The Presence and Absence of Category Knowledge in LSA

Christian D. Schunn (cschunn@gmu.edu)
Department of Psychology; MSN 3F5
George Mason University
Fairfax, VA 22030-4444

Abstract

How much information about meaning is contained in the statistical structure of the environment? LSA is a theoretical and practical tool that is challenging previous notions about what is contained in the statistical structure of the environment. This paper examines what kind of category knowledge can be obtained from the environment using LSA. In particular, two experiments are conducted with LSA to test what kind of category structure it embodies. LSA ratings about the relatedness of categories to their properties are compared with human judgments regarding the centrality of properties to the categories. LSA is found to capture aspects of property centrality for some object and event categories. However, it is found to only capture those aspects related to typicality: how often do members of the category have that property? LSA fails to capture other aspects of centrality that can be found in human category judgments. Thus, it appears that humans do bring other constraints to bear in shaping their categories.

Introduction

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique developed by Landauer and Dumais (1997) for automatically constructing a semantic representation of terms based on how they co-occur in a large corpus of texts. In the last several years, LSA has become an exciting tool for cognitive science, both as a psychological model and as a practical tool.

As a psychological theory, LSA claims that people learn the meaning of words not from learning definitions or from innate constraints, but instead by simply observing the contexts under which terms are used. In other words, the claim is that the statistical structure of the environment contains all the information that is needed to determine the meanings of words. Within this theoretical framework, LSA has been used as a model of child word acquisition (Landauer & Dumais, 1997), subject matter knowledge (Landauer, Foltz, & Laham, 1998), and semantic priming (Landuaer & Dumais, 1997). In each case, it has provided powerful demonstrations regarding how much information can be derived purely from the statistical structure of the environment.

As a practical tool, LSA has also been used for grading essays (Landauer, Foltz, & Laham, 1998), text research (Foltz, 1996), information retrieval (Dumais, 1994), and selecting reviewers for papers (Dumais & Neilson, 1992). A longer list of applications of LSA can be found in Landauer, Foltz, and Laham (1998). Again, in each of these cases, the statistical structure of the environment was proven to be surprisingly rich.

How does LSA work?

LSA begins with a very large text corpus (e.g., an encyclopedia, a series of textbooks in an area, or a large set of smaller documents). Next, all words that occur in the corpus are found. Those that occur with extremely high frequency (e.g., 'the', 'of', 'a') are removed. For all the remaining words (potentially tens of thousands), the frequency with which they co-occur in a given context (usually defined as a paragraph) are counted. For example, how often does the word 'beak' occur in paragraphs that have the word 'tail'? This produces an enormous matrix of co-occurrence information of size N by N, where N is the number of words being examined (which can be as high as 100,000).

LSA then reduces this huge frequency matrix into a much smaller matrix by a process called singular value decomposition (SVD). This new matrix is now of size N by M, where M is some number usually between 5 and 500. This reduction in size is conceptually similar to factor analysis or multidimensional scaling: it produces a more compact representation of the important statistical regularities in the larger matrix.

In the reduced matrix, each word can be thought of as a vector or point in an M-dimensional space. Since the reduced matrix is derived from frequency co-occurrence information, it represents words that occur in similar contexts with similar representations. This process forces synonyms to have very similar representations because, even though they rarely both occur in the same context, they co-occur with the same other words.

In LSA, one can also represent groups of words, sentences and paragraphs as a point in this M-dimensional space. By adding context words, the meanings of polysemous words are disambiguated. In any case, the relative similarity in meaning of two items (two words, two sentences, two paragraphs, a word and a sentences, etc.) is defined by their proximity in the M-dimensional space (in particular, the cosine of the angle between each vector). The closer the two items, the more similar their meanings.

On the surface, this complex process may seem psychologically implausible. However, some classes of neural nets do produce an approximation of SVD. Thus, the LSA claim is that humans may be doing something computationally equivalent to SVD. A more precise description of the mechanics of LSA and SVD can be found in Deerwester, Dumais, Furnas, Landauer, and Harshman (1990).

