual behavioral data. Specifically, both true recognition and false recognition were reduced by divided attention. However, true recognition was reduced more than false recognition, which is consistent with FTT' prediction.

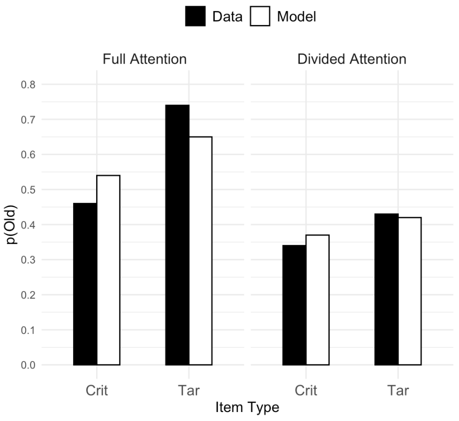

Figure 4: The simulated and actual recognition of Knott and Dewhurst (2007). Crit = critical lure. Tar = target.

Simulation 5: Effects of Retention Intervals

According to FTT, gist traces are more resilient to forgetting and interference than verbatim traces. Thus, people tend to rely more on gist traces than verbatim traces after longer retention intervals (Brainerd & Reyna, 2005). Because gist traces support false memory more than true memory, false memory should be more persistent than true memory over longer retention intervals. Indeed, Toglia, Neuschatz, & Goodwin (1999) and Seamon et al. (2002) both demonstrated that there was a shaper decline in true memory than false memory after a few weeks. Moreover, Thapar and McDermott (2001) showed that false recognition could even surpass true recognition after a 1-week delay.

Method In this simulation, we again simulated 1000 trials. In each trial, 6 DRM lists were randomly selected from the list pool of Roediger and McDermott (1995). To simulate the effect of retention interval on false recognition, we used different learning rates for shorter versus longer intervals. Because there should be more forgetting and interference with longer retention intervals, we used a lower learning rate (L = .2) for the delayed recognition test than for the immediate recognition test (L = .6). We also expect that participants should adopt a more liberal criterion in delayed tests since participants should rely more on familiarity rather than recollection compared to immediate tests. Accordingly, we used a lower decision criterion in delayed tests than in immediate tests (Cs = .6 vs. 1.2).

Results The simulation results are shown in Fig. 5, where we can see that both true recognition for targets and false recognition for critical lures decline between immediate and delayed tests, but there was a shaper declining slope for true recognition than for false recognition, as FTT predicts, suggesting that false recognition is more enduring than true recognition.

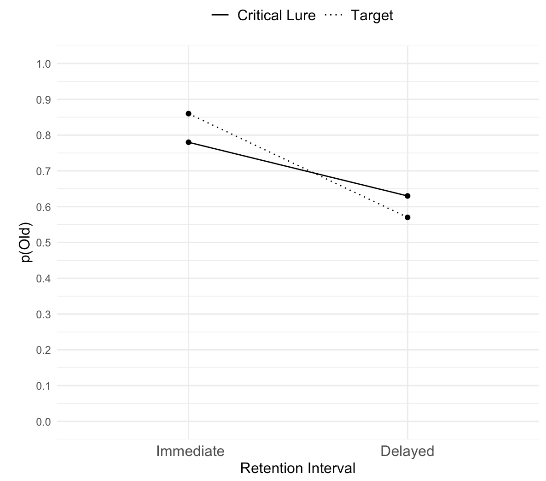

Figure 5: The simulated recognition of immediate versus delayed test.

Discussion

In the present paper, instead of using randomly generated vectors for words in the MINERVA2 model, we used word vectors derived from word2vec, which is a distributed semantic model that extracts word meaning from the word co-occurrence patterns and statistical redundancies in the natural language environment. Our results extend Arndt and Hirshman's (1998) work in showing that the MINERVA2 model provides a satisfactory account for true and false recognition in the DRM paradigm when combined with realistic semantic representations. Thus, the MINERVA2 model demonstrated robustness and flexibility in terms of accommodating real-life complexity in word representations.

Additionally, we emphasize that the integration of representational models and process models is a promising avenue and an obvious goal for the computational modeling of memory. This approach is not only more ecologically valid but also more computationally parsimonious. Regarding the latter, process models like MINERVA2 usually require separate assumptions or parameters to imitate the structure of human semantic memory with randomized representations. However, those assumptions or parameters will be unnecessary with the use of representational models (e.g., distributed semantic models), as the semantic information of the to-be-remembered items is readily capsulated in the representations themselves. In brief, we encourage future research to incorporate more realistic semantic representations instead of randomized representations into the process models.

