Estimating Conceptual Similarities using Distributed Representations and Extended Backpropagation

Peter Dreisiger[1,2], Cara MacNish[2] and Wei Liu[2]
[1] Maritime Operations Division, Defence Science and Technology Organisation
[2] Computer Science & Software Engineering, The University of Western Australia
email: {prd,cara,wei}@csse.uwa.edu.au

ABSTRACT

The ability to perceive similarities and group entities into meaningful hierarchies is central to the processes of learning and generalisation. In artificial intelligence and data mining, the similarity of symbolic data has been estimated by techniques ranging from feature-matching and correlation analysis to *Latent Semantic Analysis (LSA)*. One set of techniques that has received very little attention are those based upon cognitive models of similarity and concept formation.

In this paper, we propose an extension to a neural network-based approach called *Forming Global Representations with Extended backPropagation (FGREP)*, and show that it can be used to form meaningful conceptual clusters from information about an entity's perceivable attributes or its usage and interactions. By examining these clusters, and their classification errors, we also show that the groupings identified by FGREP are more intuitive, and generalise better, than those formed using LSA.

Categories and Subject Descriptors:

I.2.4 [**Artificial Intelligence**]: Computing Methodologies — Knowledge Representation Formalisms and Methods

Keywords:

Subsymbolic processing, Knowledge representation, Dimensional reduction, Distributed representations, Neural networks

1. INTRODUCTION

The ability to perceive similarities, and form meaningful, or 'natural', groups underlies some of our most important mental processes. In remembering, it allows us to go beyond superficial correspondences and identify precedents based upon structural, or deeper, similarities. In problem solving, it allows us to draw upon our past observations and experiences, and to find solutions in a timely and context-appropriate manner. And in learning, these groupings determine the nature of the associations and the quality of our generalisations.

In the fields of artificial intelligence and data mining, the process of placing similar objects or observations into groups is called *cluster analysis*, and its goal is to find an arrangement that maximises both the *intra*-cluster similarities and the *inter*-cluster differences. While the most common estimates of these differences are also measures of distance, choosing the 'best' measure can be far from trivial — on the one hand, it depends upon the form and representation of our observations; on the other, it determines the make up of the clusters and the types of relationships they capture.

For symbolic data, such measures should capture salient differences in the terms' roles and features. Techniques such as feature matching and correlation analysis are commonly used to estimate the difference between feature vectors or sets of objects [2]. However, these approaches have several limitations: firstly, the vectors' high dimensionality, and their resulting sparseness, can be a problem for traditional measures of distance; secondly, they treat each variable, or dimension, as equally important — an assumption that is often incorrect; and thirdly, their focus on the presence or absence of terms makes them blind to the order of terms and, thus, their roles.

One way to provide a more realistic estimate of the differences is to manually weight each variable according to its importance, or salience. Unfortunately, this requires the weights to be known beforehand, and it assumes that they are constant across situations and contexts. Another solution is to find a transformation that better captures, or accounts for, the variations in the raw data. The most common way of finding these transforms involves the use of *dimensional reduction*, and within data mining, one of the most powerful and widely used examples of this is *Latent Semantic Analysis (LSA)*.

LSA was developed to analyse and characterise written documents [4], and it uses dimensional reduction to minimise the effects of its inputs' sparseness and biases. Not only does this produce representations that better reflect the words' average meanings, but the distances between them can be used by traditional clustering algorithms to identify groups of related terms. What it cannot do, however, is record word order or differentiate between roles.

The need to represent symbolic terms, and to capture their 'essence' in a relatively small representation space, is not confined to the fields of artificial intelligence and data mining. Within the cognitive sciences, these two tasks are largely subsumed by the study of one of *our* greatest natural abilities — concept formation; indeed, the ease with which we can recognise salient similarities, and our ability to draw in-

17th School of Computer Science & Software Engineering Research Conference, Yanchep, Western Australia, 9–10 September 2009.

dividual instances together into a common hierarchy of concepts, are subjects of ongoing research. Theories, such as Rosch et al.'s *Basic Level of Categorisation* have attempted to account for the order in which we acquire concepts [8] , while others have even tried to explain the nature of concepts, and the process of concept *formation*, in terms of conceptual spaces and geometric representations (see, for example, Gärdenfors' [1] model).

