Semantic Categorization Using Simple Word Co-occurrence Statistics

John A. Bullinaria
School of Computer Science, University of Birmingham
Edgbaston, Birmingham, B15 2TT, UK
j.bullinaria@physics.org

Abstract

This paper presents a series of new results on corpus derived semantic representations based on vectors of simple word co-occurrence statistics, with particular reference to word categorization performance as a function of window type and size, semantic vector dimension, and corpus size. A number of outstanding problems and difficulties with this approach are identified and discussed.

1 Introduction

There is now considerable evidence that simple word co-occurrence statistics from large text corpora can capture certain aspects of word meaning (e.g., Lund & Burgess, 1996; Landauer & Dumais, 1997; Bullinaria & Levy, 2007). This is certainly consistent with the intuition that words with similar meaning will tend to occur in similar contexts, but it is also clear that there are limits to how far this idea can be taken (e.g., French & Labiouse, 2002). The natural way to proceed is to optimize the standard procedure as best one can, and then identify and solve the problems that remain.

To begin that process, Bullinaria & Levy (2007) presented results from a systematic series of experiments that examined how different statistic collection details affected the performance of the resultant co-occurrence vectors on a range of semantic tasks. This included varying the nature of the 'window' used for the co-occurrence counting (e.g., type, size), the nature of the statistics collected (e.g., raw conditional probabilities, pointwise mutual information), the vector space dimensionality (e.g., using only the d highest frequency context words), the size and quality of the corpus (e.g., professionally created corpus, news-group text), and the semantic distance measure used (e.g., Euclidean, City-block, Cosine, Hellinger, Bhattacharya, Kulback-Leibler). The resultant vectors were subjected to a series of test tasks: a standard multiple choice TOEFL test (Landauer & Dumais, 1997), a larger scale semantic distance comparison task (Bullinaria & Levy, 2007), a semantic categorization task (Patel et al., 1997), and a syntactic categorization task (Levy et al., 1998). It was found that the set-up producing the best results was remarkably consistent across all the tasks, and that involved using Positive Pointwise Mutual Information (PPMI) as the statistic to collect, very small window sizes (just one context word each side of the target word), and the standard Cosine distance measure (Bullinaria & Levy, 2007).

That study was primarily conducted using a 90 million word untagged corpus derived from the BNC (Aston & Burnard, 1998), and most of the results presented could be understood in terms of the quality or reliability of the various vector components collected from it: Larger windows will tend to contain more misleading context, so keeping the window small is advantageous. Estimations of word co-occurrence probabilities will be more accurate for higher frequency words, so one might expect that using vector components that correspond to low frequency context words would worsen the performance rather than enhance it. That is true if a poorly chosen statistic or distance measure is chosen, but for PPMI and Cosine it seems that more context dimensions lead to more useful information and better performance. For smaller corpora, that remains

true, but then larger windows lead to larger counts and better statistical reliability, and that can improve performance (Bullinaria & Levy, 2007). That will be an important issue if one is interested in modeling human acquisition of language, as the language streams available to children are certainly in that regime (Landauer & Dumais, 1997; Bullinaria & Levy, 2007). For more practical applications, however, much larger and better quality corpora will certainly lead to better results, and the performance levels are still far from ceiling even with the full BNC corpus (Bullinaria & Levy, 2007).

The aim of this paper is to explore how the results of Bullinaria & Levy (2007) extend to the ukWaC corpus (Ferraresi, 2007) which is more than 20 times the size of the BNC, and to test the resultant semantic representations on further tasks using the more sophisticated clustering tool CLUTO (Karypis, 2003). The next section will describe the methodology in more detail, and then the word categorization results are presented that explore how the performance varies as a function of window size and type, vector representation dimensionality, and corpus size. The paper ends with some conclusions and discussion.

