Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition

Article *in* Topics in Cognitive Science · September 2010

DOI: 10.1111/j.1756-8765.2010.01117.x

Danielle McNamara
Arizona State University
441 PUBLICATIONS **15,862** CITATIONS

TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2010) 1–15
Copyright © 2010 Cognitive Science Society, Inc. All rights reserved.
ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2010.01117.x

Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition

Danielle S. McNamara

Department of Psychology, University of Memphis

Abstract

Over the past two decades, researchers have made great advances in the area of computational methods for extracting meaning from text. This research has to a large extent been spurred by the development of latent semantic analysis (LSA), a method for extracting and representing the meaning of words using statistical computations applied to large corpora of text. Since the advent of LSA, researchers have developed and tested alternative statistical methods designed to detect and analyze meaning in text corpora. This research exemplifies how statistical models of semantics play an important role in our understanding of cognition and contribute to the field of cognitive science. Importantly, these models afford large-scale representations of human knowledge and allow researchers to explore various questions regarding knowledge, discourse processing, text comprehension, and language. This topic includes the latest progress by the leading researchers in the endeavor to go beyond LSA.

Keywords: Sematic models; Computational techniques; Meaning extraction; Cognition; Memory; Embodiment; Latent representations; LSA

1. Introduction

One method of scientifically investigating human cognition is to examine what we write. This endeavor is facilitated by current technologies that allow scientists to conduct discourse analyses on extremely large scales. Statistical models are applied to large linguistic corpora (i.e., collections of written text) to extract (what we think is) the meaning of the text and, moreover, to enhance our understanding of how the human mind works. A fundamental assumption among this collection of articles is that human cognition manifests itself in our

Correspondence should be sent to Danielle S. McNamara, Department of Psychology/Institute for Intelligent Systems, University of Memphis, 202 Psychology Building Memphis, TN. E-mail: dsmcnamara1@gmail.com

writings. Large text corpora combined with computational techniques for analyzing these corpora allow scientists to extract meaning from text and, by consequence, to explore various aspects of the human mind and culture that manifest in text. To a large extent, this volume of *topiCS* is dedicated to showing the value of using corpus analytical techniques to understand cognition.

A well known statistical method of extracting meaning from text is latent semantic analysis (LSA). In Landauer, McNamara, Dennis, and Kintsch (2007), we provided basic information about LSA as well as a number of examples of its use in furthering our understanding of cognition and in contributing to applied technologies. LSA was a groundbreaking method in which word meanings are extracted by determining the company the word keeps across large corpora of texts. Collections of word meanings then comprise sentence, paragraph, and document meanings. More specifically, given a large corpus of text with millions of words and thousands of documents, a matrix is created that indicates the context in which each word occurs. The context of a word is the *document* that it occurs in, which may be the sentence, paragraph, or entire text. Each word or term is defined in terms of a vector (i.e., one dimensional array of values) of the documents in which it occurs. This is a sparse matrix because most terms occur in few documents and it is a large matrix because there are many terms across many documents. Thus, the matrix is reduced to discover its latent properties. In LSA, it is reduced by applying singular value decomposition (SVD) and then truncating the number of dimensions to include hundreds rather than thousands of dimensions in the matrix. This process creates an LSA space that is multidimensional, where each term is located in this multidimensional space. Similarity metrics such as the cosine between words, or collections of words, are indicative of how related they are. A higher cosine indicates that the words are more related. By uncovering these latent relations between words, the meanings of the words, sentences, and texts can be derived.

Latent semantic analysis has been extremely successful in helping researchers understand a wide range of theoretical issues concerning meaning and also in scaling up theoretical insights to applied tasks (see e.g., Landauer et al., 2007; Louwerse, 2010, this volume). However, LSA alone is not the focus of this issue of *topiCS*. Here, we focus on going beyond LSA. In this topic, the featured articles describe new methods of using LSA as well as other mathematical and computational approaches that are comparable and sometimes exceed the capabilities of LSA. Our objective is not to reinforce the multiple uses of LSA, but more so to show how current research goes beyond LSA.

