Inferring Contextual Semantics from Text using a Model of Human Episodic Memory and Conceptual Knowledge Formation

Shane T. Mueller, Ph.D.
Cognitive Science Group
Klein Associates Division, ARA Inc.
1750 Commerce Center Blvd.
Fairborn, OH 45324
smueller@ara.com

Abstract

Research on human memory has shown that connotation of meaning and polysemy are important for representing concepts. Connotation also has practical consequences in computational linguistics for disambiguating homographs and ambiguous meanings in written text. REM-II is a model of human episodic and semantic memory formation that enables such connotations to be represented based on contextual semantics. REM-II maintains separate distinct meanings for concepts in a composite feature-based trace. Although designed to model memory phenomena, it can be deployed to process text corpora and develop realistic semantic representations with connotation of meaning. Several demonstrations will be shown that illustrate the ability of REM-II to learn realistic semantic representations from the Mindpixel projects GAC-80K corpus, as well as a enable cross-linguistic analysis by training on parallel text corpora.

Paper read at the 4th *Midwest Computational Linguistics Consortium*, April 2007.

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 numerous applications in computational linguistics. In text processing applications, connotation is especially evident for polysemous words and homographs. But connotation of meaning is present in lesser forms for almost every concept. For example, a kitchen can be either a room in a house, or a place where food is prepared. These different aspects of the concept kitchen go beyond the notion of presence or absence of features, because they represent constellations of features that consistently co-occur when one version or another of kitchen is being considered. The connotation of a concept is often determined by the context it appears in. Consequently, connotation is an important aspect of contextual semantics.

Contextual semantics refers to the types of semantic information that can be inferred about words, objects, or concepts by the contexts the concepts appear in. Contextual semantics have been shown to be quite powerful ways to infer meaning. This probably occurs for two reasons. First, concepts that are similar are likely to occur together in similar contexts. But more importantly, concepts that appear together may be viewed as being similar, because common context may play an important role in defining semantic similarity. Thus, the human cognitive system may learn semantics by inferring that concepts that appear in the same context share common contextual features.

Although numerous systems allow latent semantic features to be inferred from the context words appear in (e.g., LSA, Landauer & Dumais

1997; HAL, Burgess & Lund, 1997), such systems typically rely on knowledge prototypes to represent encapsulated knowledge: average values or typical weights across a set of latent features. Yet prototypes fail to capture connotative or contextual meaning, and do not allow distinct semantic meanings to be recovered out of context. Consequently, most models of contextual semantics fail to allow contextual distinct representations.

1 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. Like several earlier systems, REM-II infers contextual semantic representations by examining the contexts words appear in. However, it uses a feature-based memory trace that allows separate connotations to be maintained by using a set of conditional representations over a set of latent features. Contextual meaning is then learned by accruing latent feature co-occurrences in experienced concepts. This model can be applied beyond experimental memory phenomena, and allows for meaningful representations to develop by processing text corpora. Several such applications of the model will be shown.

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 (or document). An episode is formed by encoding or interpreting each event through past knowledge, and is represented as a set of features that were present in the event. The semantic representations for an individual component of the event (or a word in a document) 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. As more and more episodes are experienced, contextual semantic representations emerge. These representations are both defined by the local semantic context, and conditional on the context, so that distinct but related meaning branches can be maintained separately.

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 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 determined by computing a likelihood ratio based on a probabilistic model of memory encoding. The model assumes that 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, a similar process determines which encoded features are important carriers of the unique information about the episode. For each encoded trace, its distribution is compared 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.

2 Demonstration: The GAC Corpus

The REM-II model was originally designed to model memory phenomena from laboratory experiments. However, we have adapted the model to enable naturalistic text corpora to be processed and contextual semantic representations to emerge. Next, we will describe the results of training the model on one such corpus.

We attempted to identify a text corpus which could provide fairly dense information, in order to reduce the processing requirements. One of the better 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, 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.

To demonstrate that the model can 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.

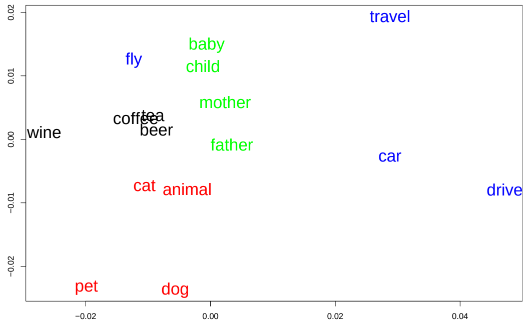

Figure 1: Multi-dimensional scaling solution for four clusters of four words, based on REM-II learning of the complete GAC corpus. Semantically similar words tend to cluster in similar regions of space.

