Proceedings of the Fourth Neural Computation and Psychology Workshop, pp. 199-212. London: Springer-Verlag, 1997.

Extracting Semantic Representations from Large Text Corpora

Malti Patel
Department of Computing, Macquarie University
Sydney, Australia

John A. Bullinaria & Joseph P. Levy
Department of Psychology, Birkbeck College
London, UK

Abstract

Many connectionist language processing models have now reached a level of detail at which more realistic representations of semantics are required. In this paper we discuss the extraction of semantic representations from the word co-occurrence statistics of large text corpora and present a preliminary investigation into the validation and optimisation of such representations. We find that there is significantly more variation across the extraction procedures and evaluation criteria than is commonly assumed.

1 Introduction

How to represent semantics has been a difficult problem for many years, and as yet there is no consensus as to exactly what is stored and how. With the rise of cognitive modelling, the problem of representing semantic information must now be addressed if any headway is to be made. Although semantics obviously plays a very important role in language, cognitive models concerned with language have either not attempted to implement this component [2, 20], or implemented it only on a small-scale [3, 4, 6, 7, 16, 17, 18]. If the experimental results from tasks such as reading and lexical decision are to be simulated, there must be serious investigations into how semantics can be represented on a large scale, e.g. for thousands of words.

Recently, work has begun on using large corpora to extract semantic information in the form of vectors of word co-occurrence statistics. In this paper, we shall discuss the results obtained from a preliminary study of extracting co-occurrence vectors from the British National Corpus (BNC) – a large corpus consisting of 100 million words, both written and spoken [9]. These vectors are obtained by counting how often words occur near each other in a corpus to give a vector of probabilities for each word with components corresponding to the different words in the corpus. There are a number of parameters which specify the vector creation process and their values will affect the resultant vectors. We describe some simple evaluation procedures with the aim of optimising these parameters to give the best semantic representations.

This kind of analysis seeks to investigate the degree to which aspects of the meaning of a word are reflected in its tendency to occur around certain other words.

This may give insights into how semantics may be learnt by humans through exposure to language and stored in the brain [8]. These vectors will also be of great use for representing semantics in models of various psychological processes, such as reading and lexical decision. The current methods are somewhat inadequate since the semantic representations are randomly generated or hand-crafted. Randomly generated vectors clearly have no relation to real semantics. Hand-crafted representations are subjective in that the modellers concerned decide how the meaning of a word, and also what features of the meaning, should be stored. For example, different people will have different ideas on how to best represent the meaning of *dog*. Also, creating semantic representations for thousands of words would be a time-consuming task. Hence, we need a technique which captures meanings in an objective fashion and one which would allow us to easily create semantic representations for many words.

We shall first briefly describe work already performed in this area of corpus analysis and then describe how semantics has been represented previously in various psychological models and why the corpus based approach for obtaining semantic representations has advantages over these. Our main focus will be on how the various parameters involved in the corpus analysis can be optimised to produce the best co-occurrence vectors. This will hopefully lead to a greater awareness of which parameter values give what types of results. In the process, we will define evaluation procedures which can be used by other researchers working on corpus analysis.

2 Previous Corpora Work

Various relevant results have already been obtained from corpus analysis [8, 11, 12, 19]. Lund and Burgess [11, 12], for example, derived co-occurrence vectors with 200 components from a 160 million word corpus, based on words occurring within a weighted window of ten words around the target word. Amongst other things, their analyses showed that vectors derived for semantically related words tended to be closer in Euclidean space than was the case for semantically unrelated words, e.g. the semantic vector for *cat* was closer to other co-occurrence vectors representing animals, such as *lion*, than to vectors representing body parts, such as *ankle*. Schütze has carried out numerous experiments on extracting semantics from corpora. His initial work [19] involved creating co-occurrence vectors from letter four-grams as opposed to words. He showed that semantically related vectors tended to be close in distance and demonstrated successful semantic disambiguation. Together, these investigations have indicated that useful semantic representations can be produced from corpora. Moreover, Bullinaria and Huckle [5] have already used semantic vectors of this form with some success in connectionist models of lexical decision.

Although these studies have shown this approach to be useful, no systematic and rigorous evaluations have yet been performed. There are a variety of parameters which specify how the co-occurrence vectors are created, for example, different window shapes and sizes, different numbers of vector components, different corpora sizes, and so on. Here we shall create co-occurrence vectors for the same groups of words for different parameter values. These vectors will be then be evaluated using two different criteria to optimise the parameter values and assess how good the best resultant semantic representations really are.