2. Statistical semantic models

Statistical models of semantics have been a focus of research for over a half century (e.g., Osgood, Suci, & Tannenbaum, 1957; Smith, Shoben, & Rips, 1974). However, the availability of technologies that could handle large corpora and high-dimensional spaces (e.g., Salton, Wong, & Yang, 1975) helped to change the field of semantics. These technological advances helped to spur LSA, which has been largely responsible for the growth in interest in and use of statistical models of semantics over the last two decades. LSA was immensely successful both as a theoretical model (e.g., Landauer, 2007; Landauer & Dumais, 1997) and as a practical tool (e.g., Landauer et al., 2007). Albeit successful and effective, a good deal of research has been devoted to improving LSA and to developing and improving semantic models in general. The majority of the articles in this volume are concerned with this objective. Over the past few decades, numerous semantic models have been developed and thus there is a large scope of models available that are capable of (more or less) successfully extracting meaning from text.

Riordan and Jones (2010, this volume) describe nine statistical models of semantics, which they refer to as *distributional* models:

- LSA (Landauer & Dumais, 1997)
- Probabilistic Topic Model (Steyvers, Chemudugunta, & Smyth, 2010, this volume; Steyvers & Griffiths, 2007)
- Correlated Occurrence Analog to Lexical Semantics (COALS; Rohde, Gonnerman, & Plaut, 2005)
- Contextual Similarity Log-Likelihood (CS-LL; Dunning, 1993)
- Contextual Similarity Log-Odds (CS-LO); Lowe, 2001)
- Hyperspace Analog to Language (HAL; Lund & Burgess, 1996)
- High Dimensional Explorer (HIDEx; Shaoul & Westbury, 2006)
- Positive Pointwise Mutual Information (PosPMI; Church & Hanks, 1990)
- Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007)

Stone, Dennis, and Kwantes (2010, this volume) describe three additional models:

- Vectorspace Model (Salton et al., 1975)
- Sparse Nonnegative Matrix Factorization (SpNMF: Xu, Liu, & Gong, 2003)
- Constructed Semantics Model (CSM; Kwantes, 2005)

In addition, Howard, Shankar, and Jagadisan (2010, this volume) describe their transformation of an episodic memory model temporal context model (TCM) into a new approach to semantic modeling:

- Predictive Temporal Context Model (pTCM)

One means of subdividing semantic models is into two categories, *context word* and *context region* (see e.g., Riordan & Jones, 2010, this volume). These two approaches differ in terms of the matrix that comprises the initial representation. Context *word* models use a word-by-word (term-by-term) matrix that is defined using a moving window approach. Essentially, the document size is reduced and defined only by words surrounding the target word. Words' co-occurrences within a defined window (e.g., two or more consecutive words) comprise the matrix. These models differ in the size of the window and whether the window considers both prior and subsequent words within the window. In addition, the words' co-occurrences are differentially weighted (e.g., by multiplying the co-occurrence matrix by another matrix that includes weights). They may be weighted by distance from the target words (HAL, COALS), frequency within the corpus (HIDEx), or their conditionalized co-occurrences (COALS, CS-LL, CS-LO, PosPMI). BEAGLE is also a context word model, but it incorporates convolution (e.g., Murdock, 1992) as a means to compress n-gram information, which captures both syntactic (word order) and semantic (contextual) information about each word.

Context *region* models, including the Vectorspace model, LSA, Topic, SpNMF, CSM, and pTCM, use a word-by-context (or term-by-document) matrix. In models such as the Vectorspace model, LSA, and SpNMF, the context of a word is the *document* in which it occurs; as described earlier, the document may be the sentence, paragraph, or entire text. In the Topic model (see Steyvers et al., 2010, this volume), the context is defined by latent topics identified in the corpus. Each document is represented by multiple topics as a probability distribution, and each topic is represented by a probability distribution of words. LSA and SpNMF are much like the Vectorspace model in that they both weight the vector representations of the terms by multiplying the vector by the log of the terms' frequency within the document and then dividing by the inverse document frequency (IDF)—these two techniques ensure that words that are important to a document are heavily weighted but the impact is reduced for words that are frequent across documents (and thus contain less discriminatory information). LSA departed from the Vectorspace model by its use of SVD to reduce the matrices to their latent dimensions. SpNMF differs from other models in that it is constrained to contain only nonnegative values.