2 Methodology

The basic word co-occurrence counts are the number of times in the given corpus that each context word c appears in a window of a particular size s and type w (e.g., to the left/right/left+right) around each target word t, and from these one can easily compute the conditional probabilities $p(c|t)$. These actual probabilities can then be compared with the expected probabilities $p(c)$, that would occur if the words were distributed randomly in the corpus, to give the Pointwise Mutual Information (PMI):

$$I(c,t) = \log \frac{p(c|t)}{p(c)} \qquad (1)$$

(Manning & Schutze, 1999, Sect. 5.4). Positive values indicate that the context words occur more frequently than expected, and negative values correspond to less than expected. The study of Bullinaria & Levy (2007) showed that setting all the negative values to zero, leaving the Positive Pointwise Mutual Information (PPMI), reliably gave the best performing semantic vectors across all the semantic tasks considered, if the standard Cosine distance measure was used. Exactly the same PPMI Cosine approach was used for all the investigations here. The window type and size, and the number of frequency ordered context word dimensions, were allowed to vary to explore their effect on the results.

The raw ukWaC corpus (Ferraresi, 2007) was first preprocessed to give a plain stream of about two billion untagged words, containing no punctuation marks apart from apostrophes. Then the list of potential target and context words contained within it was frequency ordered and truncated at one million words, at which point the word frequency was just five occurrences in the whole corpus. This process then allowed the creation of a one million dimensional vector of PPMI values for each target word of interest. The full corpus was easily split into disjoint subsets to explore the effect of corpus size.

The quality of the resultant semantic vectors was tested by using them as a basis for clustering the sets of nouns and verbs specified for the Lexical Semantics Workshop at ESSLLI 2008. Vector representations for the n words in each word-set were clustered using the CLUTO Clustering Toolkit (Karypis, 2003), with the direct k-way clustering algorithm and default settings. The quality of clustering was established by comparison against hand-crafted category labels using standard quantitative measures of *entropy* E and *purity* P, defined as weighted averages over the cluster entropies E_r and purities P_r:

$$E = \sum_{r=1}^{k} \frac{n_r}{n} E_r \;\; , \;\; E_r = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \qquad (2)$$

$$P = \sum_{r=1}^{k} \frac{n_r}{n} P_r \;\; , \;\; P_r = \frac{1}{n_r} \max_i (n_r^i) \qquad (3)$$

where n_r and n_r^i are the numbers of words in the relevant clusters and classes, with r labelling the k clusters, and i labelling the q classes (Zhao & Karypis, 2001). Both measures range from 0 to 1, with 1 best for purity and 0 best for entropy.

3 Results

It is convenient to start by looking in Figure 1 at the results obtained by instructing the clustering algorithm to identify six clusters in the semantic vectors generated for a set of 44 concrete nouns. The six

