Incorporating Connotation of Meaning into Models of Semantic Representation: An Application in Text Corpus Analysis

Shane T. Mueller (smueller@ara.com)
Klein Associates Division
A. R. A. Inc.
1750 Commerce Center Boulevard North
Fairborn, OH 45434 USA

Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47404 USA

Abstract

Connotation of meaning is an important aspect of human semantic knowledge, and it cannot be captured in simple prototype representations of concepts. Yet models of human episodic memory typically rely on prototype representations, as do statistical techniques for extracting meaningful representations from text corpora (such as LSA). We will demonstrate how REM-II (a model of human episodic and semantic memory) allows connotation of meaning to be represented, and demonstrate that model can be develop and learn reasonable semantic representations by processing the Mindpixel project's 80,000-statement GAC corpus. The success of the model at developing meaningful and contextual representations from a text corpus provides a demonstration of the importance and utility of our assumptions.

Keywords: episodic memory; semantic memory; text corpus analysis

Connotation of meaning has been shown to be important in language learning (Corrigan, 2002), meaning disambiguation (e.g., Swinney, 1979) and even latent emotional content (e.g., Cato et al., 2004). As a rough guide to its prevalence in English, the Merriam-Webster's Collegiate Dictionary, 11th edition (2003) contains 165,000 entries with 225,000 definitions. Thus, there are approximately 1.36 meanings for each word, even though homonyms are given distinct entries and the dictionary is likely to contain large numbers of infrequent and specialized terms with only one definition.

Connotation of meaning describes the fact that the concepts we understand have multiple context-specific forms. If we consider linguistic concepts, extreme versions of connotation encompass homophony, homonymy and polysemy: single word forms sharing multiple distinct meanings. Words exhibiting these properties make connotation a challenge for automated systems attempting to understand language, because the context of the word must be considered in order to understand its proper meaning. But even subtler forms of connotation can be important, and this importance can transcend purely linguistic contexts. For example, consider how taxi cabs in different cities and countries differ substantially from one another. In Manhattan, a typical taxi is a yellow four-door sedan built by an American car company; in Mexico City, a typical taxi may be a small green compact vehicle. Thus, what we are calling connotation of meaning is an important aspect of our knowledge, for linguistic and non-linguistic stimuli and for extreme and subtle cases.

Yet many psychological models of knowledge and concept representations and fail to capture connotation. For example, prototype approaches typically consider information to be encoded as a set of features, and accumulate average or typical feature values across many individual events to form a composite, ignoring systematic variation and correlation among features. Such an approach is not unreasonable, because it allows a rich composite of central tendency to be formed from a set of noisy individuals. But if there are consistent patterns in the co-occurrence of features, a prototype will not be sensitive to them and will not be able to regenerate these distinct contextual representations. A prototype for the concept taxi would be a concept that never occurs in the world: a vehicle that is a mixture between a sedan and a compact car in a color somewhere between yellow and green. And consider adding rickshaws, airport shuttles, limousine services, and horse-drawn carriages to the prototype: the result is nearly impossible to imagine.

Despite the inadequacy of prototype techniques for representing knowledge, techniques for extracting meaningful representations from text corpora typically use prototypes. For example, HAL (Burgess & Lund, 1997) uses a graded word co-occurrence vector to represent semantic space; LSA (Landauer & Dumais, 1997) uses co-occurrences as input and projects this information onto a lower dimensional space using statistical optimization procedures similar to factor analysis. Likewise, the Topics model (Griffiths & Steyvers, 2004) uses a bayesian approach to place constraints on the statistical distribution taken by features, and as a byproduct generates features that are often interpretable. And recently, Jones and Mewhort (2007) demonstrated that order and meaning can be incorporated into a composite holographic trace using a convolution/correlation process. Of these, only Jones and Mewhort (2007) use a representation of knowledge that is not a simple prototype; instead they use a complex holographic representation in which information is distributed.