The CSM model (Kwantes, 2005) differs from other context region models by using retrieval mechanisms in lieu of dimension reduction techniques. The retrieval mechanisms in CSM are based on the MINERVA 2 model of episodic memory (Hintzman, 1984). Accordingly, memories for events are feature-based representations stored as individual traces, or episodes. Each event or item is stored as a vector of values associated with features, which in turn have a probability value or learning rate. Retrieval depends on similarity-based matching rules and prior probabilities. Semantic similarity in CSM is a function of resonance between a probe vector and the context vectors. The context vectors with the stronger resonance to the probe contribute more strongly to the ultimate outcome. Different from other distributional models, CSM is an incremental learning model because it updates the representation as the sequence of words is presented.

The pTCM introduced by Howard et al. (2010, this volume) similarly updates the representation incrementally. pTCM is an extension of TCM, a model of episodic memory (Howard & Kahana, 2002). TCM is a distributed memory model that specifies the effects of temporal context on episodic memory (usually memory for words embedded in lists). Rather than using co-occurrences in memory to explain memory phenomena such as primacy and recency, TCM provides an account that is based on mechanisms related to contextual drift and contextual retrieval. Thus, TCM has similarities to models such as LSA due to the overlapping assumptions regarding the effects of context on memory and learning. pTCM models semantic memory by using the current state of context (i.e., a weighted sum of the semantic representations of recently presented items) as a cue to generate predictions about what word will be presented next. The semantic representation of a word is formed by averaging all of the *predictions* that precede its presentation. The model forms these representations incrementally over text rather than using the entire text representation as do most semantic memory models (and thus the time to run the model is somewhat cumbersome). Howard et al. show that the pTCM model performs comparably to LSA in the task of identifying semantically related words. Although a disadvantage stems from the time to run the model, its clear advantage is that it is based on a model of episodic memory; thus, this research shows that similar architectures may be used for both episodic memory and semantic representations in memory (see also the BEAGLE model; Jones & Mewhort, 2007).

The *deep generative* model described by Hinton and Salakhutdinov (2010, this volume) is also iteratively trained and inspired by cognitive models, but it is quite different from other semantic models described in this volume. Hinton and Salakhutdinov describe and test a nonlinear, multilayer model that yields binary codes that can serve to describe documents' meaning. The lower layers of the model are hidden layers with distributed representations that are iteratively trained using a *Restricted Boltzman Machine* (Hinton, Osindero, & Teh, 2006). After the features of one layer are learned, those vectors are used as data to train the next hidden layer. This process of iterative learning yields an *encoder* network that converts the word-count vectors to compact codes. Then the process is reversed such that the Restricted Boltzman Machine acts as a *decoder* network that reconstructs the word-count vectors from the compact code vectors. These are combined to produce a multilayer *auto-encoder* network. Backpropagation is used to iteratively improve the solution. This learning process is relatively slow. By contrast, after learning, retrieving solutions is rapid, much more rapid than techniques such as LSA. This difference would be particularly noticeable with large data sets. In addition, the new data can be added to the network iteratively, so the system can continue to *learn*. Thus, this type of approach would be particularly suitable for handling extremely large corpora such as the Wikipedia. The initial training would be very slow, but thereafter, new data could be added, and retrieval based on *meaning* would be extremely fast.

3. Comparing semantic models

In the quest to improve computational models of semantics, one current research question regards which models more effectively capture meaning and cognition. This topic includes two studies that compared models in terms of their ability to account for cognitive phenomenon.

Stone et al. (2010, this volume) compared six models' ability to judge similarity between paragraphs (i.e., LSA, SpNMF, Topic, Topic-JS, Vectorspace, and CSM). They found that all of the models fared relatively dismally in comparison to a word overlap model that used the cosine between term frequency vectors. That is, the model that used the nonlatent, overt representation of the text fared best in simulating humans' judgments of paragraph similarity. The latent models, however, showed improved performance with when smaller, topic-focused corpora were used to train the models (see also Shapiro & McNamara, 2000).

Constraining the knowledge space helps to avoid contamination from multiple word meanings across contexts. For example, whereas humans can quickly discern the meaning of *bank* in the context of finances, most statistical models of semantics will retrieve all senses of words (Kintsch & Mangalath, 2010, this volume). Thus, more focused, topic-specific spaces avoid the problem of retrieving multiple senses of words that are likely to be irrelevant.