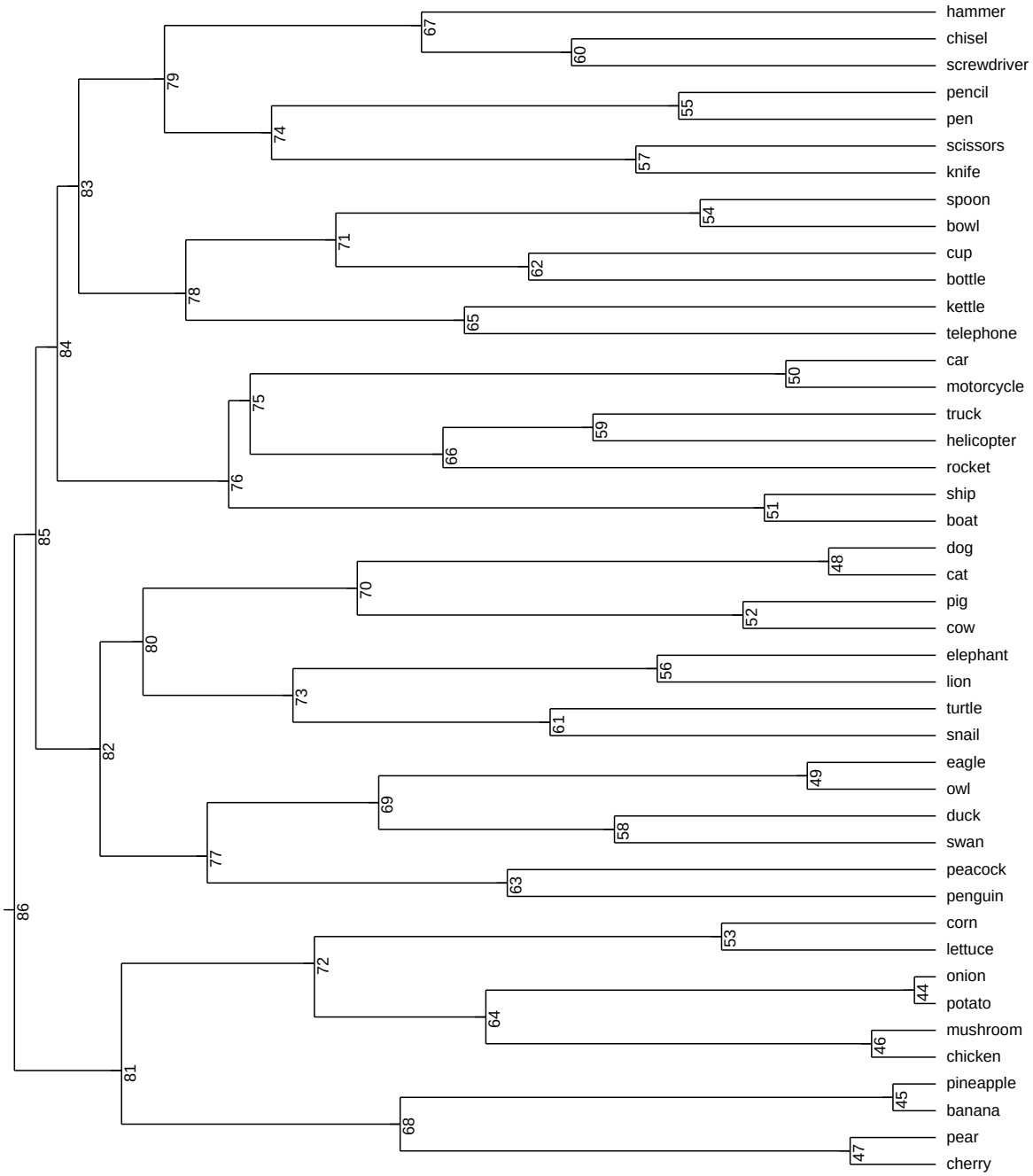

Figure 1: Noun categorization cluster diagram.

hand-crafted categories {‘birds’, ‘ground animals’, ‘fruits’, ‘vegetables’, ‘tools,’ ‘vehicles’} seem to be identified almost perfectly, as are the higher level categories {‘animals’, ‘plants’, ‘artifacts’} and {‘natural’, ‘artifact’}. The purity of the six clusters is 0.886 and the entropy is 0.120. Closer inspection shows that the good clustering persists right down to individual word pairs. The only discrepancy is ‘chicken’ which is positioned as a ‘foodstuff’ rather than as an ‘animal’, which seems to be no less acceptable than the “correct” classification.

Results such as these can be rather misleading, however. The six clusters obtained do not actually line up with the six hand-crafted clusters we were looking for. The ‘fruit’ and ‘vegetable’ clusters are combined, and the ‘tools’ cluster is split into two.

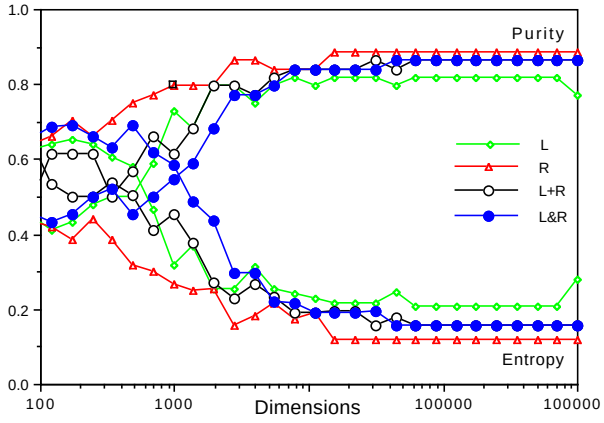

Figure 2: The effect of vector dimensionality on noun clustering quality.

