Distributional Similarity, Phase Transitions and Hierarchical Clustering

Fernando Pereira
2D-447, AT&T Bell Laboratories
PO Box 636, 600 Mountain Ave
Murray Hill, NJ 07974-0636
pereira@research.att.com

Naftali Tishby
Department of Computer Science
Hebrew University
Jerusalem 91904
tishby@cs.huji.ac.il

Abstract

We describe a method for automatically clustering words according to their distribution in particular syntactic contexts. Words are represented by the relative frequency distributions of contexts in which they appear, and relative entropy is used to measure the dissimilarity of those distributions. Clusters are represented by "typical" context distributions averaged from the given words according to their probabilities of cluster membership, and in many cases can be thought of as encoding coarse sense distinctions. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data.

Motivation

Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in building statistical language models, particularly in models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example of frequencies of pairs of a transitive main verb and the head noun of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle [1990] proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear that his notion of similarity can be used directly to construct word classes and corresponding models of association.

Our research addresses some of the same questions and uses the same raw data, but we want to investigate how to factor word association tendencies in terms of associations of words to certain hidden *senses classes* and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes [Resnik, 1992], in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts c with corresponding membership probabilities $P(w \in c)$ for each word w, with $\sum_s P(w \in c) = 1$. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes [Brown *et al.*, 1990]. Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, V and N (for the verbs and nouns in our experiments) and a single relation between them (the main verb-head of direct object relation in our experiments) sampled by the frequencies f_{vn} of occurrence of particular pairs (v, n) in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiments was derived from newswire

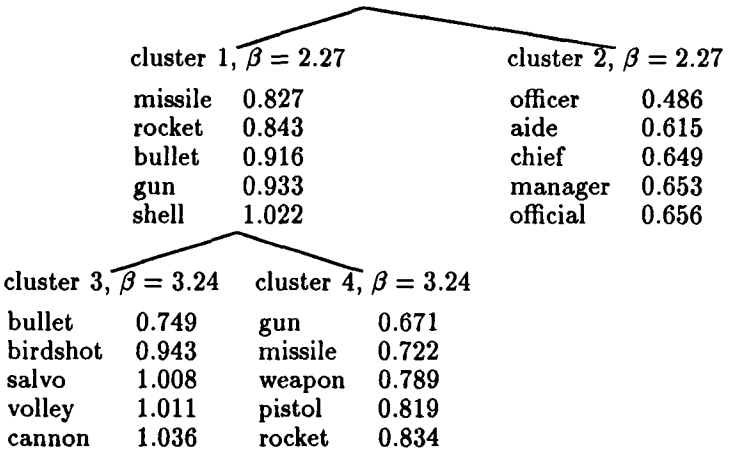

Figure 1: Direct object clusters for *fire*

text automatically parsed by Hindle's parser Fidditch. More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger [Church, 1988] and of tools for regular expression pattern matching on tagged corpora. We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. The empirical distribution of a noun n can then be given by the conditional density $p_n(v) = f_{vn}/f_{\cdot n}$ where $f_{\cdot n} = \sum_v f_{vn}$ is the marginal frequency of n. The problem we will study is how to use the p_n to classify the $n \in N$. Our classification method will construct a set C of clusters and cluster membership probabilities $P(n \in c)$, $\sum_c P(n \in c) = 1$. Each cluster c is associated to a cluster *centroid* p_c, which is discrete density over V obtained by averaging appropriately the p_n.

Measuring Distributional Similarity

We will work here with a measure of distributional *dissimilarity* rather than similarity. The dissimilarity between two nouns n and n' will be simply the relative entropy (Kullback-Leibler distance) of the corresponding densities

$$D(p_n \parallel p_{n'}) = \sum_v p_n(v) \ln \frac{p_n(v)}{p_{n'}(v)}$$

This is a well known measure of dissimilarity between densities, which is zero just in case the densities are identical and increases as the likelihood of the first density being an empirical sample drawn according to the second density decreases. In information-theoretic terms, $D(f \parallel f')$ measures how inefficient on average it would be to use a code based on f' to encode a variable distributed according to f. With respect to our problem, $D(p_n \parallel p_c)$ thus gives us the loss of information in using cluster centroid p_c instead of the actual distribution for word p_n when modeling the distributional properties of n.