3 Implementation of the Semantic Component of Reading and Lexical Decision Models

Psychological models of reading and lexical decision have been implemented using neural networks with varying degrees of success [2, 3, 4, 7, 16, 17, 18, 20]. A major problem has been in implementing the semantic component of such models since there is no established theory of what should be represented or how. Modellers have tended towards using simple notions of semantic micro-feature representations as a practical way of implementing the lexical semantics of small sets of words. For example, Hinton & Shallice [7] generated their own semantic micro-features by hand such that each stood for a specific concept such as *has-legs* or *indoors*. A semantic vector consisted of 30 components with each representing one semantic micro-feature. Similarly, Plaut & Shallice [17], used 86 semantic micro-features split into categories such as *visual characteristics*, *where found*, etc. Others have shown that realistic patterns of performance can be obtained simply by using randomly generated semantic representations, for example, both Plaut [16] and Bullinaria [3] in their lexical decision models and Bullinaria [4] in his reading model.

A somewhat different approach investigated by Patel [15] involved using WordNet definitions to represent semantics. WordNet is a dictionary based on psycholinguistic principles, developed at Princeton University by Miller et al. [13], that contains approximately 57,000 nouns, 21,000 verbs and 19,500 adjectives. For each word, WordNet gives all possible meanings in terms of a number of definitions. For example, the WordNet representation for Sense 2 of *hand* is:

```
HAND : Sense 2 : hired hand, hand, hired man -- (a hired laborer on a farm or ranch)

=>  laborer, manual laborer, labourer -- (works with hands)
   => workman, working man, working person
    => employee -- (a worker who is hired to perform a job)
      => worker -- (a person who has employment)
         => person, individual, someone, man, mortal, human, soul -- (a human being)
            => life form, organism, being, living thing -- (any living entity)
               => entity -- (something having concrete existence; living or nonliving)
```

In the semantic vectors developed by Patel, each component corresponded to one WordNet definition. The component was *on* if the meaning contained that definition otherwise it remained *off*. Although, some promising results were obtained with this approach, problems did occur occasionally with the total number of definitions used to define a meaning. In some cases for polysemous words, the wrong meaning had a higher activation that the correct meaning simply because it consisted of more WordNet definitions than the correct meaning. Hence, for these cases, the more components a vector had *on*, the greater advantage the corresponding meaning had of gaining more activation.

The appealing factor of the above vector based approaches is that they are simple and intuitive. However, in the long-term, they have no external validity, except perhaps the WordNet approach which is at least based on psycholinguistic data. The corpus based approaches for representing semantics may prove to be better if, after rigorous evaluation, it can be shown that co-occurrence vectors do have some interesting and psychologically realistic properties. A technique will then have been found which has many advantages over the usual hand-crafted approach towards semantic representation. For example, it does not rely on subjective judgements, it is automatic and it produces data that are derived from genuine linguistic performance. These co-occurrence vectors could then be used reliably in the semantic components of connectionist language processing models.

4 Optimising the Vector Creation Parameters

The semantic vectors derived from corpus analysis are produced by simply counting the occurrences of neighbouring words, e.g. by counting the number of occurrences of the context words which neighbour *flower* to create the semantic vector for *flower*. There are clearly a number of parameters that need to be specified to uniquely determine this counting and vector creation process. In this section we briefly discuss five of the main parameters that we wish to vary in this preliminary study.

4.1 The Vector Creation Process

We begin with a simple illustration to show the roles the various parameters play in creating the vectors. Suppose we are producing the semantic vector for the word *girl* using a *window size* of two words on either side of the *target* word *girl*. Then suppose that the phrase "the little girl said that ..." is the next one to appear in the corpus. The values below show the increments that will be given to the already accumulated frequency counts of these words. For a *rectangular* window, the current total for each word around *girl* will be incremented by one, whereas for a *triangular* window, the increment is larger the closer the word is to *girl*.

- *Rectangular Window* (each word carries the same weight)

the	little	girl	said	that
1	1	0	1	1

- *Triangular Window* (closer words carry more weight)

the	little	girl	said	that
1	2	0	2	1

Then, how we use these increments depends on the *window types*:

- *Left only* - count words to left of target, e.g. "the little".

- *Right only* - count words to right of target, e.g. "said that".

- *Left plus Right* - count words on both sides of target, e.g. "the little said that".

- *Left and Right* - concatenate left only and right only vectors from above.

The final value of these frequency counts will then be used to calculate the co-occurrence vector for *girl*, first by normalising to take account of the total window size, and then dividing by the target word frequencies to give the probabilities of co-occurrence. We can now look at the main parameters in more detail.

4.2 The Main Parameters

Window Size

This defines the number of neighbouring words that we count as occurring "near" to the target word, e.g. do we count the two words immediately next to it, or the five words next to it, or the fifty words next to it, etc. One might conjecture that a large window size gives more semantic information whereas a small window size gives more syntactic information.