In sum, LSA develops a semantic representation of terms based on underlying (or latent) information found in the statistical regularities in the environment. While the success of its past applications to a variety of domains is impressive, the nature of the representation that it develops requires further investigation. Of particular focus in this paper is the following question: what kind of category knowledge does LSA develop? If LSA develops a human-like semantic representation of terms, then categorical structure of human knowledge should also be found in its representations.

This research question about LSA is also a more general question about how much information can be derived from the statistical structure of the environment (specifically of text and speech input that humans receive). If LSA can produce a good account of category structure, then the statistical structure of the environment may include more information than we may have originally thought.

Property centrality

Almost all categories have a structure such that some instances of the category are considered more central than other instances. For example, a robin is considered a more central instance of the category of birds than is an ostrich. Moreover, some properties of the category are considered more central than other properties. For example, having wings is considered more central to the category of birds than having a beak. In this paper, I will focus on the centrality of properties.

First, it is important to distinguish between property typicality and property centrality. Property typicality is how often members of the category have that property (e.g., what proportion of birds have wings?). By contrast, property centrality is the importance of the property to a person's concept. For example, imagine something that was like a bird, but didn't have wings. How good an example of the concept of birds would that be? How much would it change your concept? How likely would you be to call it a bird? All of these questions address the centrality of the property to the category, or how important the property is to the category.

What makes a property central? Certainly the property's typicality is very important (Rosch, 1975). However, centrality is not synonymous with typicality. Consider the case of curvedness and bananas and boomerangs. All bananas and all boomerangs that people will have seen are curved (i.e., curvedness is equally typical of both boomerangs and bananas). Yet, curvedness is considered more central to the category of boomerangs than it is to the category of bananas (Murphy & Medin, 1985). It was argued that curvedness is especially important to boomerangs because people have a theory about the role that curvedness plays to the fundamental activities of a boomerang.

In a more recent study, Schunn and Vera (1995) found that it is specifically causal factors that play a very important role in determining property centrality. For both objects (e.g., pigeons and cars) and events (e.g., birthday parties and elections), the properties that were considered to play an important causal role in the functioning of the object/event were considered central to the category. In fact, they were considered more central than properties considered part of a formal definition or properties used to recognize the object/event. Of course, property typicality was also an important predictor of property centrality.

In sum, property centrality in humans appears to be a combination of property typicality and causal theories (see also Ahn and Lassaline, 1995). What kind of account of categorization does LSA provide? If LSA is deriving true word meaning, then it should understand the relationship between a category term and its properties. Since LSA derives its semantic structure from frequency information, it should at least be able to provide a good account of the typicality effects in category structure.

Yet, it is unclear whether LSA will be able to account for more of centrality (meaning) than just typicality. Or, asked another way, what information is found in the statistical structure of the environment? Certainly first order statistics (as measured by typicality) do not fully capture centrality. However, the complete statistical information found in textual corpuses may produce patterns in LSA that mimic what appears to be the role of causal theories in human categorization. This added information may occur either because textual corpuses contain a bias towards mentioning causal information or because higher-order statistics contain additional information.

To examine these questions, I compared categorization data from Schunn and Vera (1995) (henceforth called SV) against the responses produced by LSA to the exact same categories and properties. In Experiment 1, I report the results for analyses of object categories (both artifacts and biological kinds). In Experiment 2, I report the results for analyses of event categories.

Experiment 1: Objects

Methods

Human Data Source The human data for Experiment 1 were taken from SV Experiments 1b, 1c, and 1d. In particular, the data from the six familiar objects were used. These objects included three familiar biological kinds (cats, pigeons, and mice) and three artifacts (cars, toilets, and power lawn mowers). Each object had 6 properties. These properties were the ones most commonly generated by people as properties for each of these categories (SV, Experiment 1a). Table 1 presents an example of one of the categories and its properties. The full list of properties can be found in the original paper.[1]

Table 1: An example object (pigeons) and its properties and z-scored property centrality ratings.