In order to move beyond a simple prototype knowledge representation, we propose that knowledge accumulates in the form of feature co-occurrences. Thus, if one considers all

experienced exemplars of a concept, one would determine for each pair of features how many times those features occurred together to form a composite trace. Such a representation maintains a set of conditional representations, and enables each distinct meaning to be maintained independently. We have implemented these notions in a computational model of human we describe next. Following this, we will demonstrate the utility of our assumptions by allowing the model to read text corpora and develop meaningful representations based on information in the text.

REM-II: A Bayesian Model of Episodic Memory Retrieval and Semantic Knowledge Formation

REM-II (Mueller & Shiffrin, 2006) is an extension of REM *Retrieving Effectively from Memory*, Shiffrin & Steyvers, 1997), a bayesian model of human episodic memory. REM-II was developed in to explain empirical phenomena in which polysemous words encoded with bias would activate earlier memories associated with one, but not both connotations of the probe word. A more precise mathematical description of the model is available in Mueller and Shiffrin (2006), but for the present demonstrations in corpus analysis, we will highlight three critical assumptions. First, rather than using a global optimization process to produce representations, we implement a psychological model of sensemaking (e.g., Klein, Moon, and Hoffman, 2006) that interprets events according to its knowledge and grows knowledge because of those events. The second assumption is related to a result of the corpus techniques described earlier: words that are similar to one another tend to appear in close proximity. We assume that the opposite relation holds as well: concepts that appear in the same context grow more similar because of this co-occurrence. And finally, we assume (as discussed before) that knowledge accrues as feature co-occurrences, enabling connotation and contextual meaning to be represented. The basic steps involved in allowing REM-II to interpret a sentence of text is shown in Figure 1.

In REM-II, an event consists of a set of concepts that occur at the same time and place. In the context of corpus analysis, we treat each individual sentence or statement as a distinct event. An episode is formed through a "sensemaking" process by which each event is interpreted through past knowledge, and is represented as a set of features that were present in the event. In contrast to this flat representation, semantic knowledge of a concept is maintained as a symmetric matrix that encodes the co-occurrence of features within individual events. Each row of that matrix keeps track of a prototype of a conditional representation of that concept, conditioned on the presence of each feature.

To encode a new episode, we assume that the proper semantic knowledge matrix is identified based on perceptual and contextual information. The model then samples additional features from the knowledge matrix to enhance and give meaning to the representation. Sampling is biased by

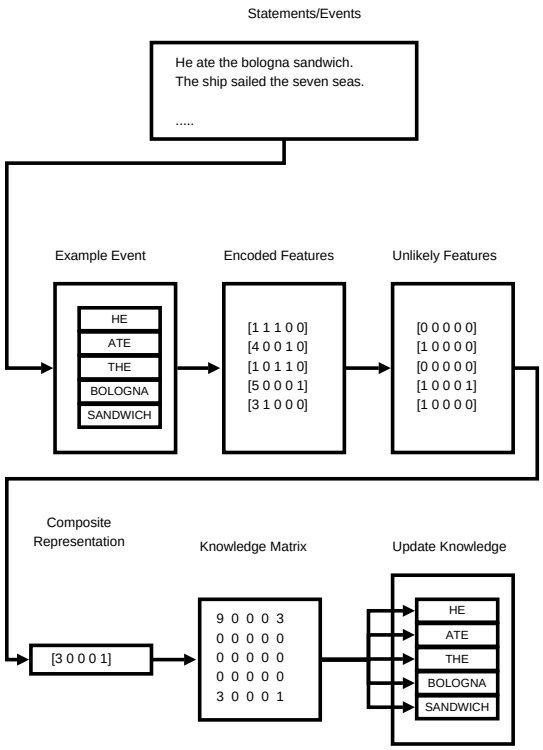

Figure 1: Basic steps performed by REM-II to process statements in language or events in the world. (1) the physical identity of the component objects in the event activate knowledge structures and (2) generates traces by sampling features. (3) These traces are compared to the base rate distribution to determine which are unlikely to have occurred by chance. (4) Then, a composite representation of the local semantic context is formed from these unlikely features, and (5) a co-occurrence matrix is formed from that composite representing the features that occurred together in the current context. (6) Finally, this composite matrix is added back into the semantic knowledge matrix for each word in the event.

the current semantic context, at first by sampling a feature from the current context, selecting that row in the knowledge matrix and sampling a feature from the selected row. We assume that greater study time would allow more features to be sampled, generating a richer representation of the concept.