Window Type

This refers to which side of the target word we count the neighbouring words:

- *Left only* - count only words occurring to the left of the target word, producing vectors with one component for each of the D different words in the corpus.
- *Right only* - count only words occurring to the right of the target word, again producing vectors with D components.
- *Left plus Right* - count words occurring to the left and right of the target word, still producing vectors with D components.
- *Left and right* - concatenate the vectors formed by looking at just the left and right sides, i.e. the vectors from left only and right only, producing vectors with 2 × D components.

We shall not investigate the possibility here, but one might also wish to consider treating the left and right contexts asymmetrically.

Window Shape

It might be appropriate to treat the context words differently depending on how far away they are from the target words, so we have windows of different shape:

- *Rectangular/Flat* - each neighbouring word around the target word is given the same weight.
- *Triangular/Weighted* - as a neighbouring word gets further from the target word, it is given linearly less weight - a technique used by Lund and Burgess [8].

and one can imagine other possibilities that we shall not consider in this paper.

Number of Vector Components

Clearly we generally do not want to use all D components of our vectors, because D will be a very large number and the resultant vectors would be too large and very difficult to process. Hence some analysis must be carried out to determine how many vector components are appropriate to obtain the best results, e.g. does restricting ourselves to 100 components give better results than, say, using 1000 components? In this paper we shall use the components corresponding to the context words of the highest frequency. In future work we shall need to consider if it is more appropriate to use the components with the highest variance, or if we should use something like principal component analysis to reduce the dimension of the space.

Corpus Size

We would expect the vectors produced from a large corpus to be better than those produced from a smaller one, simply because the relative noise in the frequency counts will fall with the counts themselves and these will clearly increase with the corpus size. We need to determine how crucial this factor is and how it depends on the frequencies of the target and context words and on the evaluation criteria.

5 Evaluation Criteria

We thus have five main parameters whose values can be varied. Obviously, changing these values will produce differing co-occurrence vectors for the same set of words. Hence, evaluation techniques need to be formulated to decide which parameter values give the best set of co-occurrence vectors and how good these best vectors really are. In this section we begin a systematic investigation by describing two simple criteria for evaluating the different sets of co-occurrence vectors and in the next section we present some preliminary results.

The most natural conjecture is that if we define some distance metric on our semantic vector space, then the vectors corresponding to semantically unrelated words should be further apart than those for related words. Since there are numerous normalisation artefacts that may arise when we compare vectors derived using different values of the above parameters (e.g. different amounts of baseline noise), the natural dimension free quantity to compare is the relatedness ratio:

$$R = \frac{\text{Mean distance between control words}}{\text{Mean distance between related words}}$$

The larger this ratio, the relatively closer are the related words, and the better our semantic representations. We chose a representative set of 100 pairs of words that had been judged by human subjects to be near synonyms [14] and for each pair we took eight frequency matched random pairs of words to act as our controls. We then created the co-occurrence vectors and calculated the ratio R using simple Euclidean distances for these pairs for a range of parameter sets.

To check the extent to which our results depended on the details of the chosen evaluation criterion, our second criterion was based on a somewhat different idea. Given a set of words which human subjects have assigned to different categories, we can define category centres in the semantic vector space and ask how many of the vectors do actually fall closer to the correct category centre than to any of the other centres. If our vectors really are a good representation of semantics, we would expect all the vectors to fall closer to an appropriate category centre than to an inappropriate

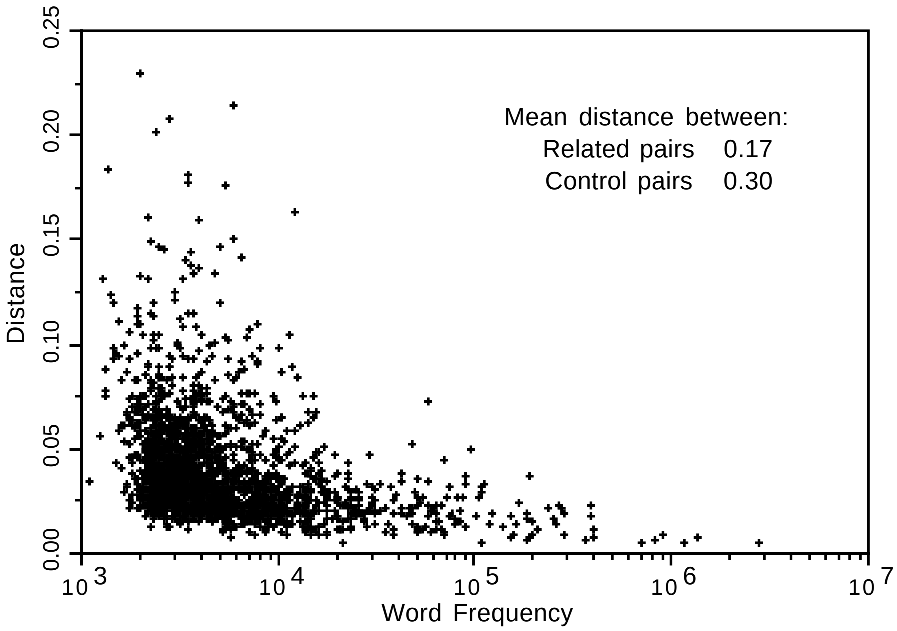

Figure 1: The distances between vectors created from different halves of the corpus showing that the higher frequency word vectors tend to be more reliable.

centre. We took ten words for each of ten Battig and Montague [1] categories which had minimal category overlap and counted the number of correct classifications for each parameter set.