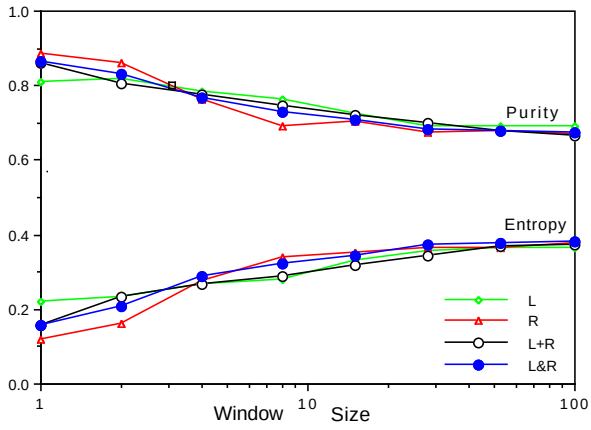

Figure 3: The effect of window size on noun clustering quality.

This contributes more to the poor entropy and purity values than the misplaced 'chicken'. If one asks for seven clusters, this does not result in the splitting of 'fruit' and 'vegetables', as one would hope, but instead creates a new cluster consisting of 'turtle', 'snail', 'penguin' and 'telephone' (which are outliers of their correct classes), which ruins the nice structure of Figure 1. Similarly, asking for only three clusters doesn't lead to the split expected from Figure 1, but instead 'cup', 'bowl' and 'spoon' end up with the plants, and 'bottle' with the vehicles. It is clear that either the clusters are not very robust, or the default clustering algorithm is not doing a particularly good job. Nevertheless, it is still worth exploring how the details of the vector creation process affect the basic six cluster clustering results.

The results shown in Figure 1, which were the best obtained, used a window of just one context word to the right of the target word, and the full set of one million vector dimensions. Figure 2 shows how reducing the number of frequency ordered context dimensions and/or changing the window type affects the clustering quality for window size one. The results are remarkably consistent down to about 50,000 dimensions, but below that the quality falls considerably. Windows just to the right of the target word (R) are best, windows just to the right (L) are worst, while windows to the left and right (L+R) and vectors with the left and right components separate (L&R) come in between. Increasing the window size causes the semantic clustering quality to deteriorate as seen in Figure 3. Large numbers of dimensions remain advantageous for larger windows, but the best window type is less consistent.

That large numbers of dimensions and very small window sizes are best is exactly what was found by Bullinaria & Levy (2007) for their semantic tasks using the much smaller BNC corpus. There, however, it was the L+R and L&R type windows that gave the best results, not the R window. Figure 4 shows how the clustering performance for the various window types varies with the size of corpus used, with averages over distinct sub-sets of the full corpus and the window size kept at one. Interestingly, the superiority of the R type window disappears around the size of the BNC corpus, and below that the L+R and L&R windows are best, as was found previously. The differences are small though, and often they correspond to further use of different valid semantic categories rather than "real errors", such as clustering 'egg laying animals' rather than 'birds'. Perhaps the most important aspect of Figure 4, however, is that the performance levels still do not appear to have reached a ceiling level by two billion words. It is quite likely that even better results will be obtainable with larger corpora.

While the PPMI Cosine approach identified by Bullinaria & Levy (2007) produces good results for nouns, it appears to be rather less successful for verb clustering. Figure 5 shows the result of attempting five-way clustering of the verb set vectors obtained in exactly the same way as for the nouns above. No reliably better results were found by changing the window size or type or vector dimen-

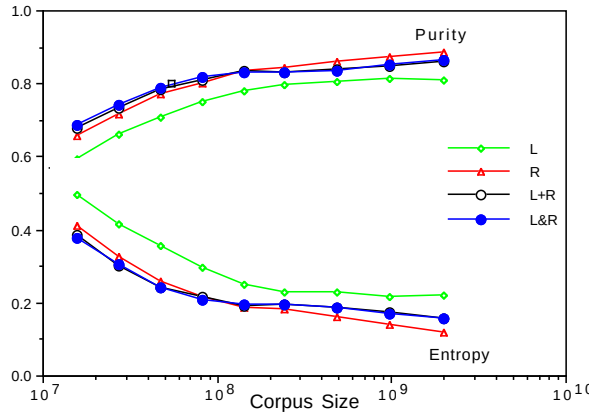

Figure 4: The effect of corpus size on noun clustering quality.