Riordan and Jones (2010, this volume) compared models' ability to capture semantic clusters in comparison to humans' judgments of concept features. They found that among the nine models compared, a context word model, COALS, was best at capturing the semantic clusters. The model's success was attributed to how it conditionalizes the co-occurrences of words. COALS is based on HAL, but it reduces the influence of high-frequency words such as function words (e.g., the, a, that) to reduce the impact of chance co-occurrence. It does so by calculating the conditional rate of co-occurrence (i.e., the frequency of co-occurrence in comparison to its base rate co-occurrence with other words). This is achieved in COALS using a correlation matrix in which the negative values are discarded and the positive values square rooted (to increase their weight). However, this explanation for COALS relative success may not provide an entirely convincing explanation because other models also control for chance co-occurrence of words and virtually all models control for highly frequent words. Thus, why this model faired so well in terms of detecting semantic clusters remains a question (but see also Rohde et al., 2005).

The success of COALS in Riordan and Jones (2010, this volume) supports the notion that the meaning of words is in the company they keep (Firth, 1957)—but further implies that the company only includes *close neighbors who are not friends with everyone else*. Hence, local word co-occurrence (corrected for chance) goes a long distance in extracting meaning. Likewise, Stone et al. (2010, this volume) found that the nonlatent word overlap model faired far better than the six statistical models. One important issue regards whether meaning need be extracted using a latent or second-order representation or by contrast whether that meaning can be extracted with equal success from the words alone.

4. Improving semantic models

Comparing semantic models improves our understanding of which aspects of the models contribute most to successful meaning extraction. Another approach to improving methods of extracting meaning from text is to augment statistical models with other techniques. Steyvers, Chemudugunta, and Smyth (2010, this volume) do so by combining the Topic model with human judgments. They propose that models of the learning process can be enhanced by using corpus techniques to represent background knowledge of the learner. As such, by combining information from corpora and human data, statistically derived relationships can complement human-defined concepts by filling in missing information, and *vice versa*, the combination of the techniques may indicate the relative importance of preexisting and new knowledge extracted from a text.

Another means of improvement regards the analyses that are conducted with their output and how the results are interpreted. Louwerse (2010, this volume) uses multidimensional scaling (MDS) to extract the representation yielded by the LSA analysis, reducing it from hundreds of dimensions to a two-dimensional visualization. One advantage of MDS (and other similar methods) is that it provides an indication of the distance between concepts. For Louwerse, this technique affords the visualization of spatial and temporal relations present in the text corpora. Similarly, Riordan and Jones (2010, this volume) use MDS to convey the nature of the clusters of concepts. The distance between clusters represents similarity of the clusters; the height represents the internal consistency of the cluster; and the area indicates the number of terms in the cluster. Thus, MDS can provide additional information as well as facilitate interpretation of underlying relationships in the data. Other researchers have also used this technique. For example, O'Rourke and Calvo (2009) have incorporated the use of MDS in combination with LSA to examine the relationships between paragraphs in essay writing. MDS essentially allows the researcher to uncover the clusters of related concepts and the distances between text sections, much like principal component and factor analyses.

5. Embodiment

One important and hot topic in computational semantics regards the potential importance of perceptual simulations and embodiment in cognitive processing. It is clear to most that our experiences are constrained by the world and how our body fits into that world. By consequence, cognition is shaped by our experiences in the material world. Similarly, it is clear that the mind often represents the world, and our experiences in that world, in some sort of analog form, such as images. If that is the case, some argue that modeling human cognition using mathematically based or symbolic representations is futile because these models cannot capture a fundamental aspect of cognition, embodiment (e.g., Barsalou, 1999; De Vega, Glenberg, & Graesser, 2008; Glenberg, 1997). Moreover, some argue that meaning as it relates to cognitive processes cannot be extracted from representations such as text because text, being verbal and symbolic, cannot provide a complete picture of human cognition. One potential value of this extremist view is that it has served to fuel the debate. In addition, some researchers seek elegant models or attempt to see how far one can go with a simple model. Pushing the limits of simple models can reveal a good deal about the phenomena. Nonetheless, most recognize that there is value to both viewpoints (e.g., Paivio & Sadoski, in press). Most researchers and theorists recognize that cognition is comprised of symbolic, abstract, and verbal thought and reasoning, in addition to concrete and embodied representations. The notion that we have only one form or the other to represent meaning in the mind is, frankly, absurd.