6 Results

The first thing we need to consider is the reliability of the vectors we create. We are estimating probabilities by counting the word occurrences in a finite corpus, and we therefore expect the random variations in the vectors to be smallest for the high frequency words and for very large corpora. The important questions are how small can the corpus be and how low can the frequencies be before we start running into problems. We begin by checking that our full corpus of 89 million written words is large enough and that the frequencies of our chosen words are high enough, and consider what happens for smaller corpora at the end of this section. We generated vectors of the *left & right* type using a weighted window of size two with components corresponding to the 128 highest frequency words. Figure 1 shows that the distances between vectors for the same word created from different halves of the full corpus are small compared with the mean distances between different related and control words within the corpus. Figure 2 confirms this for the actual word pairs we used and shows the distribution of related and unrelated distances which is our first indication that we really are extracting semantic effects. Together these Figures give us confidence that our results are not going to be swamped by statistical noise.

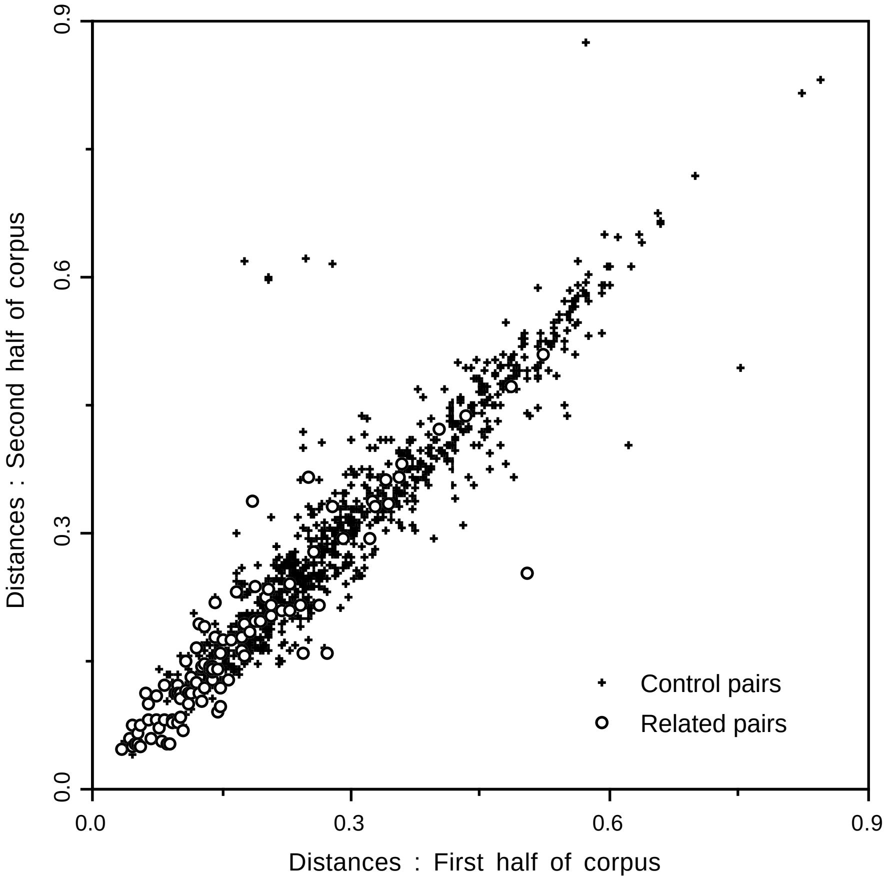

Figure 2: A comparison of the inter-word distances for vectors derived from different halves of the corpus. As one would hope, semantically related word pairs tend to be closer together than random control word pairs.

The following graphs showing how our two performance measures vary with our main parameters are fairly self explanatory. Figures 3 and 4 show how our criteria vary with the number of frequency ordered vector components. We used *left & right* type vectors for flat and weighted windows of size two. We see that the ratio measure has a peak at around a hundred components and then falls slowly, whereas the classification measure increases rapidly up to about 64 components and then remains fairly level. Figures 5 and 6 show the variation with the window size and between flat and weighted windows for *left & right* type vectors of 128 components. The ratio measure has a peak at window size two whereas the classification measure peaks nearer sixteen. In each of these graphs we have a trade-off between acquiring more information against more noise from the extra vector components or window positions. For both measures, we can see that large weighted windows behave equivalently to a flat window of about half the size.