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

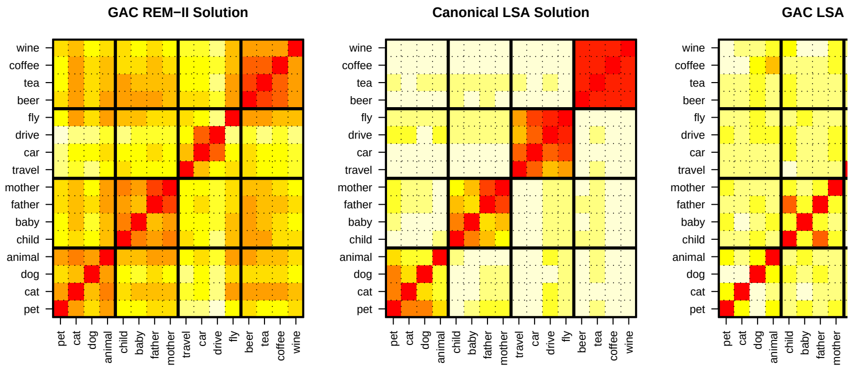

Figure 2: Depiction of dissimilarity matrices for sixteen target words, created using REM-II on the GAC corpus (left panel), LSA on the TASA corpus (middle panel), and LSA on the GAC corpus (rightmost panel)

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.

This demonstrations shows that the model is capable of inferring useful semantic representations based on contextual semantics. Past work (Mueller & Shiffrin, 2006, 2007) has shown that contextually-distinct meanings can emerge for words that appear in multiple contexts. Next, we will report an exploratory analysis that demonstrates the system's ability to infer common semantic representations from different languages.

3 Demonstration: Cross-Language Training

One useful application of automatic contextual semantic systems is the possibility of creating a cross-linguistic database such that words with corresponding meanings in different languages will appear in the same regions of semantic space. Similar techniques were demonstrated used by Littman et al. (1996). The technique has useful applications because automated monitoring of international media in different languages can be mapped into the same spaces.

3.1 Initial Exploratory Analysis

To investigate these issues further, we created a parallel corpus: English and Spanish translations of the book of Genesis. We formed a parallel corpus, in which the text from the corresponding verses in each language were copied together onto corresponding lines of a text file, so that the same verse in each language appeared together. The model treated each line as a common context, determining a local semantic meaning for that context and adding features from that context back into the individual representations of the words.

The goal of this simulation was to investigate ways in which a common language-general semantic space could be produced, while still allowing language-specific features to be used. Although this analysis can prove useful for cross-language monitoring, it highlights an important limitation of current contextual semantic models: integral features (such as language of origin) should be able to be incorporated into the representation, yet these features should be able to be ignored if need be. Thus, by addressing the problem of cross-language analysis, we will hopefully introduce the groundwork for handling a large number of other integral features that are important in linguistic applications. With a more sophisticated language parser, other integral lexical features could be introduced as well. For example, instead of the common practice of using word stems for verbs, a more intelligent natural language parser could enhance the stemmed token with integral features representing tense, number, etc. To incorporate integral language features, the encoding process examined each token and identified its language of origin by consulting a database. Two features were

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

used to determine language of origin, with the first feature indicating English, and the second Spanish.

The following simulations were performed using 20 features per word. The book of Genesis consisted of 1533 verses. The English translation of Genesis has 39788 words, consisting of 2501 distinct token types. The Spanish translation had 36463 words, with 3740 distinct tokens types. In all, the combined corpus contained 6051 distinct token types. This represents a relatively small corpus, and so strong contextual semantic representations do not to emerge.

In the first training simulation, the presence of the integral language features reduced match likelihoods for cross-language matches. For example, consider the top matches for two corresponding words (years and años), which are shown in Table 2. The top matches tend to be exclusively within the same language as the probe, and primarily include words with time and genealogical connotations (names, numbers, time periods, etc.). These occurred because large sections of Genesis are devoted to describing familial lines and ages of important historical figures. This pattern occurred for nearly all words in the corpus: the most similar targets to an English probe were almost exclusively English, and the most similar targets to a Spanish probe were almost exclusively Spanish.

So, when an integral feature encoding the language source is used, the words from the distinct languages fall into distinct areas of the semantic space. This is useful because it codes one important property of the words, but presents a problem because words having the same meaning should appear in the same areas of space. Thus, the question becomes how we can incorporate language source (or other integral features) into the knowledge structure, while still allowing words from multiple languages to map into the same semantic space. A solution to this can have implications for other types of integral features, such as other linguistic features (part of speech, verb tense, noun numerosity), or physical features like shape, sound, or color.

To address this issue, we have introduced the notion of attention into the likelihood match process. Attention can be used to determine which features are considered when computing a match likelihood. So, one might want to consider all features in some case (when trying to identify similar words in a specific language), but one may want to ignore some features in other cases (when attempting to map a document from some other language into a common semantic space).