Centrality	Properties
0.628	Pigeons have wings.
0.612	Pigeons have feathers.
0.392	Pigeons fly.
-0.429	Pigeons are gray.
-0.555	Pigeons coo.
-0.648	Pigeons live in big cities.

For all six properties of all six objects, SV obtained five different ratings from five different groups of subjects. First, property centrality ratings (the importance of the property to the category) were obtained by asking subjects how much their concepts were changed by negating each property. For example, subjects were asked how much does your concept of pigeons change if a particular instance did not have wings. Each subject's ratings were z-score transformed. Then a mean normalized rating for each property was determined. This mean for each property served as the input for the current analyses. Table 1 presents the centrality ratings for each property of pigeons. Higher ratings represent more central properties.

Second, property typicality ratings (the frequency with which category members have the property) were obtained by asking subjects either what proportion of those objects in the world had each property (or what proportion of objects that they had seen had each property). The two variations produced highly correlated responses, and the average frequency rating across both variants was computed for each property.

Third, subjects were asked to rate each property of each object in terms of how important it was for a scientific or expert definition of the object category. Fourth, subjects were asked to rate each property of each object in terms of how important it was for recognizing members of the object category. Fifth, subjects were asked to rate each property in terms of how important the property was for the object to be successful at what it did. Mean ratings were determined on each scale for each property. These last three ratings were called definition, recognition, and cause, respectively.

In sum, the human data consisted of ratings of each property of each object on five dimensions: centrality, typicality, definition, recognition, and cause. SV determined that centrality judgments were best predicted by typicality and cause. At issue now is to what LSA judgments correspond.

[1] An online version of the paper can be found at http://hfac.gmu.edu/~schunn.

LSA Data The LSA data was gathered using the web version of LSA available at http://lsa.colorado.edu.[2] As previously used in many tests of LSA, a data space derived from the TASA college corpus was selected. Of the available data spaces, this one best represents general knowledge and experiences of a college student (who was the source of the human data). The TASA college corpus uses a variety of texts, novels, newspaper articles, and other information. It includes 37,651 documents and 92,409 terms.[3] The default setting of 300 dimensions was used and those results are reported below. However, all analyses were also redone using only 100 dimensions to examine the role of number of dimensions.

To get judgments of semantic similarity or relatedness from LSA, LSA was asked to compare each category name with its properties. For example, LSA was asked to compare "pigeon" with "has wings". The same object and property names were given to LSA as were given to the human subjects. For each of these comparisons, LSA produced a number between -1 and 1 representing the perceived similarity (with 1 being perfectly similar and -1 being perfectly dissimilar).

Results

How Well Does LSA Predict Property Centrality? To see how well LSA values predicted property similarity, the 36 LSA values were correlated against the 36 human centrality means (6 objects and 6 properties per object). Across the 36 properties, LSA correlated $r=.26$, $p<.1$ with property centrality. The correlations were slightly larger when done separately for artifacts ($r=.29$) and biological kinds ($r=.32$). Computing the correlations separately for each object, two of the three artifacts (cars $r=.42$, toilets $r=.91$) and one of the three biological kinds (pigeons $r=.99$) had noticeably positive correlations. Correlations for the other three objects were: cats $r=-.37$, mice $r=.04$, and power lawn mowers $r=-.46$. While some of these correlations are based on extremely low Ns and are correspondingly noisy, it appears that LSA did predict some aspects of property centrality, but quite poorly for some artifacts and biological kinds.

What Predicts Which Objects Will Have Centrality Well Predicted By LSA? Why did LSA predict centrality quite well for some objects but terribly for others? In particular, what generally determined whether LSA could predict property centrality for a particular object? The most important determinant was how well LSA predicted property typicality. For several of the objects, LSA did not predict property typicality

[2] In particular, the matrix comparisons application was used, with the "term to term" comparison method

[3] At the web location, this topic space is called "General reading up to 1st year college"

(i.e., how often members of the category have the given property). This can be readily seen by comparing the correlation between LSA and typicality for each object to the correlation between LSA and centrality for each object. The correlation among these correlation values was r=.97 (N=6, p<.01). In other words, if LSA did not account for typicality effects, it did not account for centrality effects. But, this explanation does not tell us whether there was something inherently different between the three objects LSA did predict well and the three objects LSA did not predict well.