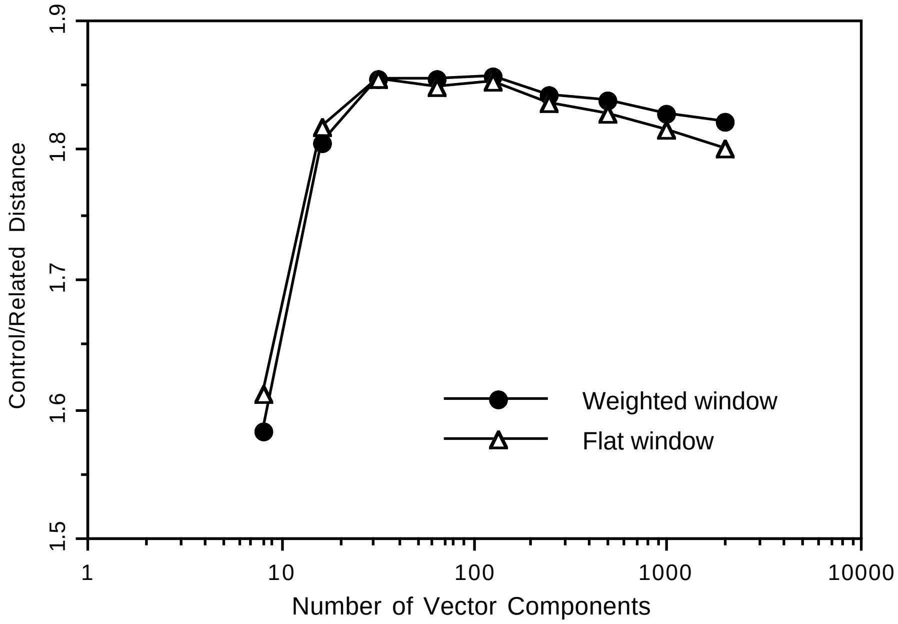

Figure 3: The plot of our Control/Related distance ratio as a function of the number of frequency ordered vector components has a maximum and then falls.

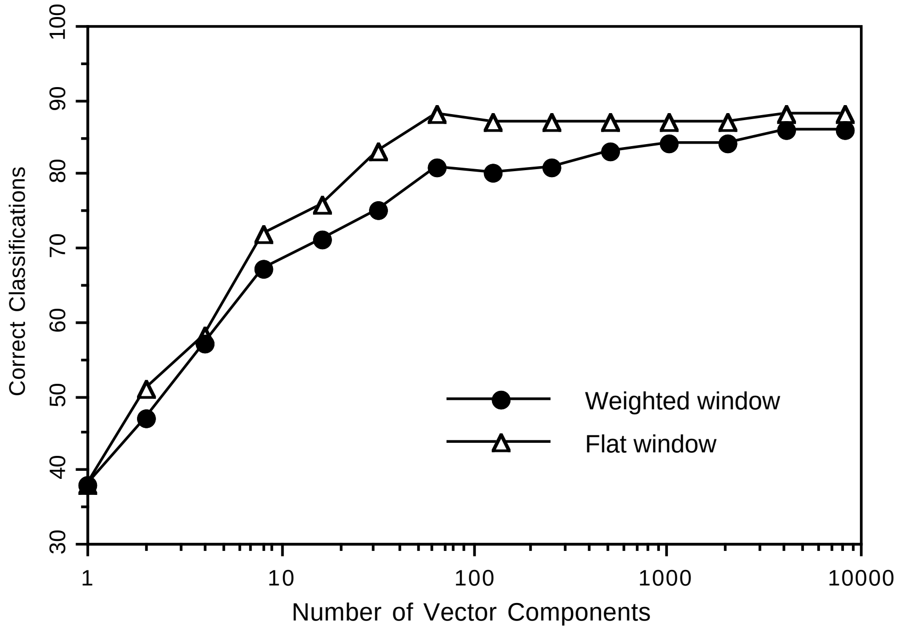

Figure 4: The plot of the number of correct classifications as a function of the number of frequency ordered vector components rises and eventually levels off.