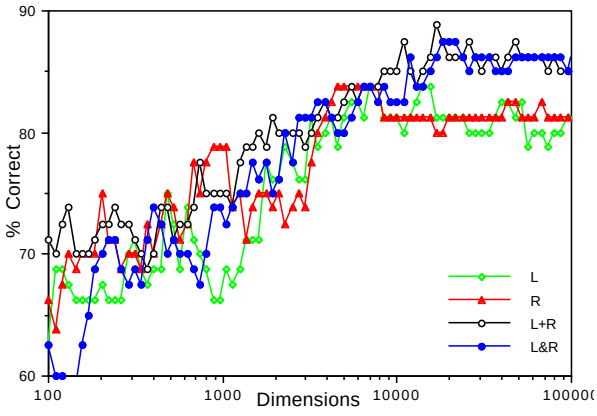

Figure 6: The effect of vector dimensionality on TOEFL performance.

sionality. There is certainly a great deal of semantic validity in the clustering, with numerous appropriate word pairs such as 'buy, sell', 'eat, drink', 'kill, destroy', and identifiable clusters such as those that might be called 'body functions' and 'motions'. However, there is limited correspondence with the five hand crafted categories {'cognition', 'motion', 'body', 'exchange', 'change-state'}, resulting in a poor entropy of 0.527 and purity only 0.644.

Finally, it is worth checking how the larger size of the ukWaC corpus affects the results on the standard TOEFL task (Landauer & Dumais, 1997), which contains a variety of word types. Figure 6 shows the performance as a function of window type and number of dimensions, for the optimal window size of one. Compared to the BNC based results found by Bullinaria & Levy (2007), the increased corpus size has improved the performance for all window types, and the L+R and L&R windows continue to work much better than R or L windows. It seems that, despite the indications from the above noun clustering results, it is not true that R type windows will always work better for very large corpora. Probably, for the most reliably good overall performance, L&R windows should be used for all corpus sizes.

4 Conclusions and Discussion

It is clear from the results presented in the previous section that the simple word co-occurrence counting approach for generating corpus derived semantic representations, as explored systematically by Bullinaria & Levy (2007), works surprisingly well in some situations (e.g., for clustering concrete nouns), but appears to have serious problems in other cases (e.g., for clustering verbs).

For the verb clustering task of Figure 5, there is clearly a fundamental problem in that the hand-crafted categories correspond to just one particular point of view, and that the verb meanings will be influenced strongly by the contexts, which are lost in the simple co-occurrence counts. Certainly, the more meanings a word has, the more meaningless the resultant average semantic vector will be. Moreover, even if a word has a well defined meaning, there may well be different aspects of it that are relevant in different circumstances, and clustering based on the whole lot together will not necessarily make sense. Nor should we expect the clustering to match one particular set of hand crafted categories, when there exist numerous equally valid alternative ways of doing the categorization. Given these difficulties, it is hard to see how any pure corpus derived semantic representation approach will be able to perform much better on this kind of clustering task.

Discrepancies amongst concrete nouns, such as the misplaced 'chicken' in Figure 1, can be explored and understood by further experiments. Replacing 'chicken' by 'hen' does lead to the correct 'bird' clustering alongside 'swan' and 'duck'. Adding 'pork' and 'beef' into the analysis leads to them being clustered with the vegetables too, in a 'foodstuff' category, with 'pork' much closer to 'beef' and 'potato' than to 'pig'. As we already saw with the verbs above, an inherent difficulty with testing