Two articles in this topic address this issue (Louwerse, 2010, this volume; Riordan & Jones, 2010, this volume). Riordan and Jones (2010, this volume) compare symbolic, distributional models (e.g., LSA) and feature-based models (that rely on human judgments of features) in their ability to represent semantic clusters. Feature-based models have been argued to more aptly represent perception and action in cognition than do purely symbolic, distributional models such as LSA because they represent words' meanings in terms of their descriptive features (e.g., McRae, Cree, Seidenberg, & McNorgan, 2005; see Riordan &

Jones, 2010, this volume, for additional citations). These features are assumed to be closer to sensorimotor experiences, which in turn are assumed to comprise a fundamental aspect of words' meanings (e.g., Barsalou, 2003).

Riordan and Jones (2010, this volume) compare feature-based models to the nine distributional models listed earlier (i.e., LSA, Topic, COALS, CS-LL, CS-LO, HAL, HIDEx, PosPMI, and BEAGLE). They examine the ability of these nine models to extract semantic clusters in comparison to human generated norms on word features. In the first two studies, the models were trained using the TASA corpus and then compared on their ability to cluster words according to semantic class labels used in WordNet. Their performance was compared to McRae feature norms (McRae et al., 2005) on concrete nouns in the first study and Vinson and Vigliocco (2008) feature norms on nouns and verbs in the second study. In the third study, they compared models using semantic classes from the MacArthur-Bates Communicative Development Inventories (Fenson et al., 1994) and trained the models using the CHILDES database (MacWhinney, 2000) of caregiver speech (adults' utterances to children, 12–48 months).

Across the three studies, Riordan and Jones (2010, this volume) found that several models rivaled the human-based feature norms. However, across datasets, COALS was the most consistently found to be comparable to the feature-based norms in reproducing semantic classes. However, they also found that the distributional and featural information were not redundant. The distributional models tended to pick up on actions, functions, and situations, and less about perceptual attributes such as color or texture. Nonetheless, the distributional models use of linguistic cues in the language rivaled performance by feature norms (produced by humans). One implication of this research is that a model that combines human derived norms within a distributional approach may be more successful (see Steyvers et al., 2010, this volume). On the whole this research indicates that the statistical models are able to pick up on information associated with embodied thought, but not in the same way as do humans.

Louwerse (2010, this volume) also argues that both aspects of processing (i.e., symbolic and embodied) are important to describing human cognition. He refers to this as the Symbol Interdependency hypothesis, which supposes that language and language comprehension depend on interdependencies of abstract linguistic symbols as well as the references these symbols make to perception and action. Furthermore, he provides evidence that perceptual and modal aspects of cognition can be extracted from text corpora, which has been considered purely symbolic by the embodied theorists. The underlying assumption of Louwerse's argument (and this issue of *topiCS*) is that how we think manifests itself in the language we use. As such, the effects of perception and action, or embodied thought, on cognition can be extracted from language.

Louwerse (2010, this volume) shows that a wide array of results that have been used to support the embodied perspective can be replicated using techniques such as LSA. For example, he demonstrated that word pairs from the same sensory modality (motoric, smell, sound, taste, touch, and vision) have higher LSA cosines than do words from different modalities. In addition, concepts have a stronger relationship according to LSA cosine values to their properties or features than to properties descriptive of other concepts. Thus, like

Riordan and Jones (2010, this volume), Louwerse shows that features can be detected using computational models such as LSA, even if those features might be considered embodied and thus beyond computation by many researchers.

6. Where is meaning in text?

We assume that there is meaning in text and that meaning resides in the words, sentences, paragraphs, and so on. However, one question is where the meaning of the text resides. Can the full scope of meaning be extracted solely from the words and their co-occurrences (i.e., the company they keep) or is more context and information needed, such as syntax or human-derived data?

Most if not all text comprehension models assume that comprehension occurs at various levels that together produce the reader's mental representation (see e.g., Kintsch, 1998; McNamara & Magliano, 2009). For example, the Construction-Integration model assumes at least three levels of representation, the surface code (words and syntax), the propositional textbase (deeper meaning of the text), and the situation model (a combination of the text with prior knowledge). Because readers' comprehension is multidimensional, a more complete picture of it is provided when comprehension is assessed in various ways and at multiple levels (e.g., using both text-based and situation model questions; McNamara & Kintsch, 1996).

One cause for multiple levels of comprehension can be attributed to cognitive processing mechanisms. However, another is the signal itself. Language is comprised of multiple levels of information—it is multidimensional. If language comprises different levels of meaning, then statistical models that seek to extract meaning from text should also assume that the meaning should be extracted based on these levels. For example, Jones, Kintsch, and Mewhort (2006) demonstrated that including both semantic and syntactic (word order) information improves the ability of semantic models to account for a wider range of cognitive phenomena (see also, Dennis, 2005; Dennis & Kintsch, 2008).