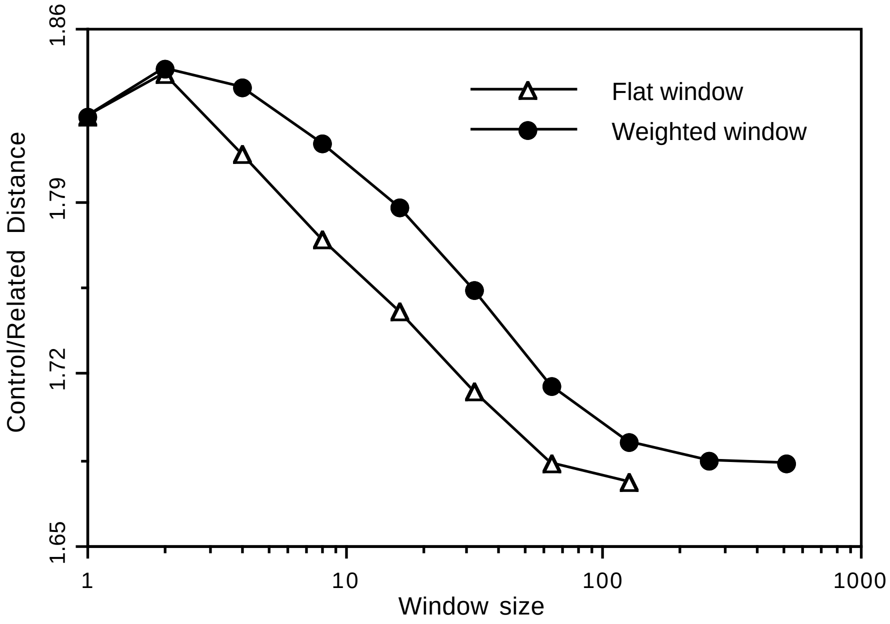

Figure 5: The plot of our Control/Related distance ratio as a function of window size has a maximum at 2 and then falls till it levels off at around 100.

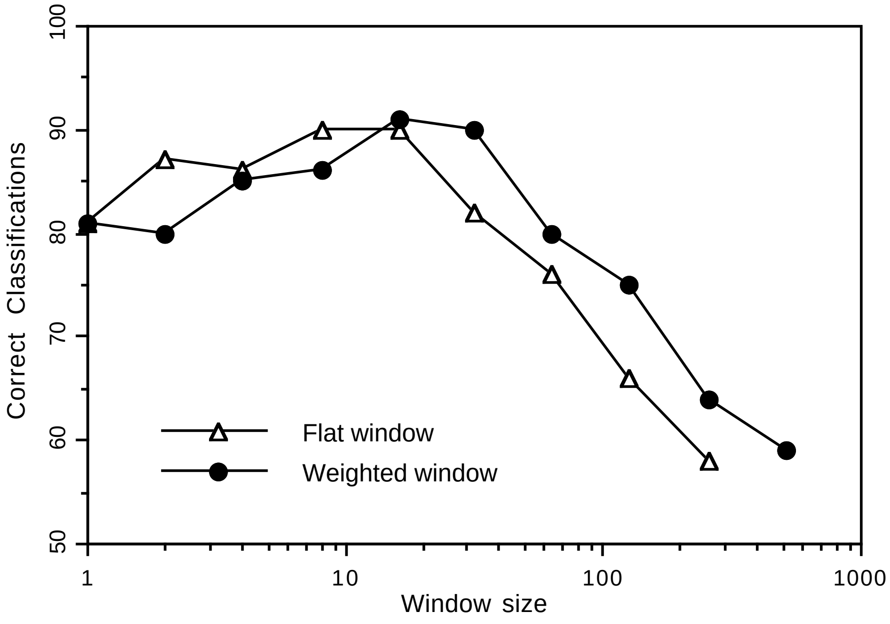

Figure 6: The plot of the number of correct classifications as a function of window size has a maximum around 16 and falls for larger windows.

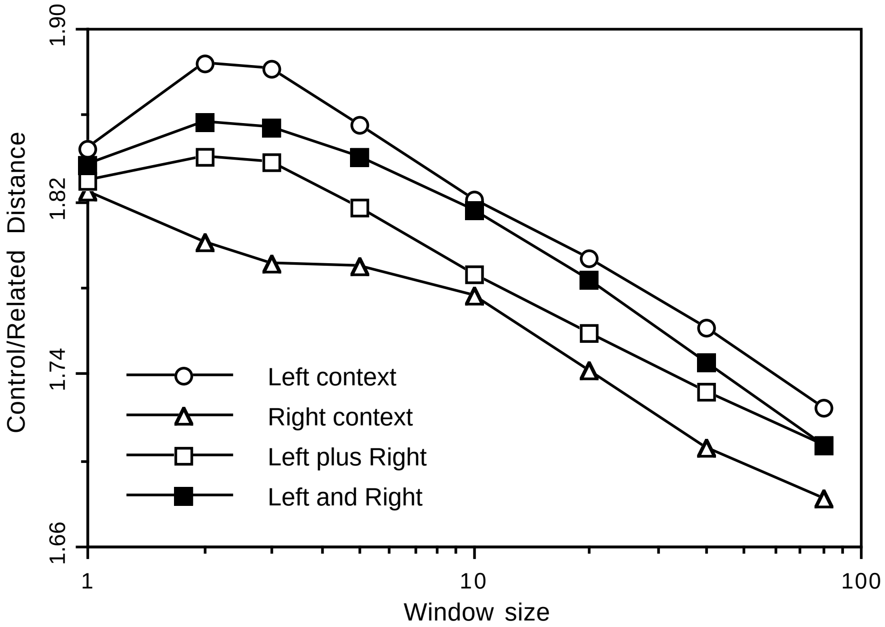

Figure 7: The plot of our Control/Related distance ratio for the four main window types as a function of window size.