In the original REM model, memory matches are computed by computing a likelihood ratio based on a probabilistic model of memory encoding. The model assumes that a features can appear in a memory trace either because they are were correctly encoded, or because an error was made. The distribution of errors is assumed to follow the base rate of features in the environment, and so for any memory probe, one can compute the probability that it "matches" an episodic trace by computing the likelihood that the trace arose from the memory structure associated with the probe. When events are encoded, we go through a similar process to determine which encoded features are important carriers of the unique infor-

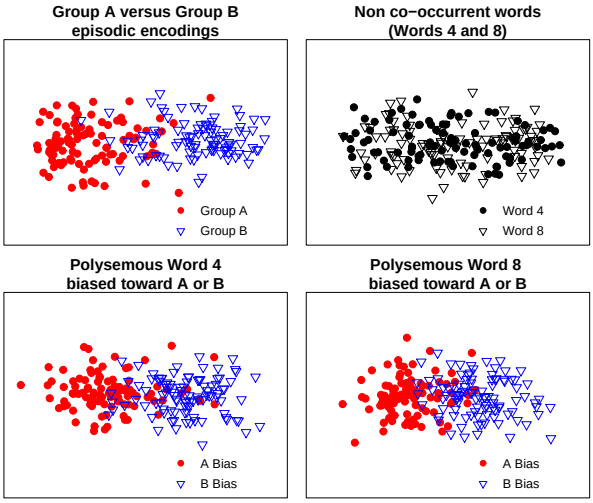

Figure 2: MDS solution for biased and unbiased episodes encoded from each meaning group and from the two polysemous words.

mation about the episode. For each encoded trace, we compare its distribution to the base rate distribution of features across the entire history of the model. Only those features with density greater than expected by chance are selected. A co-occurrence matrix is formed from the outer product of the index features, and this co-occurrence matrix is added back into the semantic knowledge matrix for each concept occurring in the episode.

Although this is a model of the interpretation of events and formation of knowledge from those events, we have found that it can go beyond modeling simple laboratory experimental situations, and be deployed on meaningful text to learn useful representations. In the remainder of the paper, we will describe several demonstrations in which the model was allowed to read a corpus of text and develop semantic representations based on the co-occurrence patterns in the text.

Application: Text Corpus Analysis

If the assumptions of REM-II are accurate, we should be able to present information to the model and have it grow representations that produce natural semantic spaces. We first tested some of the assumptions using a small hand-generated corpus. We then scaled the model to a large targeted corpus, and finally to a broad corpus of knowledge. Results from each demonstration are described below.

Demonstration 1: Small Polysemous Corpus

We began with a small toy corpus generated with simple probabilistic rules. The corpus contained eight distinct words with two sets of three words that tended to appear together, and two polysemous words that appeared with each set but not together. So, if *A*, *B* and *P* denote whether a word was from Set A, B, or a polysemous word, a typical automatically set of sentences might look like: $A_1A_3P_1A_2A_1P_1.B_1B_3P_2P_2B_1B_2P2$. The model made 5000 iterations through each of four sentence types (A,P_1,A,P_2,B,P_1,A,P_2), at which point we determined that the representations had converged to be highly similar within each meaning set, and the two polysemous words had also converged to nearly identical representations.