To examine the contribution of attention, we performed similarity-based queries ignoring the language features. Now, when attention was limited to just the contextual semantic features, similarity in a language-general space was appeared. To illustrate this, we chose 20 English words from the corpus, across a range of frequencies. For most probe words, the most similar word (other than itself) was the Spanish translation of the English word.

This demonstrates how one can infer cross-language semantics without losing distinct linguistic categories. Such techniques require further development to enable other integral features to be incorporated, but with careful application of psycholinguistics and attention, we hope that such representations can be augmented in useful and important ways.

Table 2: Most similar word to two probes (*years* and *años*) are primarily within-language when language features are attended to. For words marked with a *, their corresponding translation also appears in Table 2. Words marked with a + are names, which are typically different for each language and appear rarely, typically in versus describing the age and lineage of a familial line.

English	Spanish
YEARS	AÑOS
*BEGAT	TENIA
HUNDRED	E
+TERAH	*HIJAS
SONS	CUANDO
SEVEN	*ENGENDRO
DAUGHTERS	+NOE
NINE	*DESPUES
DIED	SEM
+ARPHAXAD	ESTOS
+ENOCH	SEGUN
+NAHOR	*MURIO
EIGHT	+JAFET
DAYS	+VIVIO
NINETY	*TRES
AFTER	COSTUMBRE
SEVENTY	ADQUIRIDO
+SHEM	LUGAR
+LAMECH	GANADO
+METHUSELAH	BENDIJO
+JAPHETH	ENVIO
+SHALEM	MATARLE
LIVED	JARDIN
FIVE	PALABRAS
THREE	VINIERON

4 Summary and Conclusions

A number of statistical and machine learning techniques have been developed that infer contextual semantic representations via text co-occurrence. Despite their reliance on context to infer semantics, they have typically failed to account for the fact that different connotations of a concept depend upon the context a word appears in. Thus, they have capitalized on just one of the important aspects of contextual semantics.

This report describes new developments of REM-II, a model of episodic memory that allow richer contextual semantic representations to be developed. The model allows multiple conditional representations to be formed, and uses an encoding process by which important features from the local semantic context are used to infer semantic representations. Initially, this model was used to account for memory effects in laboratory experiments (Mueller & Shiffrin, 2006). More recently, we have shown how its principles can be used to infer useful contextual semantic representations from natural text (Mueller & Shiffrin, 2007). This report shows how the model can be augmented to incorporate non-contextual semantic features, using a cross-linguistic text base as a test case. These developments require the notions of attention to be incorporated into the matching process, to allow flexible likelihood-based memory trace matching based on different types of questions.

Although systems based on contextual semantics have proved very useful in the past, they have often ignored–and sometimes deemed irrelevant–semantic knowledge gained through other means. Such semantic information is important, but vector-space representations such as LSA have been unable to easily incorporate them. In contrast, REM-II capitalizes on context to develop useful semantic representations, but can incorporate other types of information as well. We hope that by considering other aspects of semantic information, further progress in natural language processing technologies can be made.

References

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

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.

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

Griffiths, T. & Steyvers, M. 2004. Finding Scientific Topics. *Proceedings 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.

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.

Littman, M. L., Dumais, S. T. , & Landauer, T. K.. 1996. Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing In Grefenstette, G., editor, Cross Language Information Retrieval. Kluwer.

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.

Table 3: Twenty English Probes, and the rank similarity of the corresponding spanish word when the language features were unattended.

	Absolute	Rank	Spanish	Target	Target Rank:
English	Frequency	Frequency	Target	Rank	Spanish Words
GOD	230	29	DIOS	2	1
FATHER	169	38	PADRE	2	1
SONS	158	41	HIJOS	2	1
JOSEPH	138	48	JOSE	2	1
EARTH	121	52	TIERRA	4	2
YEARS	113	58	ANOS	2	1
NAME	102	66	NOMBRE	2	1
WIFE	101	67	MUJER	2	1
HOUSE	86	77	CASA	2	1
BROTHER	81	84	HERMANO	2	1
PHARAOH	80	85	FARAON	2	1
BRETHREN	80	87	HERMANOS	7	4
EGYPT	77	90	EGIPTO	2	1
BEGAT	67	101	ENGENDRO	2	1
MEN	64	103	HOMBRES	2	1
DAUGHTERS	63	107	HIJAS	2	1
DAYS	63	106	DIAS	2	1
SISTER	22	225	HERMANA	8	1
GARDEN	14	318	JARDIN	2	1
SERPENT	6	609	SERPIENTE	2	1

In the Spanish text, numbers were typically represented as numeral sequences (e.g., 700), whereas in the English translation, numbers were written out (e.g., seven hundred). This accounts for some asymmetry between language probes.