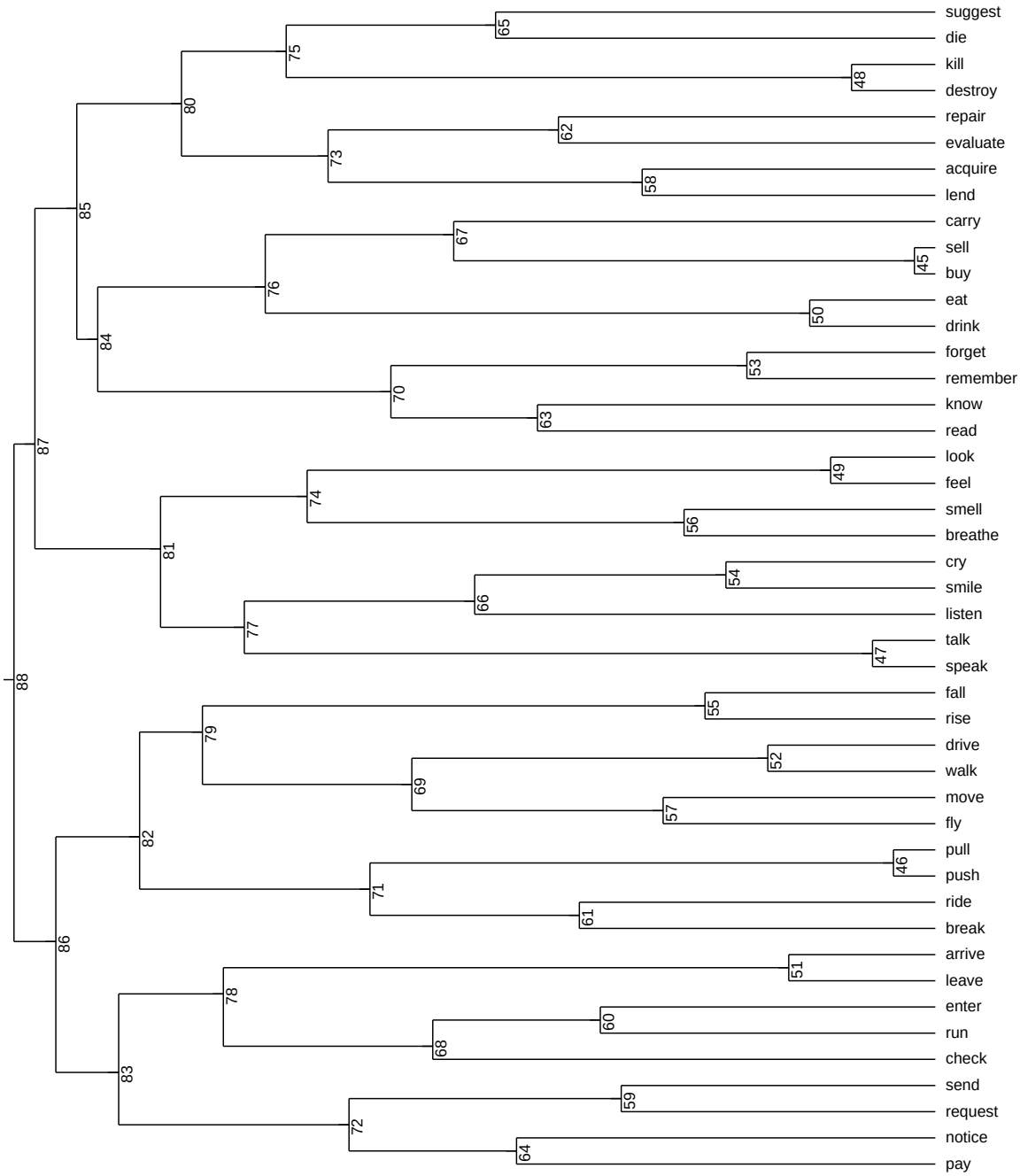

Figure 5: Verb categorization cluster diagram.

semantic representations using any form of clustering is that words can be classified in many different ways, and the appropriate classes will be context dependent. If we try to ignore those contexts, either the highest frequency cases will dominate (as in the 'foodstuff' versus 'animal' example here), or merged representations will emerge which will quite likely be meaningless.