We were especially interested in whether the representations of the polysemous words would indeed keep the meanings associated with the distinct contexts separate, or whether the representation would simply converge to an average of the two contexts. To test this, we used probabilistic encoding process described earlier to generate biased and unbiased episodic traces from different words in this small corpus. To encode an unbiased representation, a row is initially sampled unconditionally from the base rate distribution, and a feature is sampled from that row, but for following samples row are chosen probabilistically from the he representation being built. We encoded 100 unbiased episodes from group A, group B, and the two polysemous words, and 100 biased episodes from the two polysemous words, biased by "A" or "B" contexts. Once encoded, we computed a distance matrix over the complete set of sampled episodes by calculating the root-mean-square deviation between episodes normalized to sum to 1.0. We then submitted this distance matrix to a single multi-dimensional scaling (MDS) solution using the `isoMDS` function of the R statistical computing language. We present the data from the global MDS solution in multiple panels of Figure 2 to assist visualization.

The upper left panel of Figure 2 shows that unbiased encoding of pure A or B words segregate in the space, with little overlap. At the same time, unbiased episodes encoded from the two polysemous words (upper right panel) cover the entire space and are indistinguishible from one another (even though they never appeared together). When episodes were encoded from the polysemous words biased by either A or B (lower panels) the resulting episodes clustered in the spaces corresponding to unbiased encodings of words from those two groups. Thus the two polysemous words which appeared in two distinct contexts retained the information separately, and appropriate versions of these traces could be extracted using a biased encoding process.

Demonstration 2: "Fly" Subset of the GAC Corpus

We next attempted to scale up the model to a larger naturally-occurring corpus. To increase the efficiency of the learning process, we replaced the sampling process used to generate an episodic trace in Demonstration 1 with the probabilistic computation of the expected distribution. This is simply a weighted sum of a normalized context vector and a normalized knowledge matrix. This was compared to the base rate distribution for features, and only unlikely features were selected, and so this remained a fairly similar process, but increased the efficiency of the process substantially.

We attempted to identify a text corpus which could provide fairly dense information, to reduce the processing requirements for this exploratory project. One of the better

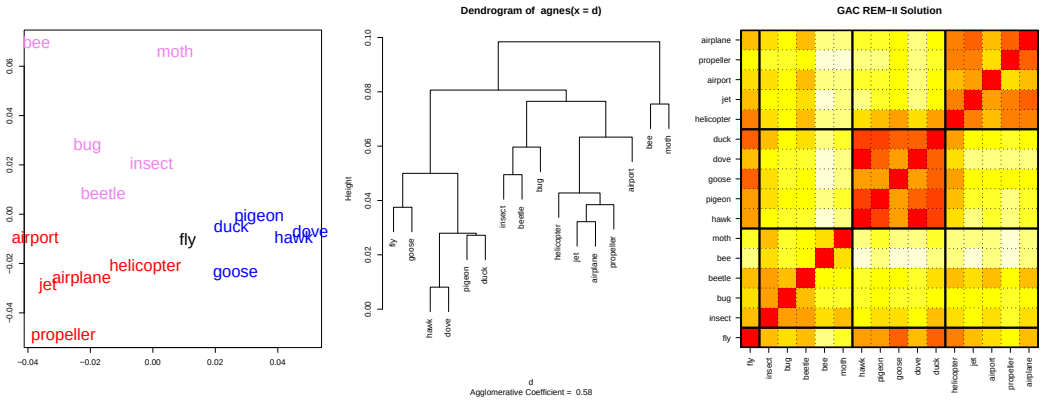

Figure 3: Results of Demonstration 2: An MDS solution, agglomerative hierarchical clustering tree, and visual depiction of similarity matrix for words related to three connotations of “fly”.