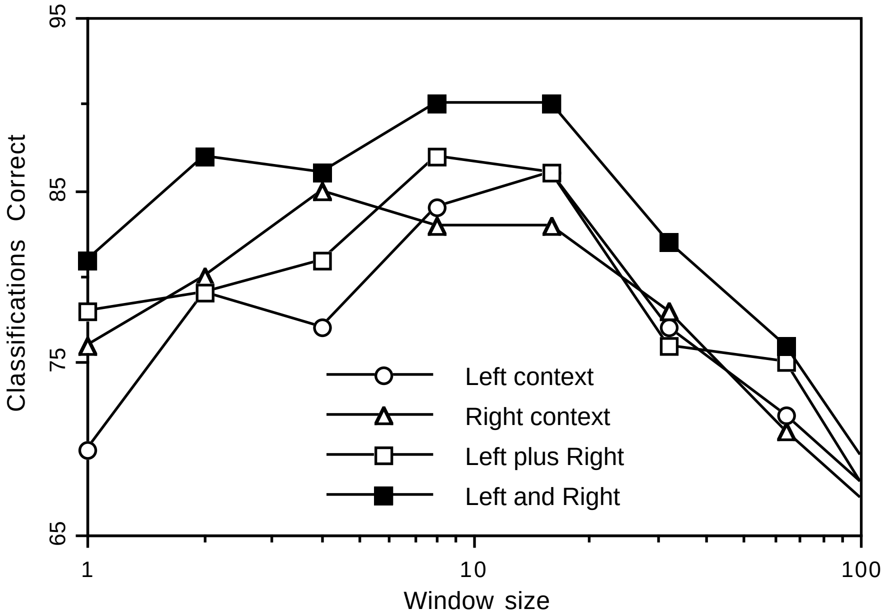

Figure 8: The plot of the number of correct classifications for the four main window types as a function of window size.

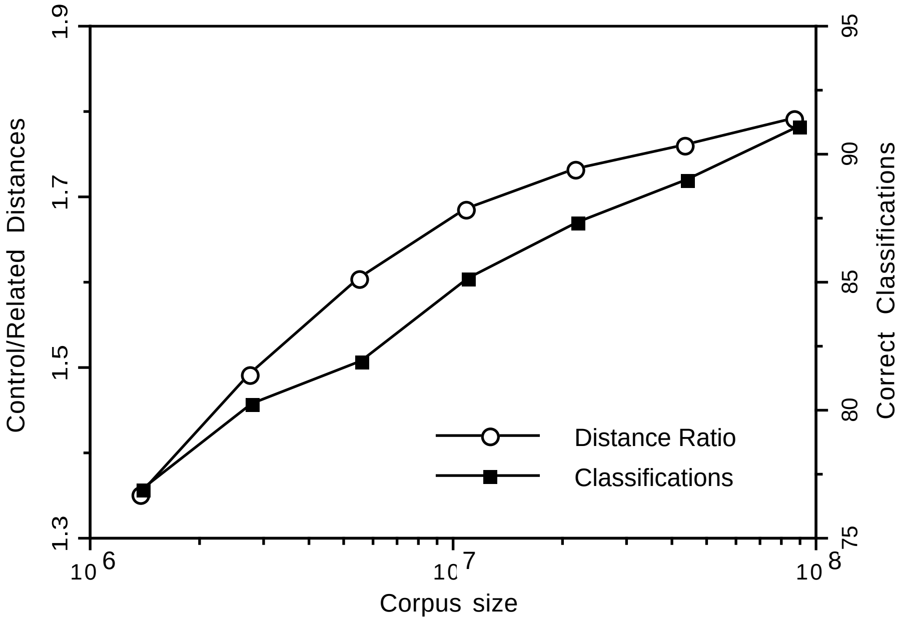

Figure 9: Both criteria for validating our semantic vectors show the expected improvement in vector quality as we increase the size of the corpus.

Figures 7 and 8 show the differences between the window types for weighted windows of different sizes for vectors of 128 components. For the ratio measure, the performance ordering is independent of the window size with the *left* contexts always giving the best vectors and the *right* contexts the worst. For the classification measure, the combined *left & right* context vectors are always the best, but the window size has a more variable effect on the others. Again, we see that the optimal choice of parameters depends on what you want to do.