There will certainly be dimensions or sub-spaces in the semantic vector space corresponding to particular aspects of semantics, such as one in which 'pork' and 'pig' are more closely related than 'pork' and 'potato'. However, as long as one only uses simple word co-occurrence counts, those will not be easily identifiable. Most likely, the help of some form of additional supervised learning will be re-

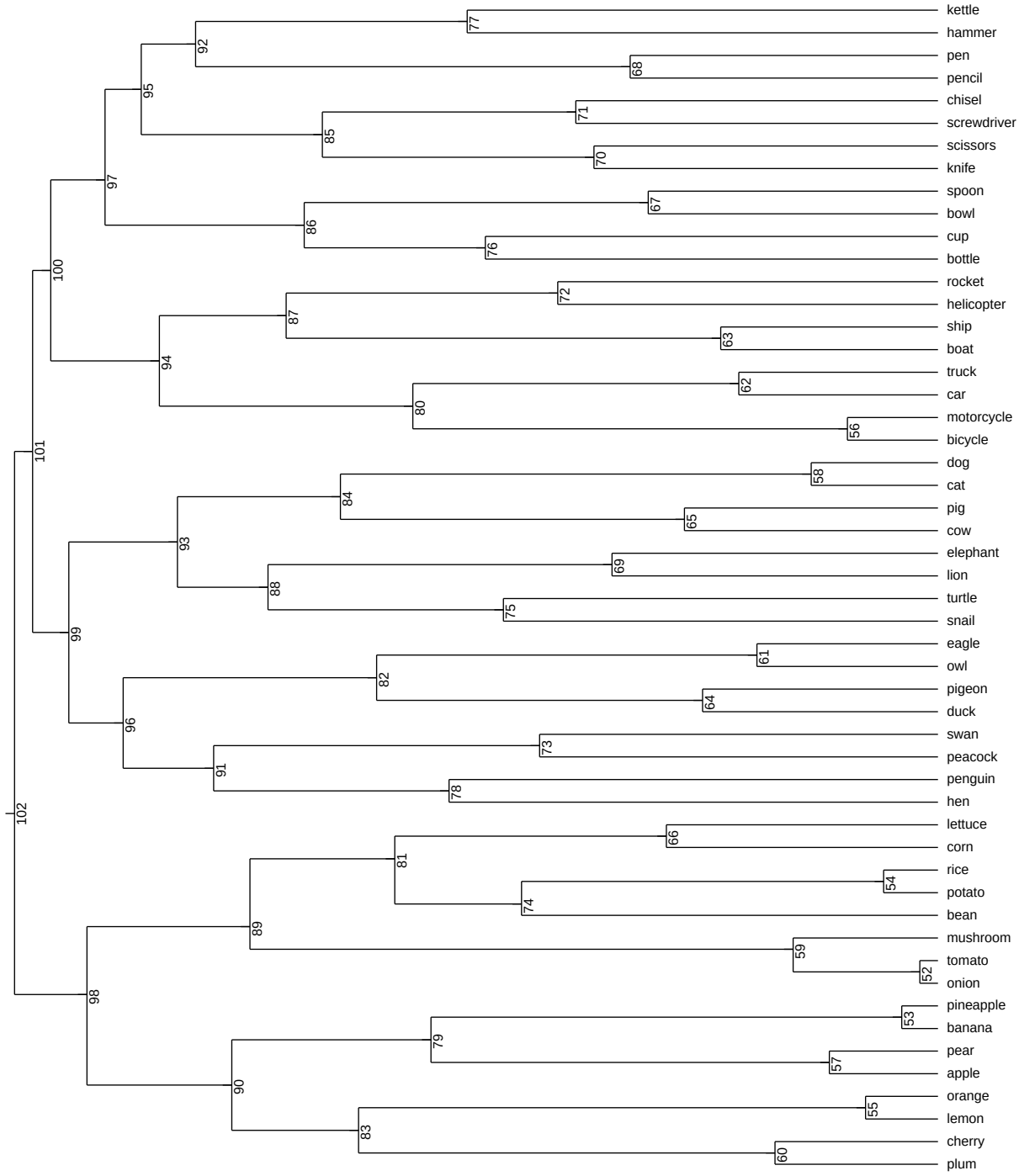

Figure 7: Extended noun categorization cluster diagram.

quired (Bullinaria & Levy, 2007). For example, appropriate class-labelled training data might be utilized with some form of Discriminant Analysis to identify distinct semantic dimensions that can be used as a basis for performing different types of classification that have different class boundaries, such as 'birds' versus 'egg laying animals'. Alternatively, or additionally, external semantic information sources, such as dictionaries, could be used by some form of machine learning process that separates the merged representations corresponding to word forms that have multiple meanings.