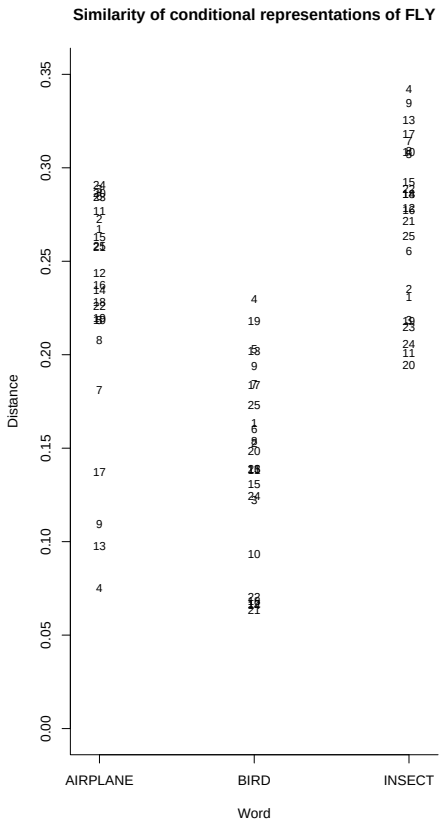

Figure 4: Dissimilarity of each conditional representation of “fly” to three key words: “airplane”, “bird”, and “insect”. Each conditional representation is indicated by the index of the feature used to form the conditional representation. Results demonstrate that the matrix representation segregates meanings related to each context, enabling connotation and polysemy to emerge.

sources we identified was a corpus produced by the Mindpixel project. The Mindpixel project was an internet-based collaborative project to generate verifiable statements about the world. Users submitted statements or questions about the world (e.g., “Is a dog is a mammal?” and other users would verify if the statement was correct. Each such statement was considered a “mindpixel”. The project began in the year 2000, and had putatively collected 1.4 million “mindpixels” by 2004, in a database called GAC (General Artificial Consciousness). Although the project appears to have been abandoned with the death of its founder in 2006, a database of 80,000 verified statements was released on the internet. We view these statements as a rich yet broad source of semantic content that could be used by our REM-II model to grow representations resembling human knowledge.

We have found that when the model is applied to typical text corpora, common function words which appear in many contexts end up developing representations that resemble the base rate distribution substantially, and so their information is ’filtered out’ by the likelihood comparison process. Thus, in order to further increase the speed of the algorithm, we performed some simple pre-processing to the GAC corpus, eliminating common function words and mapping distinct word forms onto the same base word according to the lemmas in the CELEX database. As a result of this preprocessing, the 80,000 statement corpus containing approximately 660,000 tokens and 29,000 unique words was reduced to 78,745 statements containing 269,000 tokens and 11,859 unique words.

Our initial target for corpus analysis was a subset of the GAC corpus: statements which contained either the word “fly” or one of its close associates (e.g., airplane, bird, insect, etc.). This resulted in 3992 statements containing a total of 15,583 tokens and 2907 unique words. To monitor progress, we selected 16 words in three groups describing three connotations of “fly”: the word fly, words related to insects, airplanes, and birds. We used 25 features as a basis for the representation. Reasonable representations for these 16 words developed fairly quickly, but the meaning groups continued

to develop over more than 100 consecutive readings of the text. We computed a similarity matrix across the 16 words, and show three methods for visualizing it in Figure 3. On the left, an MDS solution shows that the word "fly" is in the center of the space, and the three different meanings cluster together in three corners of the space. The center panel shows how an agglomerative hierarchical clustering solution tends to cluster the like words together. Finally, the rightmost panel shows the pairwise similarity between each word, with darker entries indicating greater similarity. These solutions compared favorably to the ones produced by LSA using the TASA corpus ($r = .584$), and the solution produced by LSA on the exact same GAC corpus ($r = .403$). In fact, the REM solution was more similar to both LSA solutions than they were to one another ($r = .202$).

These graphical visualizations show that REM-II placed "fly" and its semantic neighbors into a reasonable semantic space, with related subsets of words clustering together and "fly" being somewhat similar to all concepts. Yet such a phenomenon could occur even if a prototype representation was used. To determine whether the representation of "fly" contains conditional representations that segregate these different meanings, we examined each row of the matrix representation. Recall that each row (or column) of a co-occurrence matrix can be interpreted as a conditional representation, conditioned on the presence of a specific feature. For the 25 features in this demonstration, there were thus 25 conditional representations for "fly". We examined each of these in turn, comparing them to the composite representations for "airplane", "bird", and "insect" using a root-mean-square deviation over normalized vectors. This produced 25 distance scores for each comparison word, which are all shown in Figure 4, denoted by the index of each feature.

