Modeling canonical and contextual typicality using distributional measures

Louise Connell[1] and Michael Ramscar[2]
[1]*Dept. Computer Science, University College Dublin, Dublin 4, Ireland <louise.connell@ucd.ie>*
[2]*ICCS, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland*

The underlying assumption in much of categorization research is that effects such as typicality are reflective of stored conceptual structure. This paper questions this assumption by simulating typicality effects by the use of a distributional model of language, Latent Semantic Analysis (LSA). Despite being a statistical tool based on simple word co-occurrence, LSA successfully simulates participant data relating to typicality effects and the effects of context on categories. Moreover, it does so without any explicit coding of categories or semantic features. In the light of the findings reported here, we question the traditional interpretation of typicality data: are these data reflective of underlying structure in people's concepts, or are they reflective of the distributional properties of the linguistic environments in which they find themselves.

INTRODUCTION

How do humans pick out regularities in the stuff of experience and index them using words? Here, we wish to consider the idea that language itself is part of the environment that determines conceptual behavior. A growing body of research indicates that distributional information may play a powerful role in many aspects of human cognition. Saffran, Newport and Aslin (1996) have demonstrated that infants and adults are sensitive to simple conditional probability statistics, suggesting one way in which the ability to segment the speech stream into words may be realized. Redington, Chater & Finch (1998) suggest that distributional information may contribute to the acquisition of syntactic knowledge by children.

The objective of this paper is to examine the extent to which distributional measures can model human categorization data: What is the relationship between typicality judgements and distributional information? Are the responses people provide in typicality experiments more reflective of the distributional properties of their linguistic environments than they are of an underlying conceptual structure?

Typicality Effects and Distributional Measures

Rosch (1973) provided the first empirical evidence of typicality effects by giving participants a category name with a list of members and asking them to rate how good an example each member was of its category. The results showed a clear trend of category gradedness – e.g. apples are consistently judged a typical fruit, while olives are atypical. Roth & Shoben (1983) later showed that the context a concept appears in affects the typicality of its instances. A typical bird in the context-free sense may be a robin, but in the context "The bird walked across the barnyard", chicken would instead be typical. They found that measures of typicality in isolation do not play a predictive role once context has been introduced.

According to Rosch (1978), typicality ratings predict the extent to which the

member term is substitutable for the superordinate word in sentences. This has a parallel in distributional approaches (e.g. Landauer & Dumais, 1997; Burgess & Lund, 1997). In a distributional model of word meaning such as Latent Semantic Analysis (LSA), a contextual distribution is calculated for each lexeme in the corpus by counting the frequency with which it co-occurs with every other word. In this way, two words that tend to occur in similar linguistic contexts will be positioned close together in semantic space. By using this proximity of points as a measure of their contextual substitutability, LSA offers a tidy metric of distributional similarity

EXPERIMENT 1 – CANONICAL TYPICALITY

The purpose of this experiment is to examine whether data from typicality studies can be modeled using a distributional measure. Specifically, it was predicted that participant typicality scores from previous studies would correlate with a distributional measure (LSA; Landauer & Dumais, 1997) when comparing similarity scores for category members against their superordinate category name.

Method

Each set of typicality data was divided up according to the original study: Set A was taken from Rosch (1973), B from Armstrong, Gleitman & Gleitman (1983), C from Malt & Smith (1984). Within these three data sets, 18 sets of typicality ratings existed, across 12 separate categories. For each category in each data set, all items were compared to the superordinate category name and LSA similarity scores noted. The LSA corpus used contains texts thought to represent readings up to college age. LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the studies.

