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

Permalink
https://escholarship.org/uc/item/1cc8p34v

Journal

ISSN
1069-7977

Authors
Mueller, Shane T.
Shiffrin, Richard M.

Publication Date
2007

Peer reviewed
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 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 feature 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 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.
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 applications is the GAC corpus.
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.
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