Another explanation that focuses on properties of the objects themselves is how well the different components of centrality correlated with one another. There were four measures (from SV) that all played some role in centrality to varying degrees (in decreasing order of importance): typicality, cause, recognition, and definition. For each object, the correlations were computed between each of these measures. Then the average of the 6 pairwise correlations was computed for each object. This average correlation represents how consistent the various components of centrality were with one another. For example, for some objects, properties high in typicality were also high in causal importance, high in recognitional importance, and high in definitional importance, whereas for other objects, properties high in typicality may not have been high in causal importance or not high in recognitional importance. This average correlation is purely a property of the objects and is not logically tied to LSA. It is truly an independent predictor of LSA's quality of fits to centrality. Thus, it is impressive that it did predict well how LSA was related to property centrality for each object: the correlation among correlations (average intercorrelation vs. LSA to centrality correlation) was r=.82 (N=6, p<.02). Moreover, since each individual correlation was based on such a small N and so is likely to be quite noisy, the strength of this correlation is noteworthy.

To make this last analysis more concrete, let us consider the examples of objects with centrality well and poorly predicted by LSA. The six properties of pigeons are listed in Table 1. The two most central properties of the category are having feathers and wings. Those properties are also in the top three for ratings of typicality, definition, recognition, and cause. Thus, LSA could receive information about any of these factors from the text corpus structure and be able to predict centrality well.

By contrast, consider the case of cats. For the category of cats, having fur and four legs was considered central. While these two properties were also rated as most typical of cats, the relatively less typical property of having claws and meowing were viewed as very important in ratings of definition, recognition, and cause (although also tied with having four legs). Thus, the various sources of information that LSA may be abstracting conflict with one another, and thus produce poor judgments of centrality.

In sum, one can say that LSA did a good job of predicting centrality when it did a good job of predicting typicality, or one can say that LSA was able to predict centrality for objects that had a simple and consistent property structure.

Does LSA Capture More Than Typicality? For humans, centrality is more than just typicality (Murphy & Medin, 1985; Schunn & Vera, 1995). In this human data set (as SV showed), cause was a very important predictor of centrality above and beyond typicality. For example, partialling out the correlation of typicality with centrality, cause still has a high partial correlation with centrality (partial r=.48, N=36, p<.01).

Does LSA also have more to it than typicality? In particular, can it predict aspects of centrality above and beyond typicality effects? To examine this issue, LSA predictions were correlated with centrality after partialling out the correlation of typicality with centrality. Unfortunately, LSA predicted very little of centrality beyond typicality (partial r=.06, N=36, p>.5).

Discussion

At least for objects, it appears that LSA can only account for typicality aspects of category structure. It does not predict aspects of centrality beyond the effects of typicality. These results can be viewed in a positive or negative light. On the positive side, LSA can predict property centrality quite well for some objects. Give that is simply uses word co-occurrence from a textual corpus, this result is not trivial. On the negative side, its correlations overall are quite weak, and for some objects, it does not at all predict property centrality.

The analyses provided some insight into the circumstances under which LSA can predict centrality. When LSA does not predict typicality well, it does a terrible job of predicting centrality. Moreover, those objects that have complex property structure appear to be the most difficult for LSA to capture.

In testing LSA against human judgments, the default (and recommended) value of 300 dimensions was used. To examine whether the results depended on the number of dimensions, the analyses were redone LSA values based on 100 dimensions. The new values produced slightly larger correlations for some objects and slightly smaller correlations for others. Overall, results remained the same. Thus, the results appear to generalize across a range of dimension settings.

The results also do not appear to be very specific to minor variations in how the text is presented to LSA. LSA produces similar numbers when the singular form of the category names were used (e.g., mouse versus mice or car versus cars). Correlations between singular and plural ranged between .72 and .98.