Similarly, different aspects of a representation can be emphasized by varying parameters within the algorithms. For example, McNamara, Cai, and Louwerse (2007) evaluated variations of the LSA algorithm to examine whether the performance of LSA could be improved by varying two factors: emphasis on high- versus low-frequency words, and similarity strictness. Overall, the study indicated that different algorithms may be more apt to detect differences in meaning depending on the level of analysis. Thus, different algorithms may be more or less appropriate and effective depending on the cognitive processes that are targeted in the particular study. Indeed, this is an underlying assumption of Coh-Metrix (Graesser & McNamara, 2010, this volume; McNamara, Louwerse, McCarthy, & Graesser, 2010). Coh-Metrix provides information about various levels of language to support investigations of how and when these levels come into play in various situations. Coh-Metrix provides indices on words (e.g., frequency, concreteness, homonymy, polysemy), sentences (e.g., length, noun phrase density, number of words before the main verb), and cohesion (lexical diversity, referential cohesion, semantic cohesion). Cohesion is the level of connectivity in a text—the degree to which clauses, sentences, ideas, and larger sections of a text are explicitly tied together. If text is conceptualized as a network of nodes and links, then the nodes would represent the words (or the parts of words) of the text. The words of a text can be predominately abstract or concrete, familiar or unfamiliar, ambiguous or unambiguous. These are characteristics of the words that result from and are thus signals for the meaning of a text.

Indeed, articles in this topic support the notion that a good deal of a text's meaning can be detected just on the basis of the words in the text. For example, Stone et al.'s (2010, this volume) findings indicate that meaning can be extracted from the nonlatent information available in large text corpora (see also Recchia & Jones, 2009). LSA was a groundbreaking technique because it demonstrated that the text could be reduced to fundamental components comprising a latent representation, which contained the essentials of the text's semantics. However, dimension reduction is not always crucial to successful meaning extraction. Along these lines, Louwerse (2010, this volume) argues that it is not computational models such as LSA that, first and foremost, afford the ability to extract meaning from text, but rather this ability emerges from the organization of language itself.

Cohesion is one aspect of that organization and, in particular, connectivity. Cohesion is the glue that allows the ideas to stick together. At the surface level, words units serve as a glue to connect phonemes and morphemes, and syntax serves as glue to combine the words into a meaningful unit. At the textbase level, verbs serve as glue to connect ideas together to form meaningful clauses. Overlapping words and concepts among sentences serve to tie the ideas together. Likewise, at the situation model level, connectives serve as signals of the relationships between ideas and between the larger concepts that are being expressed across the text. Further, rhetoric, pragmatics, and global cohesion cues tie the text or discourse into a meaningful unit of discourse. These cohesion cues are important to comprehension because when those cues are not present in the text, then the reader must use reasoning and prior knowledge or retrieve prior text to infer the missing connections (e.g., McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007). It is the combined information from the words, sentences, and relations that afford the most complete picture of language. Thus, picking up on the multidimensionality of text and its deeper meaning likely depends on assessing the various dimensions of how those words are connected, beyond their particular characteristics, proximity, and co-occurrence (latent or nonlatent).

If meaning is present at multiple levels, then a more complete statistical model of semantics would benefit by taking into account multiple levels of processing. Kintsch and Mangalath (2010, this volume) do just that. They present a model that makes use of both context-word and context-region matrices as well as syntactic information. These sources of information are combined within a modified version of the Construction-Integration model of comprehension (CI-II). The word-document matrix generated using the Topic Model provides relational, gist information (called the *gist trace*), whereas a word-word matrix provides information representative of a surface level of processing (called the *explicit relational trac*e). Syntactic information is provided by a dependency grammar parser, which provides two *explicit sequential traces* (one for each side of the dependency unit). Thus, this model potentially captures both textbase (gist) and surface (explicit) level representations of text, as well as syntactic information. The CI-II model randomly samples from the explicit sequential trace with the constraint that the sample be both semantically and syntactically relevant as well. Kintsch and Mangalath show that the conditionalized combination of these three sources of information is more powerful across of a range of semantic, lexical, and sentential tasks compared to using only one or two sources or using LSA.