Table 1
Rank Correlation Coefficients *rho* (With Significance *p*) Between LSA And Participant Scores

Category	Set A	Set B	Set C
sport	1.000 *(p<0.01)*	0.811 *(p<0.01)*	-
fruit	0.886 *(p<0.05)*	0.539 *(p<0.10)*	0.157 *(insignif)*
vehicle	0.829 *(p<0.10)*	0.788 *(p<0.01)*	-
crime	0.814 *(p<0.10)*	-	-
bird	0.714 *(p<0.10)*	-	0.375 *(insignif)*
science	0.414 *(insignif)*	-	-
vegetable	0.371 *(insignif)*	0.580 *(p<0.10)*	-
female	-	0.346 *(insignif)*	-
trees	-	-	0.705 *(p<0.01)*
clothing	-	-	0.521 *(p<0.05)*
furniture	-	-	0.466 *(p<0.05)*
flowers	-	-	-0.499 *(insignif)*

Note— '-' appears where category was not present in set

Results

Spearman's rank correlation (rho) was used to compare scaled LSA and participant

scores. The global rank correlation between the participant ratings and LSA scores across all Sets (193 items) was rho=0.515 (2-tailed $p<0.001$). See Table 1 for full LSA results. It must be noted that the same rank correlation coefficient results in differing levels of significance. With small data sets (5 to 20 items), the power of the tests is restricted and sensitive to individual data points. Thus, given the constraints of the data, those results where $p<0.10$ are considered marginally significant.

In this experiment, LSA scores correlated significantly with participant typicality ratings. Without any hand-coding of category membership or salient features, LSA's semantic space successfully modeled gradients of typicality within categories. With some variation between categories, this experiment successfully shows a distributional measure modeling human typicality data with a global correlation significant to $p<0.001$ (rho=0.515).

EXPERIMENT 2 – CONTEXTUAL TYPICALITY

The first experiment indicates that a co-occurrence model such as LSA can be used to model typicality judgements in canonical (context-free) categories. However, categorization is also subject to linguistic context, whose capacity to skew typicality has been demonstrated by Roth & Shoben (1983).

The purpose of Experiment 2 was to test if LSA could be used to predict participant responses for typicality in context. The hypothesis was that LSA could predict human judgements of exemplar appropriateness (typicality) for given context sentences. LSA similarity scores for each context sentence and respective category members were used to form significantly different clusters of appropriate (high scores / similarity) and inappropriate (low scores / similarity) items. It was predicted that participant ratings of typicality in context for these items would fall into the same clusters, and that these clusters would also be significantly different.

Method

Materials consisted of 7 context sets, each of which contained a context sentence and 10 possible members of the category. 3 of the context sentences were taken from Roth & Shoben (1983), the other 4 created for this experiment. Category members were chosen in two ways, to form the appropriate and inappropriate clusters for the context. First, appropriate items were found by randomly selecting 4-5 high-level category members (e.g. cow, not calf, for category animal) that appeared in the list of the context sentence's 1500 near neighbors. This list corresponds to the 1500 points in LSA's high-dimensional space that would receive the highest similarity scores. Second, inappropriate items were found by compiling a large list of category members and selecting the 5-6 of those that had the lowest (preferably negative) LSA similarity score against the context sentence.

These materials were then split into two sections. Each consisted of 7 context sets containing 5 items, selected so that there were at least 2 of both appropriate and inappropriate items in the set and so that each category member appeared only once per section. Participants received one section apiece, with presentation of section 1

or 2 alternated between participants. All 35 items within each section were presented in random order, resampled for each participant. 19 native speakers of English volunteered to participate in this experiment via an electronic questionnaire.

The scores were calculated in LSA by comparing the context sentence to each item in the list, using the same corpus and scaling as for Experiment 1. Participants read instructions that explained typicality and the 7-point scale as per Rosch (1973), and were asked to rate the appropriateness of the member in each given context sentence.

Results

Participants agreed with LSA's predictions of typicality for 62 of the total 70 items – 10/10 items in 3 context sets, 9/10 items in 3 further context sets, and 5/10 in the remaining set. Significant difference in clusters, not rank correlation, is the important factor here, because even participant data with low correlation to the LSA score may fall into the two specified clusters (thus supporting the main prediction).