Experiment 1 focused on object categories. Events are another important class of categories, and have a wide variety of subtypes. For example, there are socially defined events (weddings, parties) that are similar to artifacts. There are also more naturally defined events (thunderstorms, car accidents) that are similar to biological kinds. Experiment 2 examines how well LSA can predict category structure for events.

Experiment 2: Events

Methods

Human Data Source The human data for Experiment 2 were taken from SV Experiments 2b, 2c, and 2d. The data included 16 familiar events, including social, formal, and physical events of varying durations: birthday party, birth, breakfast at diner, car accident, getting dressed to go out, elections, getting a haircut, grocery shopping, having a cold, using an ATM machine, making coffee, making photocopies, making a phone call, taking a final exam, thunderstorm, and wedding. There were between 6 and 10 properties for each event. As with Experiment 1, these properties were the ones most commonly generated by people (SV, Experiment 2a). Table 2 presents an example of one the categories and its properties.

Table 2: An example event (weddings) and its properties and z-scored property centrality ratings.

Centrality	Property
1.006	groom
.913	bride
.303	relatives
.150	rings
.078	cake
.009	priest
-.093	gifts
-.426	throw bouquet
-.440	drinking

As with Experiment 1, the human data consists of norms on the same five dimensions using similar procedures: centrality, typicality, definition, recognition, and cause. Again, mean subject ratings on each dimension will be used as the input for the current analyses.

LSA Data The LSA was gathered in the same manner as for Experiment 1: the web implementation of LSA using the semantic space derived from the TASA database with 300 dimensions. Of the 118 properties, LSA did not know two of the property terms: photocopier (for making photocopies) and mudslinging (for elections), and thus those two properties were deleted from the analyses.

Results

How Well Does LSA Predict Property Centrality? To see how well LSA values predicted property similarity, the 116 LSA values were correlated against the 116 human centrality means (16 events and 6–10 properties per event). Across the 116 properties, LSA correlated $r=.21$, $p<.05$ with property centrality. To see how well it predicted property centrality within each of the events, the correlations were computed separately for each event. Eight of the events produced noticeably positive correlations with centrality (in decreasing order of predictiveness): weddings $r=.88$, phone call $r=.88$, thunderstorm $r=.85$, births $r=.74$, using ATM machine $r=.44$, making coffee $r=.34$, having a cold $r=.26$, and dressed to go out $r=.25$. The other eight events produce very small or negative correlations with centrality (in decreasing order of predictiveness): car accident $r=.18$, elections $r=.11$, getting a haircut $r=-.12$, breakfast at diner $r=-.23$, making photocopies $r=-.39$, birthday parties $r=-.48$, taking final exam $r=-.53$, grocery shopping $r=-.54$. In sum, LSA did not predict centrality well overall, but did predict centrality for half the events.

What Predicts Which Objects Will Have Centrality Well Predicted By LSA? As with the objects, one can examine what determines when LSA will predict event property centrality. The same two factors proved to be important. First, when LSA did not predict property typicality, it did not predict centrality. The correlation of correlations (LSA vs. typicality and LSA vs. centrality) was $r=.86$, ($N=16$, $p<.01$). Second, when the events had a complex property structure, LSA did not predict centrality. The correlation between the average intercorrelation for each event (mean pairwise intercorrelation among typicality, cause, recognition, and definition) and the LSA-centrality correlation for each event was $r=.47$ ($N=16$, $p<.05$).

Does LSA Capture More Than Typicality? As with objects, for humans cause is a very important predictor of centrality. In this SV data set, cause predicts centrality significantly above the contributions of typicality (partial $r=.46$, $N=118$, $p<.001$). By contrast, LSA only predicts a very small amount of centrality beyond typicality (partial $r=.18$, $N=116$, $p<.1$).

Discussion

As with objects, it appears that LSA can only account for typicality aspects of category structure. When LSA does not predict typicality, it does not predict centrality. Again, when objects had complex property structure, LSA also did not predict centrality well. However, LSA did predict centrality relatively well for at least half of the events, and this accomplishment should not be minimized.