This analysis reveals several things. Similar to Figure 2, it shows that conditional representations of "fly" are on average closer to the word "bird" than "airplane" or "insect", and although several are close to "airplane", none are very close to "insect". Additionally, representations that were close to "airplane" tended to be farther from "bird" and "insect". Overall, the dissimilarity of conditional representations of "fly" to "airplane" was negatively correlated with the dissimilarities of "fly" to "bird" ($r = -.5$) and insect ($r = -.8$), whereas the dissimilarities of "fly" to "bird" was slightly positively correlated with those of "fly" to "insect" ($R = .17$). This indicates that several of the features (e.g., 4, 9, 13, and 17) in "fly" tended to encode an "airplane" connotation, whereas others (e.g., 10, 12, 21, 22) tended to encode "bird" or "insect" connotations. The features of "airplane" that had the greatest densities were 4,9,10, 13, 17, 18, and 22, and the features of "bird" that had the greatest densities were 10, 12, 14, 18, 21, and 22, which maps closely onto those conditional representations of "fly" that were similar to the each word. Interestingly, the two connotations of "fly" appear to map onto a natural/man-made distinction fairly nicely.

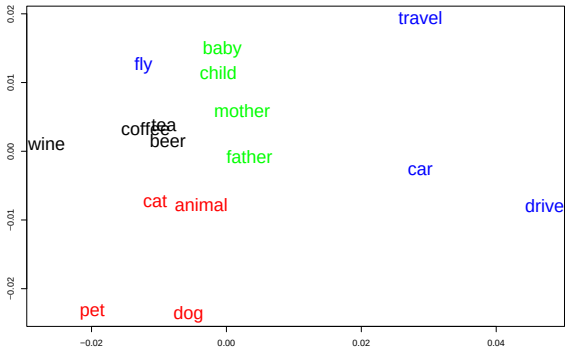

Figure 5: Multi-dimensional scaling solution for four clusters of four words, based on REM-II learning of the complete GAC corpus. Semantically similar words tend to cluster in similar regions of space.

Demonstration 3: Complete GAC Corpus

Finally, we wanted to demonstrate that the model could be used to learn representations of wider knowledge in the complete GAC corpus. In this demonstration, we used 40 features, and allowed the model to read the complete 80,000-statement database multiple times, randomizing the order of the statements between each pass, and monitoring the intermediate solutions.

To assess whether the representations of different words humans judge as similar grow similar to one another, we selected 16 high frequency words in four target areas to monitor as the representations grew. These included: pet, cat, dog, animal, child, baby, father, mother, travel, car, drive, fly, beer, tea, coffee, and wine. A multi-dimensional scaling solution for these target words is shown in Figure 5 after twelve passes through the database. Semantically similar words tended to cluster together, with the curious exception of the word "fly". This apparent anomaly was completely unrelated to its role in the earlier analysis.

Considering just the 16 target words, REM-II was able to produce between-word similarities that compared well with those produced by LSA on the large TASA corpus ($r = .439$), in ways similar to that produced by LSA on the same GAC corpus ($r = .459$). The between-word similarities from REM-II and LSA analyses of the GAC corpus were also correlated, but to a lesser extent ($r = .324$). The dissimilarity matrices for these three analyses are depicted visually in Figure 6.

Finally, because the analysis was completed for a larger corpus with approximately 29,000 words, we are able to generate similarity-based queries and evaluate their fitness qualitatively. To demonstrate, we present the closest ten representations to a variety of key probes in Table 1.

