LABELING A LIST WITH UNIQUE IDENTIFIERS

A. ROSS ECKLER
Morristown, New Jersey

Suppose that one wishes to replace words in a list by abbreviations in order to save space. If one allows abbreviations of variable length, a very simple rule is: use as many letters as needed to ensure that a word is not confused with any other word. For example, the states Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware... can be replaced by ALAB, ALAS, ARI, ARK, CA, COL, CON, D ...

However, if abbreviations must all be the same length, the problem becomes more interesting. One can, of course, simply use AA, AB, ... ZZ to abbreviate lists up to 676 in size, but if the list is much smaller it ought to be possible to relate the abbreviations to the words. In particular, one asks: is it possible to devise a set of abbreviations that can be unambiguously related to the words in the list by a single simple rule? For example, can one take the first two letters of each word, or the first and last letters of each word to deduce the words? Neither strategy works for the states, for AL represents both Alaska and Alabama, and AA represents Alaska, Alabama and Arizona.

In the April 2000 issue of Wordsworth, Ted Clarke lists the 40 Vehicle Registration Offices (VROs) for British automobiles, showing that the first-and-last-letter rule fails twice: LN can be either Lincoln or Luton, and PH, either Peterborough or Portsmouth.

> Aberdeen, Bangor, Beverley, Birmingham, Bournemouth, Brighton, Bristol, Cardiff, Carlisle, Chelmsford, Chester, Dundee, Edinburgh, Exeter, Glasgow, Inverness, Ipswich, Leeds, Lincoln, Luton, Maidstone, Manchester, Middlesbrough, Newcastle, Northampton, Norwich, Nottingham, Oxford, Peterborough, Portsmouth, Preston, Reading, Sheffield, Shrewsbury, Sidcup, Stanmore, Swansea, Truro, Wimbledon, Worcester

Is his near-success a fluke, a fortuitous property of this particular list? One can ask a more general question: for what list size is an unambiguous set of abbreviations likely to be found? This can be recognized as a linguistic version of the well-known "birthday problem" which states that in a group of 22 people there is a 50-50 chance that at least one pair will share the same month-and-day birthday. Since one is dealing with a 676-day "year" in the abbreviations problem, one might expect the corresponding group size for a 50-50 chance of success (that is, every two-letter abbreviation uniquely points to the corresponding word on the list) to be somewhat larger than 22. However, this is more than compensated for by the fact that abbreviations, unlike birthdays, are not equiprobable; some (BE, AN) are much more likely to turn up than others (QZ, JX).

Mike Keith has supplied a calculation based on the assumption that each letter is drawn at random from letters with English-language text probabilities, and that these independently combine to form abbreviations (for example, if the probability of E is 0.1 and of T is 0.08, then the probability of ET or TE is 0.008). A more precise calculation would take into account correlations (likely for adjacent letters), but these are probably not large enough to change the conclusions reached below.

The mathematically-challenged reader can skip this paragraph and move on to the conclusion. Let P(i) denote the probability of the letter i in text. The probability that two letters selected at random match is p(a)p(a) + p(b)p(b) + + p(z)p(z) = 0.065. Suppose we form n two-letter abbreviations. What is the probability that all n are distinct using some rule (like the ones above), meaning that we can reconstruct the list from the abbreviations? First we must calculate P, the probability that a selected pair of two-letter abbreviations is distinct. This is equal to

1 - prob(not distinct) = 1 - prob(first letter matches, second letter matches) =
1 - (0.065)(0.065) = 0.995775

(Since 1/(1-P) is about 237, we are doing the birthday problem with a 237-day year, not 676.) Now proceed as in the birthday problem. Number the n items 0,1,2,...n-1. The probability that the one numbered k is different from the k that precede it on the list is

1 - prob(it is the same as one of the preceding k)
1 - k (prob(it is the same as a selected one of them))
1 - k(1 - P)

The total probability that all n are distinct is the product of this expression as k goes from 1 to n-1. With the aid of a computer the probabilities for various n are:

1 to 10: 1.00, .996, .983, .967, .946, .922, .895, .865, .832, .797
11 to 20: .760, .721, .681, .641, .601, .560, .520, .480, .442, .404
21 to 30: .368, .334, .302, .271, .242, .216, .191, .169, .148, .129
31 to 40: .112, .097, .084, .072, .061, .052, .044, .037, .031, .025
41 to 50: .021, .017, .014, .012, .009, .008, .006, .005, .004, .003

If one has a list of size 18, the chance is 50-50 that an abbreviation scheme is workable. For a list of 40, there is only one chance in 40 that it will work, so one should not be surprised that the British VROs don't lead to a unique set of abbreviations.

However, there is still hope. What if one tries not just one but many possible abbreviation rules? If all words in the list are at least m letters long, one can form (1) abbreviations using the ith and jth letters, where i and j run from 1 through m, and (2) abbreviations using the ith letter from the beginning and the jth letter from the end of a word, with the same ranges. These yield $2m^2$ different rules (50 for the VRO). Suppose that all these abbreviation rules are uncorrelated with each other--that is, the probability that one rule works is independent of whether or not the others do. It is simple to calculate the probability that at least one rule out of the whole set will be successful:

prob(at least one rule is successful) = 1 - (prob(no rules are successful)
= 1 - [prob(rule i is not successful)]^{50}

which yields 0.75 for the VRO list. In fact, all such abbreviations failed.

However, Mike Keith extended the universe of possible abbreviations beyond the limitation imposed by the shortest word in the list. He proposed that when one reaches the end of a word, one continues to use the last possible abbreviation for that word (for LUTON, if one takes the first letter and the kth letter from the end, one successively forms the abbreviations LN LO LT LU LL LL LL ...) He was able to find a unique set of abbreviations for two of the generalized rules: fifth letter from start, seventh letter from end, and ninth letter from end, fourth letter from end.

Instead of constructing abbreviations anchored to letter-positions in the word, one can construct abbreviations that "float" through the word but whose parts maintain a fixed relationship to each other. The simplest abbreviation rule of this nature is to assign to each word a unique bigram contained in that word (unique, in the sense that it appears in none of the other words in the list). The shortest words on the list have the fewest bigrams and consequently are the ones most likely to generate abbreviations colliding with other words. In the VRO list, one quickly finds that in LEEDS LE collides with WimbLEdon, EE with AberdEEn, ED with WimblEDon again, and DS with MaiDStone, so one must reject the bigram rule. For bigrams interrupted by one extraneous letter, all five-letter and six-letter names pass muster, but the seven-letter CHESTER cannot be distinguished from MANCHESTER. So when checking for the bigram rule, one should first ask whether or not any word is contained in another.

The birthday model does not apply to bigram rules; in fact, no simple mathematical model appears to exist because one must tailor the calculation to the distribution of word lengths. Assuming Mike Keith's idealized model of independent letter frequencies, what is the probability that every one of the 40 names in the VRO list will contain a unique bigram? (For the answer, go to the last two sentences in the next paragraph.)

Begin with the previous value of 0.995775 for the probability that two randomly-selected bigrams will not match. What is the probability that all four of the bigrams in a typical five-letter name will match bigrams in the rest of the list? There are a total of 281 bigrams in the rest of the list, so the probability that a bigram will fail to match any of them is $(.995775)^{281} = 0.3042$. The probability that all four bigrams will match one of the 281 others is $(1 - .3042)^4 = 0.234$; thus, a five-letter name has a probability of 0.766 of generating one or more unique bigrams. For words of other lengths the last equation must be raised to higher powers; for lengths of five through thirteen, the corresponding probabilities are 0.886, 0.921, 0.945, 0.962, 0.973. 0.982 and 0.987. The number of names of length five through thirteen in the VRO list is 3,5,10,5,7,6,2,1,1; multiplying the above probabilities the appropriate number of times, one finds that the probability that all 40 names have unique bigrams is only 0.018, little different from the 0.025 calculated earlier. In other words, there is little hope that floating abbreviations will do better than anchored ones, at least as far as the VRO list is concerned.