The CI-II model (Kintsch & Mangalath, 2010, this volume) also emphasizes the importance of context in deriving meaning from words and sentences. Whereas multiple meanings of words may reside in long-term memory, a generative *comprehension* model is necessary to weed out these multiple word senses when a word in understood in the context of text and discourse. Whereas a word such as *band* has multiple meanings in long-term memory, its meaning is constrained by context in sentences such as *He played in a rock band* and *He wore his wrist band*. The CI-II model narrows down the meaning of words in context by making use of multiple sources of information and basing activation on the combined conditionalized probabilities. This operates somewhat like the predication model (Kintsch, 2001, 2007, 2008). Such an approach allows the contextualized meaning of words to emerge iteratively in working memory. This is particularly important for accounting for more complex cognition, such as the understanding of metaphor.

7. Conclusion

Collectively, the research presented in this topic supports a number of conclusions. First, numerous models and variations on models have been developed over the last two decades, with a dramatic growth in interest and research in the last decade. Semantic models that apply statistical algorithms to large corpora of text are powerful means for extracting meaning from text. The computational power that has emerged over the last three decades has afforded the ability to analyze large corpora that successfully represent a good deal of human knowledge. Some have questioned the ability of such techniques in capturing the true essence of human cognition because ''bag of words'' techniques potentially miss out on important aspects of cognition. However, we see here that even nonlatent approaches to corpora analysis successfully capture much of what has been deemed well beyond the mere words in text. Indeed, Louwerse (2010, this volume) shows that a good deal of the results reported by embodiment theorists can be simulated using both nonlatent and latent (i.e., LSA) statistical models.

Second, semantic models can be augmented by combining them with data generated by humans, by accounting for word order or syntax, and by accounting for multiple levels of cognitive processing. Essentially, it seems that including multiple sources of information and assuming multiple levels of processing will be necessary for models to account for a wide range of cognitive data. For the most part, semantic models simulate human knowledge, as it resides inertly in long-term memory. The contents of a model's knowledge base can be manipulated by controlling the text that comprises the space. For example, to simulate an

8-year-old, one would use text and discourse to which a typical 8-year-old might have been exposed. The performance of semantic models improves when the text corpora used to create the semantic space is contextually and developmentally constrained (e.g., Shapiro & McNamara, 2000; Stone et al., 2010, this volume).

The importance of the ability of semantic models to simulate human knowledge should not be trivialized—it was a path-breaking achievement both theoretically and practically. Nonetheless, one challenge for these models is to go beyond the fine tuning of extracting semantic similarity based on statistical constraints in corpora, which are in turn aligned with particular mathematical properties. Semantic models must account for complex cognitive phenomena such as humans' understanding of synonymy, paraphrase, metaphor, and coherence. Indeed, several of the researchers featured in this issue have just done that—or at least paved the road to do so in future research.

To go beyond the simulation of knowledge, and account for performance on a wide variety of tasks, it seems that semantic models must use a combination of approaches. First, models that combine assumptions relevant to both episodic memory and semantic processing are successful in accounting for a variety of phenomena, including incremental learning, memory, and semantic processing (Hinton & Salakhutdinov, 2010, this volume; Howard et al., 2010, this volume; Jones & Mewhort, 2007). Second, information from syntax often plays an important role in the processing of text and discourse, and by consequence including sources of information representative of syntax (e.g., word order, grammatical dependencies) improves model performance (Dennis, 2005; Dennis & Kintsch, 2008; Jones et al., 2006; Kintsch & Mangalath, 2010, this volume; Riordan & Jones, 2010, this volume). Third, comprehension comprises multiple levels of processing, including surface, textbase, and situation model levels of understanding (Kintsch, 1998), and thus including multiple sources of information may be necessary to account for the full scope of comprehension, memory, and learning phenomena (Graesser & McNamara, 2010, this volume; Kintsch & Mangalath, 2010, this volume). Indeed, many practical applications use latent representations extracted using statistical algorithms such as LSA in combination with information from the words and syntax (e.g., McNamara, Boonthum, Levinstein, & Millis, 2007).