Discussion

In this paper, we have shown how a model of human memory can be deployed to grow the same types of representations

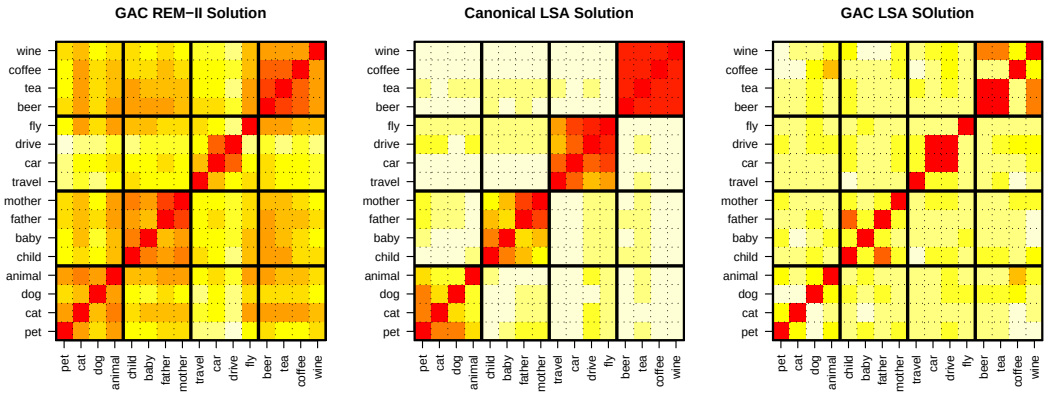

Figure 6: Depiction of dissimilarity matrices for sixteen target words, created using REM-II on the GAC corpus (left panel), LSA on the TASA corpus (middle panel), and LSA on the GAC corpus (rightmost panel)

Table 1: Ten most similar words to eight probes.

color	food	europe	man	fly	fire	car	earth	animal
color	food	europe	man	fly	fire	car	earth	animal
blue	eat	italy	woman	flap	match	drive	around	breath
violet	cereal	locate	physically	airplane	flame	motorcycle	close	worm
combination	rice	portugal	strong	air	touch	ride	spin	predator
red	regularly	rome	attractive	plane	start	driving	mars	usually
orange	restaurant	germany	virgin	balloon	wise	automobile	planet	human
primary	hamburger	belgium	average	lighter	term	form	sun	intelligent
yellow	mouth	music	naked	african	off	move	spherical	eat
combine	order	spain	crave	pop	hot	vehicle	rotate	curious
purple	usually	japan	reach	fan	lightbulb	consume	moon	cockroach

that statistical corpus analysis techniques can produce. The success demonstrated here shows that the psychological assumptions we based the model on are sufficient to develop rich knowledge representations, providing a clear demonstration of the conference theme: "CogSci in the Real World". By taking our psychological model out of the laboratory and allowing it to learn from natural artifacts, we are able to demonstrate the utility of the model, and at the same time create a tool that can be useful for automated processing and interpretation of text. By taking these psychological processes and representations seriously, we believe that new and more intelligent tools can be developed for a range of applications in knowledge management and understanding.

References

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12*, 1–34.

Corrigan, R. (2002). The acquisition of word connotations: asking 'What happened?' Journal of Child Language, 31: 381-398.

Cato, M. A., Crosson, B., Gkay, D., Soltysik, D., Wierenga, C., Gopinath, K., Himes, N., Belanger, H., Bauer R. M., Fischler I. S., Gonzalez-Rothi, L. & Briggs, R. W. (2004). Processing Words with Emotional Connotation: An fMRI Study of Time Course and Laterality in Rostral Frontal and Retrosplenial Cortices. *Journal of Cognitive Neuroscience, 16*, 167-177.

Griffiths, T. & Steyvers, M. (2004). Finding Scientific Topics. *Proceedinngs of the National Academy of Sciences, 101 (suppl. 1)*, 5228–5235.

Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*,1–37.

Klein, G. Moon, B., and Hoffman, R. R. (2006). Making Sense of Sensemaking 2: A Macrocognitive model. *IEEE Intelligent Systems, 21*, 88-92.

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*, 211–240.

Merriam-Webster's Collegiate Dictionary, 11th Ed. (2003). Springfield, MA: Merriam-Webster, Inc.

Mueller, S. T. & Shiffrin, R. M. (2006). REM-II: A Model of the developmental co-evolution of episodic memory and semantic knowledge. *Paper presented at the International Conference on Learning and Development (ICDL), Bloomington, IN, June, 2006*.

Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM–retrieving effectively from memory. *Psychonomic Bulletin and Review, 4*, 141–166.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645–659.