References

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*(3), 371–391.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs.

context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Brainerd, C. J., Chang, M., & Bialer, D. M. (2020). From association to gist. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(11), 2106–2127.

Brainerd, C. J., Nakamura, K., Chang, M., & Bialer, D. M. (2019). Verbatim editing: A general model of recollection rejection. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *45*(10), 1776–1790.

Brainerd, C. J., & Reyna, V. F. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology*, *71*(2), 81–129.

Brainerd, C. J., & Reyna, V. F. (2005). The science of false memory. Oxford University Press.

Brainerd, C. J., Reyna, V. F., & Forrest, T. J. (2002). Are young children susceptible to the false–memory illusion? *Child Development*, *73*(5), 1363–1377.

Chang, M., & Brainerd, C. J. (2021). Semantic and phonological false memory: A review of theory and data. J*ournal of Memory and Language*, *119*, 104210.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

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

Dewhurst, S. A., Barry, C., Swannell, E. R., Holmes, S. J., & Bathurst, G. L. (2007). The effect of divided attention on false memory depends on how memory is tested. *Memory & Cognition*, 35(4), 660–667.

Gallo, D. A. (2006). Associative illusions of memory: Research on false memory for related events. Psychology Press.

Gallo, D. A., & Roediger, H. L. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. J*ournal of Memory and Language*, *47*(3), 469–497.

Gallo, D. A., & Roediger, H. L. (2003). The effects of associations and aging on illusory recollection. *Memory & Cognition*, *31*(7), 1036–1044.

Gatti, D., Rinaldi, L., Marelli, M., Mazzoni, G., & Vecchi, T. (2022). Decomposing the semantic processes underpinning veridical and false memories. *Journal of Experimental Psychology: General*, *151*(2), 363–389.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: *A discussion of common misconceptions. Perspectives on Psychological Science*, *14*(6), 1006–1033.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*(2), 96–101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Hutchison, K. A., & Balota, D. A. (2005). Decoupling semantic and associative information in false memories: Explorations with semantically ambiguous and unambiguous critical lures. *Journal of Memory and Language*, *52*(1), 1–28.

Jacoby, L. L. (1996). Dissociating automatic and consciously controlled effects of study/test compatibility. *Journal of Memory and Language*, *35*(1), 32–52.

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology*, *1*(3), 174-183.

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review*, *17*(5), 662–672.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*(4), 486–518.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2021). A continuous source reinstatement model of true and false recollection. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 75(1), 1-18.

Knott, L. M., & Dewhurst, S. A. (2007). The effects of divided attention at study and test on false recognition: A comparison of DRM and categorized lists. *Memory & Cognition*, *35*(8), 1954–1965.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), Advances in neural information processing systems. Curran Associates.

Reid, N. J., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language*, *129*, 104397.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*(3), 231–237.

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., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407.

Seamon, J. G., Goodkind, M. S., Dumey, A. D., Dick, E., Aufseeser, M. S., Strickland, S. E., Woulfin, J. R., & Fung, N. S. (2003). "If I didn't write it, why would I remember it?" Effects of encoding, attention, and practice on accurate and false memory. *Memory & Cognition*, *31*(3), 445–457.

Seamon, J. G., Luo, C. R., & Gallo, D. A. (1998). Creating false memories of words with or without recognition of list items: Evidence for nonconscious processes. *Psychological Science*, *9*(1), 20–26.

Seamon, J. G., Luo, C. R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, N. S., & Schwartz, M. A. (2002). Are false memories more difficult to forget than accurate memories?: The effect of retention interval on recall and recognition. *Memory & Cognition*, *30*(7), 1054–1064.

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

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), Experimental cognitive psychology and its applications (pp. 237-249). Washington, DC: American Psychological Association.

Thapar, A., & McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory & Cognition*, *29*(3), 424–432.

Thompson-Cannino, J., & Cotton, R. (2009). Picking cotton: Our memoir of injustice and redemption. St: Martin's Press.

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, *7*(2), 233–256.

Wilkomirski, B. (1997). Fragments: Memories of a wartime childhood. Schocken.