The need for multilevel models may not always be apparent because some levels of processing overwhelm the others, depending on the situation and the targeted dependent variable (McNamara & Magliano, 2009). For example, oftentimes prior knowledge overwhelms other factors to the extent that there are few discernable contributions from the text itself (other than word frequency). Likewise, when the targeted construct is the global text understanding, or a document, then the effects of syntax may be overwhelmed by the meanings of the words and the text as a whole. This is likely an explanation for why statistical models such as LSA that ignore syntax are oftentimes able to successfully extract meaning from text, despite ignoring fundamental levels of text meaning. Nonetheless, extracting the full meaning of text, including the full glory of its multidimensionality, will require using multiple, complementary approaches. The future likely lies more in the combination of techniques, rather than determining one winning model. Better, more complete, models of semantics are likely to emerge by measuring multiple levels of meaning.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavior and Brain Sciences*, *22*, 577–660.

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Series B*, *358*, 1177–1187.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

De Vega, M., Glenberg, A., & Graesser, A. C. (2008). *Symbols and embodiment: Debates on meaning and cognition*. Oxford, England: Oxford University Press.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*, 145–193.

Dennis, S., & Kintsch, W. (2008). Text mapping and inference rule generation problems in text comprehension: Evaluating a memory-based account. In F. Schmalhofer & C. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 105–132). Mahwah, NJ: Erlbaum.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61–74.

Fenson, L., Dale, P., Reznick, S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*, 1–185.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Special volume of the Philological Society. Oxford, England: Blackwell.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*, 1–55.

Graesser, A. C., & McNamara, D. S. (2010). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01081.x

Hinton, G., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Hinton, G., & Salakhutdinov, R. (2010). Discovering binary codes for documents by learning deep generative models. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01109.x

Hintzman, D. L. (1984). MINERVA2: A simulation model of human memory. *Behavior, Research Methods, Instruments, and Computers*, *16*, 96–101.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Howard, M., Shankar, K., & Jagadisan, U. (2010). Constructing semantic representations from a gradually-changing representation of temporal context. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01112.x

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202.

Kintsch, W. (2007). Meaning in context. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 89–105). Mahwah, NJ: Erlbaum.

Kintsch, W. (2008). Symbol systems and perceptual representations. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 145–163). New York: Oxford University Press.

Kintsch, W., & Mangalath, P. (2010). The construction of meaning. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01107.x

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, *12*, 703–710.

Landauer, T. K. (2007) LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 3–34). Mahwah, NJ: Erlbaum.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2010). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01106.x

Lowe, W. (2001). Towards a theory of semantic space. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Conference of the Cognitive Science Society* (pp. 576–581). Mahwah, NJ: Erlbaum.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavioral Research Methods, Instrumentation, and Computers*, *28*, 203–208.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*, 3rd ed. Mahwah, NJ: Erlbaum.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227–241). Mahwah, NJ: Erlbaum.

McNamara, D. S., Cai, Z., & Louwerse, M. M. (2007). Comparing latent and non-latent measures of cohesion. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 379–400). Mahwah, NJ: Erlbaum.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292–330.

McNamara, D. S., & Magliano, J. P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 298–372). New York: Elsevier Science.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Murdock, B. B. (1992). Serial organization in a distributed memory model. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 201–225). Hillsdale, NJ: Erlbaum.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*, 121–152.

O'Rourke, S. T., & Calvo, R. A. (2009). Visualizing paragraph closeness for academic writing support. In S. Murugesan (Ed.), *Handbook of research on the web 2.0, 3.0, and X.0 technologies, business, and social applications* (Ch. XLVII). Hershey, PA: IGI Global.

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Paivio, A., & Sadoski, M. (in press). Lexicons, contexts, events, and images: Commentary on Elman (2009) from the perspective of dual coding theory. *Cognitive Science*.

Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 657–663.

Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01111.x

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2005). An improved method of deriving word meaning from lexical co-occurrence. Unpublished manuscript. Available at: http://tedlab.mit.edu/~dr/. Accessed January 15, 2010.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*, 613–620.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*, 190–195.

Shapiro, A. M., & McNamara, D. S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research*, *22*, 1–36.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *1*, 214–241.

Steyvers, M., Chemudugunta, C., & Smyth, P. (2010). Combining background knowledge and learned topics. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2010.01097.x

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 427–448). Mahwah, NJ: Erlbaum.

Stone, B. P., Dennis, S. J., & Kwantes, P. J. (2010). Comparing Methods for Single Paragraph Similarity Analysis. *Topics in Cognitive Science*, 10.1111/j.1756-8765.2010.01108.x.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*, 183–190.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In C. Clarke and G. Cormack (Eds.), *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 267–273). New York: ACM Press.