For all 7 context sets, Mann-Whitney (2-tailed) tests showed the LSA scores fell into two significantly different clusters. The participant scores' results for the predicted clustering varied: three context sets showed significant differences at $p<0.01$, three at $p<0.10$ and one set failed to achieve any significant difference ($p=0.69$). See Table 2 for full results.

The results support the basic hypothesis that, in the majority of cases, distributional information (in this case modeled in LSA) can predict whether members of a category will be appropriate or inappropriate in a given context. Whereas canonical typicality simulations essentially involve the comparison of individual lexemes already in the corpus, introducing context involves the ad-hoc creation of points in semantic space that are not already present. In other words, LSA can predict the more complex human judgement of typicality in context, as well as in canonical categories (Experiment 1).

Table 2

Wilcoxon's *W* And Significance Of Difference *p* Between Clusters For Each Context Sentence

Context Sentence	LSA	Participants
Stacy volunteered to milk the *animal* whenever she visited the farm *	10 ($p<0.01$)	10 ($p<0.01$)
Fran pleaded with her father to let her ride the *animal* *	15 ($p<0.01$)	15 ($p<0.01$)
The *bird* swooped down on the helpless mouse and carried it off	10 ($p<0.01$)	10 ($p<0.01$)
Jane liked to listen to the *bird* singing in the garden	15 ($p<0.01$)	18 ($p<0.10$)
Jimmy loved everything sweet and liked to eat a *fruit* with his lunch every day	15 ($p<0.01$)	18 ($p<0.10$)
Sophie was a natural athlete and she enjoyed spending every day at *sport* training	15 ($p<0.01$)	19.5 ($p<0.10$)
During the mid morning break the two secretaries gossiped as they drank the *beverage* *	15 ($p<0.01$)	25 ($p<0.70$)

Note—* Sentences taken from Roth & Shoben (1983)

GENERAL DISCUSSION

The success of these distributional modeling experiments suggests interesting

possibilities for a theory of categorization that incorporates information from the structure of language as well as from the structure of the world. Distributional models of language use a representation that is learned from the language alone, assuming that the way words co-occur with one another gives rise to clues about their semantic meaning. Gleitman (1990) has discussed a similar approach with regards to first language acquisition, where this type of representation can easily be learned from an individual's linguistic environment.

In this respect, the results reported here raise interesting questions regarding the mental representations of the meanings of words: Do people use distributional information to construct their representation of word meanings, or do the distributional properties of words merely fall out of the fact that underlying concepts share certain semantic features? Work by MacDonald & Ramscar (in press) would seem to indicate the former. They show that manipulating the distributional properties of the contexts in which nonce words are read can significantly influence similarity judgements between existing words and nonces. This indicates that not all distributional responses can be explained in terms of underlying conceptual structure, because nonce words won't have an existing conceptual structure.

What the results presented here (and other distributional research) seem to indicate is that any proper characterization of conceptual thought will have to consider more than just the information that comes from physical experience and environment. One must also consider experience of language, and the structure of the linguistic environments in which speakers find themselves.

REFERENCES

Armstrong, S. L., Gleitman, L. R. & Gleitman, H., (1983). What some concepts might not be. *Cognition*, **13**, 263-308.

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

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, **1**, 3-55.

Landauer, T. K. & Dumais, S. T., (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**, 211-240.

Malt, B. & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, **23**, 250-269.

MacDonald, S & Ramscar, M. J. A. (in press) Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. To appear in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (in press).

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, **22**, 425-469

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.) *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Rosch, E., (1978). Principles of Categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, N.J.: Erlbaum.

Roth, E. M. & Shoben, E. J., (1983). The effect of context on the structure of categories. *Cognitive Psychology*, **15**, 346-378.

Saffran, J. R., Newport, E. L. & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, **35**, 606-621.