Unique Entropy As A Model Of Linguistic Classification

Toben H. Mintz (tmintz@usc.edu)
Department of Psychology, SGM 501; University of Southern California
Los Angeles, CA 90064-1061 USA

Several researchers have proposed that young children could make use of statistically weighted distributional information as a significant source of information about the categories of words in their language (Cartwright & Brent, 1997; Mintz, Newport, & Bever, 1999; Redington, Chater, & Finch, 1998). Most of these analyses result in a hierarchical cluster analysis (HCA) which clusters words together based on their distributional similarity. HCAs do not produce categories, but rather graded clusters based on similarity. A similarity threshold must be chosen such that words in clusters which exceed the similarity threshold are said to belong to the same category. Finding a deterministic method for selecting the categorization threshold which results in optimal linguistic categorization, and which does not rely on *a priori* knowledge of the correct linguistic categories, has been problematic. In this poster, I propose a deterministic solution for choosing categorization thresholds in HCAs. I present the notion Unique Entropy which, when applied to linguistic corpora, yields an optimal categorization of words into grammatical categories.

One can characterize the notion "best categorization point" for a HCA on formal, information-theoretic grounds. Specifically, the similarity threshold which yields the highest Entropy (Equation 1, *l*=number of groups), will provide the categorization level which maximizes the intrinsic information carried by the resulting category structure. "Best categorization" in this sense means "best" in terms of the amount of information inherent in the resulting category structure, independent of whether it best approximates the linguistic categories being sought. It is an empirical question, whether the best information theoretic classification results in the best linguistic classification. I now demonstrate that it does, at least for the four corpora analyzed in Mintz et al. (1999).

$$(1)\quad E_l = -\sum_1^l \log(p(i))(p(i)),\quad p(i) = \frac{number\,of\,elements\,in\,cluster\;\;i}{l}$$

$$(2)\quad UE_l = E_l - ((-\log(\frac{m-(n-1)}{m})(\frac{m-(n-1)}{m})) + (n-1)(-\log(\frac{1}{m})(\frac{1}{m})))$$

The Entropy, or Information, in a set of categories is affected in two ways by the structure of the set. 1) For a set of a given number of categories, information contained in the category structure (Entropy) will be higher when categories contain the same number of items than when items are unevenly distributed among categories. 2) All else being equal, having more categories results in greater Entropy. In selecting an optimal categorization point based on maximum Entropy, one only wants to consider sources of Entropy that are due to the specific characteristics of the HCA in question and not which are due to merely having a certain number of categories. Therefore, to determine the unique information provided by a HCA of *m* items at a given categorization threshold, *l*, which yields *n* categories, one must subtract out the base information that would come merely from having *n* categories. The result of this subtraction I call Unique Entropy (UE, Equation 2).

Figure 1a plots Unique Entropy by number of categories for the distributional analyses of four corpora presented in Mintz et al. (1999). Mintz et al. reported that the best linguistic categorization in their HCAs was obtained when members were divided into about 30 groups. This is shown by the vertical bar in Figure 1a, and corresponds to the regions with the highest Unique Entropy points for each corpus. Thus, it appears that the best linguistic classification for these corpora is achieved by selecting the classification level with the highest Unique Entropy.

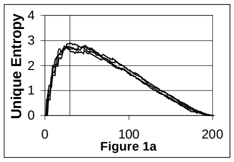

Figure 1a

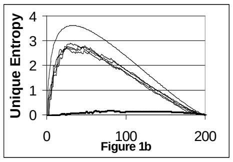

Figure 1b

Figure 1b shows that the specific character of the UE curves produced by the distributional analyses of child directed speech is not a necessary consequence of performing such an analysis on any corpus. The lowest line plots the average UE of 10 pseudo-corpora generated by randomly ordering the words in one of the four Mintz et al. corpora. This UE curve shows that any information inherent in the random pseudo-corpora HCAs is due simply to having a given number of categories. The top curve in Figure 1b shows the upper bound for UE when classifying 200 items into *n* categories. The four corpus based curves are repeated in Figure 1b. The structure of the actual corpus based HCAs are nearly maximally informative by this measure.

Further research will explore the implications of this finding for psycholinguistics, as well as investigate how it extends to other areas of human categorization. Perhaps humans have evolved categories which are structurally the most informative.

References

Cartwright, T. A., Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition. 63*, 121-170.

Mintz, T. H., Newport, E. L., & Bever, T. G. (1999). The Distributional Structure of Grammatical Categories in Speech to Young Children. Ms. under review.

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