As in Experiment 1, LSA was tested using the default value of 300 dimensions. To examine whether the results depended on the number of dimensions, the analyses were redone using LSA values based on 100 dimensions. As with objects, the overall, results remained the same. Thus, the results continue to generalize across a range of dimension settings.

Once again, the results also do not appear to be very specific to minor variations in how the text is presented to LSA. For example, correlations between the singular and plural (e.g., birth vs. births, birthday vs. birthdays) were usually in the .9 range.

General Discussion

One goal of the research presented here was to determine whether and under which circumstances LSA has a human-like category structure to its semantic space. In general, the results were a story of the glass half-full and the glass half-empty. For the glass half-full, LSA was able to produce modest correlations with centrality overall, and reasonably strong correlations with centrality for at least half the objects and events that were examined. Thus, there are some important human-like properties to the semantic space produced by LSA. For the glass half-empty, LSA was primarily restricted to typicality effects. LSA could not account for aspects of property centrality beyond typicality. This is in sharp contrast to human categories, for which causal aspects play a very important role in both objects and events. When categories had complex structure, in which typicality, definitional, recognitional, and causal factors did not all correlate highly with one another, LSA did not predict property centrality well.

The implications of these findings extend beyond LSA. They suggest that the textual environment appears not to have aspects of centrality beyond typicality hidden in simple statistical structure. If this implication is correct, then this provides further evidence that if human categories have aspects of centrality beyond typicality, these aspects must come purely from a top-down bias. Of course, it is possible that other methods for retrieving information from statistical structure may have found other aspects of property centrality.

There are important limitations of the work presented here that must be acknowledged. First, there are other kinds of category knowledge than property centrality. For example, there is also instance centrality—how central an instance is of a category (e.g., that a robin is more central an instance of the category birds than is an ostrich). Preliminary work by Laham (1998) suggests that LSA can capture some aspects of this category knowledge. Moreover, there are other ways of defining property centrality.

Second, it is possible that using another textual corpus to construct the semantic space with LSA or using different similarity metrics between items in LSA would have produced better category knowledge. Similarly, alternative schemes to LSA might provide a better extraction of information from the environment. For example, the HAL model (Burgess, 1998; Burgess & Conley, 1998) uses a different scheme that takes into account how close the words appear within a context. However, HAL can only represent the similarity between words, not words and phrases as in LSA.

Third, there are many possible variations that could have been used in presenting the properties to LSA. LSA was given the variations most similar to what the human participants saw. However, it is possible that other variations (e.g., adding the object and event names to the property sentences) might have changed the results, and this needs to be explored further.

As a final note, the research presented here illustrates the advantages of the world wide web as a research tool. Complex computational engines can be made available to other researchers without requiring the overhead of maintaining the software or specialized computers to run the software.

Acknowledgements

I would like to thank Greg Trafton, Erik Altmann, Lelyn Saner, Sherryl Miller, Audrey Lipps, and Susan Trickett for comments made on drafts of this paper.

References

Ahn, W., & Lassaline, M. (1995). Causal structure in categorization. In the *Proceedings of the 17th Annual Conference of the Cognitive Science Society.*

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30, 188-198.

Burgess, C., & Conley, P (1998). Representing proper names and objects in a common semantic space: A computational model. *Brain and Cognition.*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407.

Dumais, S. T., & Nielson, J. (1992). Automating the assignment of submitted manuscripts to reviewers. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavioral Research Methods, Instruments and Computers, 28*(2), 197-202.

Laham, D. (1998). Latent Semantic Analysis approaches to categorization. Poster presented at the *19th Annual Meeting of the Cognitive Science Society.*

Landauer, T. K. & Dumais, S. T. (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., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic *Analysis. Discourse Processes, 25*, 259-284.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Schunn, C. D., & Vera, A. H. (1995). Causality and the categorization of objects and events. *Thinking & Reasoning, 1*(3), 237-284.