One technical difficulty is that $D(f \parallel f')$ is not defined (infinite) when $f'(x) = 0$ but $f(x) > 0$. We could sidestep this problem (as we did initially) by smoothing appropriately zero frequencies [Church and Gale, 1991]. However, this is not very satisfactory because one of the main goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes. It turns out that the problem is avoided by our clustering technique, which never needs to compute the dissimilarity between individual word distributions, but only between a word distribution and average distributions (cluster centroids) that are guaranteed to be nonzero whenever the word distributions are (except in certain numeric underflow situations). This is an important advantage of our method compared with aglomerative clustering techniques that need to compare individual objects being considered for grouping.

Statistical Mechanics and Clustering

Our clustering method is an adaptation of a technique by Rose *et al.* [1990]. The basic idea is to construct each cluster centroid p_c as an average of the data densities p_n for individual nouns weighted by $\exp -\beta D(p_n \parallel p_c)$ where β is a scale parameter analogous to the inverse of a "temperature". Intuitively, the higher β (the lower the temperature), the more local is the influence of each data point on the definition of centroids. The dissimilarity plays here the role of distortion. When the scale parameter β is close to zero, the dissimilarities are almost irrelevant, all words con-

number	0.999	number	1.429	speed	1.177	speed	1.130	velocity	1.216
material	1.361	diversity	1.537	level	1.315	zenith	1.214	percent	1.338
variety	1.401	structure	1.577	velocity	1.371	depth	1.244	m	1.399
								zenith	0.862
								height	0.903
								depth	0.964
						number	1.461		
						concentration	1.478		
						strength	1.488		
				change	1.561	pollution	1.187	problem	1.109
				failure	1.562	failure	1.290	challenge	1.373
				variation	1.592	increase	1.328	crisis	1.420
								pollution	0.996
								increase	1.103
								failure	1.132
						structure	1.371		
						relationship	1.460		
						aspect	1.492		
		number	1.026	number	1.047	number	1.120		
		material	1.093	comedy	1.060	variety	1.217		
		mass	1.252	essay	1.142	material	1.275		
						essay	0.695		
						comedy	0.800		
						poem	0.829		
				material	0.976				
				salt	1.217				
				ring	1.244				
state	1.320	residence	1.082	complex	1.161	complex	1.097	region	1.326
ally	1.458	state	1.102	network	1.175	network	1.211	lake	1.393
residence	1.473	conductor	1.213	community	1.276	lake	1.360	half	1.399
								complex	0.953
								ring	1.049
								nucleus	1.077
						navy	1.096		
						community	1.099		
						network	1.244		
				conductor	0.699	state	1.279	people	1.242
				vice-president	0.756	people	1.417	animal	1.404
				editor	0.814	modern	1.418	person	1.413
								dance	1.414
								system	1.418
								combination	1.421
						conductor	0.457		
						vice-president	0.474		
						director	0.489		
		grant	1.392	improvement	1.329	program	1.459	operation	1.116
		distinction	1.554	voyage	1.338	operation	1.478	study	1.157
		form	1.571	migration	1.428	study	1.480	investigation	1.181
								year	1.434
								rest	1.442
								day	1.464
						voyage	0.861		
						trip	0.972		
						progress	1.016		
				control	1.201	form	1.110		
				recognition	1.317	explanation	1.255		
				nomination	1.363	care	1.291		
						recognition	0.874		
						acclaim	1.026		
						renown	1.079		

Figure 2: Noun clusters for Grolier's Encyclopedia

tribute about equally to each centroid, and so the lowest average distortion solution involves just one cluster which is the average of all word densities. As β is slowly increased following an appropriate "annealing schedule", one eventually reaches a "phase transition" past which the natural solution involves two distinct centroids. The new centroids will in turn split as β increases again. For a given value of β, we have a most natural set of centroids, which are the leaves of a centroid splitting tree rooted at the single centroid at low β.