Another problem for small semantic categorization tasks, such as those represented by Figures 1 and 5, is that with so few representatives of each hand-crafted class, the clusters will be very sparse compared to the "real" clusters containing all possible class members, e.g. all 'fruits' or all 'birds'. With poorly chosen word sets, class outliers can easily fall in the wrong cluster, and there may be stronger clustering within some classes than there are between other classes. This was seen in the overly poor entropy and purity values returned for the intuitively good clustering of Figure 1.

In many ways, there are two separate issues that both need to be addressed, namely:

1. If we did have word forms with well defined semantics, what would be the best approach for obtaining corpus derived semantic representations?

2. Given that best approach, how can one go on to deal with word forms that have more than one meaning, and deal with the multidimensional aspects of semantics?

The obvious way to proceed with the first issue would be to develop much larger, less ambiguous, and more representative word-sets for clustering, and to use those for comparing different semantic representation generation algorithms. A less computationally demanding next step might be to persevere with the current small concrete noun clustering task of Figure 1, but remove the complications such as ambiguous words (i.e. 'chicken') and class outliers (i.e. 'telephone'), and add in extra words so that there is less variation in the class sizes, and no classes with fewer than eight members. For the minimal window PPMI Cosine approach identified by Bullinaria & Levy (2007) as giving the best general purpose representations, this leads to the perfect (entropy 0, purity 1) clustering seen in Figure 7, including "proof" that 'tomato' is (semantically, if not scientifically) a vegetable rather than a fruit. This set could be regarded as a preliminary clustering challenge for any approach to corpus derived semantic representations, to be conquered before moving on to tackle the harder problems of the field, such as dealing with the merged representations of homographs, and clustering according to different semantic contexts and criteria. This may require changes to the basic corpus approach, and is likely to require inputs beyond simple word co-occurrence counts.

References

Aston, G. & Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Bullinaria, J.A. & Levy, J.P. 2007. Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, **39**, 510-526.

Ferraresi, A. 2007. Building a Very Large Corpus of English Obtained by Web Crawling: ukWaC. Masters Thesis, University of Bologna, Italy. Corpus web-site: http://wacky.sslmit.unibo.it/

French, R.M. & Labiouse, C. 2002. Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*, 316-322. Mahwah, NJ: Lawrence Erlbaum Associates.

Karypis, G. 2003. CLUTO: A Clustering Toolkit (Release 2.1.1). Technical Report: #02-017, Department of Computer Science, University of Minnesota, MN 55455. Software web-site: http://glaros.dtc.umn.edu/gkhome/views/cluto.

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.

Levy, J.P., Bullinaria, J.A. & Patel, M. 1998. Explorations in the Derivation of Semantic Representations from Word Co-occurrence Statistics. *South Pacific Journal of Psychology*, **10**, 99-111.

Lund, K. & Burgess, C. 1999. Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments & Computers*, **28**, 203-208.

Manning, C.D. & Schutze, H. 1996. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Patel, M., Bullinaria, J.A. & Levy, J.P. 1997. Extracting Semantic Representations from Large Text Corpora. In J.A. Bullinaria, D.W. Glasspool & G. Houghton (Eds.), *Fourth Neural Computation and Psychology Workshop: Connectionist Representations*,199-212. London: Springer.

Zhao, Y. & Karypis, G. 2001. Criterion Functions for Document Clustering: Experiments and Analysis. Technical Report TR #01–40, Department of Computer Science, University of Minnesota, MN 55455. Available from: http://cs.umn.edu/karypis/publications.