Finally, Figure 9 shows how the quality of the *left & right* type vectors vary with corpus size for 128 vector components. We used near optimal weighted window sizes of 2 for the distance ratio and 16 for the classification measure. This confirms our natural expectation that, for both performance measures, the vectors do improve as we increase the corpus size, but it is slightly worrying to see that they are still improving even at the 89 million words which is the whole of the written BNC.

7 Conclusions

We have discussed how we can derive semantic representations for use in connectionist models from the word co-occurrence statistics of large text corpora. Arguments concerning their psychological realism have been suggested elsewhere [8]. Here, we have been concerned with optimising a number of parameters that affect the quality of the representations obtained from this approach, and have suggested and investigated two simple criteria for evaluating this quality.

Already we have seen that the optimal values of the various parameters will be dependent on the criteria we use. For example, our simple ratio of mean Euclidean distances between semantically related and unrelated words tells us that a co-occurrence window of two words produces the best results, whereas our measure of correct category classification suggests that a window some eight times larger is best. Preliminary analysis of other distance measures and criteria [10] suggest there is even more variability to be found. We have also seen that even with corpora of tens of millions of words, there is still room to improve our results by using still larger corpora. Clearly there is significantly more variation across the vector creation procedures and evaluation criteria than is commonly assumed.

This work, together with the preliminary application of similar corpus derived semantic representations in a connectionist lexical decision model [5], gives us confidence in the usefulness of this approach, but clearly we need more extensive analysis to determine exactly how these representations relate to those employed in real brains. For example, it does not seem appropriate to use different semantic representations for different tasks, and it is difficult to argue that our corpus derived representations are realistic if they require a corpus much larger than the total amount of written and spoken input experienced in a whole lifetime. However, variations on this theme are already looking promising [8, 10].

References

1. Battig WF & Montague WE. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. Journal of Experimental Psychology Monograph 1969; 80
2. Bullinaria JA. Modelling Reading, Spelling and Past Tense Learning with Artificial Neural Networks. Brain and Language 1997; in press
3. Bullinaria JA. Modelling Lexical Decision: Who needs a lexicon? In Keating JG. (Ed) Neural Computing Research and Applications III, 62-69. Maynooth, Ireland: St. Patrick's College, 1995
4. Bullinaria JA. Connectionist Models of Reading: Incorporating Semantics. In Proceedings of the First European Workshop on Cognitive Modelling, 224-229, Berlin: Technische Universitat Berlin, 1996

5 Bullinaria JA & Huckle CC. Modelling Lexical Decision Using Corpus Derived Semantic Representations in a Connectionist Network. In Proceedings of the Fourth Neural Computational and Psychology Workshop 1997

6. Coltheart M, Curtis B, Atkins P & Haller M. Models of Reading Aloud: Dual-Route and Parallel-Distributed-Processing Approaches, Psychological Review 1993; 100: 589-608
7. Hinton GE & Shallice T. Lesioning an Attractor Network: Investigations of Acquired Dyslexia. Psychological Review 1991; 98:74-95
8. Landauer TK & Dumais ST. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review 1997; 104: 211-240
9. Leech G. 100 million words of English: the British National Corpus. Language Research 1992, 28:1-13

10. Levy JP, Bullinaria JA & Patel M. Evaluating the Use of Word Co-Occurrence Statistics as Semantic Representations, in preparation
11. Lund K, Burgess C & Atchley RA. Semantic and Associative Priming in High-dimensional Semantic Space. In Moore JD & Lehman JF (Eds), Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society, 660-665. Lawrence Erlbaum Associates, Pittsburgh PA 1995
12. Lund K & Burgess C. Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. Behaviour Research Methods, Instruments and Computers 1996; 2:203-208
13. Miller GA & Fellbaume C. Semantic networks of English. Cognition 1991; 41:197-229
14. Moss HE, Ostrin RK, Tyler LK & Marslen-Wilson WD. Accessing Different Types of Lexical Semantic Information: Evidence From Priming. Journal of Experimental Psychology: Learning, Memory and Cognition 1995; 21:863-883
15. Patel M. Using Neural Nets to Investigate Lexical Analysis. PRICAI'96: Topics in Artificial Intelligence 1996; 241-252
16. Plaut DC. Semantic and Associative Priming in a Distributed Attractor Network. Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society 1995; 37-42
17. Plaut DC & Shallice T. Deep Dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology 1993; 10:377-500
18. Plaut DC, McClelland JL, Seidenberg MS & Patterson KE. Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. Psychological Review 1996; 103:56-115
19. Schutze H. Word Space. In Hanson SJ, Cowan JD & Giles CL (Eds), Advances in Neural Information Processing Systems 5, 895-902. Morgan Kaufmann, San Mateo CA, 1993.
20. Seidenberg MS & McClelland JL. A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review 1989; 96:523-568