For each β value in the schedule, our clustering method attempts to minimize the free energy

$$F = -\beta^{-1} \sum_n \ln \sum_c \exp -\beta D(p_n \parallel p_c)$$

Differentiating with respect to each c under the constraint $\sum_v p_c(v) = 1$ to identify the critical points of F we obtain

$$p_c = \frac{\sum_n P(n \in c) p_n}{\sum_n P(n \in c)}$$

where $P(n \in c) = \exp -\beta D(p_n \parallel p_c)/Z_n$ and $Z_n = \sum_c \exp -\beta D(f_n \parallel p_c)$. This can be solved iteratively in a fashion analogous to the EM algorithm [Dempster *et al.*, 1977].

The annealing schedule for β is determined by repeated subdivision to try to pinpoint as closely as possible individual phase transitions.

Experiments

As a first experiment, we used our method to classify the 64 nouns appearing most frequently has heads of direct objects of the verb "fire" in one year (1988) of Associated Press newswire.[1] In this corpus, the chosen nouns appear as direct object heads of a total of 2147 distinct verbs. Thus, each noun is represented by a density over the 2147 verbs.

Figure shows the 5 words most similar to the each cluster centroid for the four clusters resulting from the first two cluster splits. It can be seen that first split separates the objects corresponding to the weaponry sense of "fire" (cluster 1) from the ones corresponding to the personnel action (cluster 2). The second split then further refines the weaponry sense into a projectile sense (cluster 3) and a gun sense (cluster 4), although that split is somewhat less sharp, possibly because not enough distinguishing contexts occur in the corpus.

Figure 2 shows the three closest nouns to the centroid of each of a set of hierarchical clusters derived from verb-object pairs involving the 1000 most frequent nouns in Grolier's Encyclopedia. [2] While we have not yet developed rigorous methods for evaluating such cluster sets, it is interesting to see how major semantic classes (locations, times, actions, physical structures, animated beings, scalar variables and extremes) emerge from the procedure.

Further Work

We are proceeding with the application of the algorithm to a wider variety of data sets, and experimenting with convergence and evaluation criteria for the clustering procedure. We plan to consider additional distributional relations (eg. adjective-noun) and to test the results of the algorithm as the class component of language models, particularly those based on stochastic lexicalized tree-adjoining grammars [Schabes, 1992].

Acknowledgments

We would like to thank Don Hindle for making available the 1988 Associated Press verb-object data set, the Fidditch parser and a verb-object structure filter, Mats Rooth for selecting the objects of "fire" data set and many discussions, Lillian Lee for developing suitable regular-expression search patterns and building the Grolier data set, and David Yarowsky for help with his stemming and concordancing tools.

References

Brown, Peter F.; Della Pietra, Vincent J.; deSouza, Peter V.; Lai, Jenifer C.; and Mercer, Robert L. 1990. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, Paris, France. 283–298.

Church, Kenneth W. and Gale, William A. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language* 5:19–54.

Church, Kenneth W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas. Association for Computational Linguistics, Morristown, New Jersey. 136–143.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.

[1] The verb-object pairs for this example were collected by Don Hindle using his parser Fidditch, and this particular subset selected by Mats Rooth.

[2] We used the June 1991 electronic version of Grolier's Encyclopedia (10 million words), tagged with part-of-speech and word stems using stochastic tagging and stemming tools developed by David Yarowsky and Ken Church. The verb-object pairs were extracted with regular-expression searches over the tagged corpus.

Hindle, Donald 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania. Association for Computational Linguistics, Morristown, New Jersey. 268–275.

Resnik, Philip 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-Based Natural-Language-Processing Techniques*, San Jose, California.

Rose, Kenneth; Gurewitz, Eitan; and Fox, Geoffrey C. 1990. Statistical mechanics and phase transitions in clustering. *Physical Review Letters* 65(8):945–948.

Schabes, Yves 1992. Stochastic lexicalized tree-adjoining grammars. In *Proceeedings of the 14th International Conference on Computational Linguistics*, Nantes, France.