In this paper, however, we will focus on a particularly promising model of sub-symbolic processing that was first described by Miikkulainen and Dyer [6]. Like LSA, their technique, called *Forming Global Representations with Extended back-Propagation (FGREP)*, represents symbolic data as vectors, uses a form of dimensional reduction to emphasise deeper correlations, and produces representations whose relative distances reflect underlying similarities. Unlike LSA, it develops these representations along a relatively small set of opaque dimensions, and it takes order into account.

Through a series of experiments, we investigate FGREP's ability to produce meaningful clusters given an entity's perceivable attributes, or its usage and interactions. We compare these groupings to the reference class hierarchies and those found using LSA, we examine how well the resulting clusters generalise and capture differences in the terms' roles, and we show that it can be used to provide a more intuitive estimate of similarity. We begin, in Section 2, with an overview of LSA and FGREP. In Section 3, we describe our variation of the FGREP model and the test data, and in Section 4, we present the results of two experiments that focus on perceptual data and term usage. We close, in Section 5, with a summary of our results and an outline of our future work.

2. BACKGROUND

LSA is a well established statistical technique that accepts a matrix of term–passage frequencies, and uses dimensional reduction to compute their 'average meanings'. More specifically, it uses reduced-rank singular value decomposition to produce estimates of the average frequencies that are based upon the terms' shared neighbourhood, and their distribution across passages. Not only do these estimates reflect the words' meanings more closely than the raw frequencies, but the distances between the resulting row and column vectors can be used by traditional clustering algorithms to identify groups of related terms.

LSA is fundamentally deterministic, it can scale to relatively large corpora, and studies have shown that its estimates of inter-document similarity are comparable to those produced by human subjects [4]. LSA does, however, have some limitations: it assumes that the inputs can, in fact, be divided into natural bundles of information such as paragraphs or documents, and it is blind to the ordering of words within a document [3]. While LSA is well suited to the analysis of *textual* documents, this latter fact, in particular, affects both its ability to form clusters from declarative knowledge, and its ability to arrange them into meaningful hierarchies.

FGREP is a cognitive model that was introduced by Miikkulainen and Dyer [6], and explored further by Miikkulainen [5]. It is built around a multi-layer perceptron, and the essence of this model is that: (1) every term is represented by a numerical vector, called a *distributed representation*; (2) these representations change, from initially random values, to ones that capture patterns in the terms' usage; and (3) they are developed *automatically* by propagating the error signals back through the hidden layer to the network's inputs. Put another way, this extended form of backpropagation changes both the network's weights *and* its inputs, and FGREP uses these additional degrees of freedom to further improve the network's accuracy.

Structurally, FGREP consists of three components: a distributed representation store, a 'routing' network, and a three-layer perceptron. The store keeps track of the terms' current distributed representations; the purpose of the routing network is to convert tuples of terms into input and target vectors, and vice versa. At the beginning of each training cycle, the routing network retrieves the appropriate representations from the store, concatenates their current patterns of activation, and presents the resulting vector to the neural network's input and target layers (see Figure 1a). The errors are then calculated, and once the inputs have been updated, the routing network reverses the process by extracting the terms' new representations and placing them back into the store (Figure 1b). It is this circulation of representations that allows them to develop.

While FGREP is, to the best of our knowledge, the first neural system that changes its own inputs to better reflect regularities in the training data, it is not the only one to use extended backpropagation. In particular, this rule, which was called *backpropagation-to-representation* by Rogers and McClelland [7], formed the basis of their model of human concept formation. What distinguishes FGREP from this approach is its use of a distributed store and a routing network. Even though backpropagation-to-representation develops a set of distributed representations, it does so across a hidden layer — the input and target patterns are still expressed as feature vectors and, for this reason, the number of terms it can support is completely determined by the size of its input and output layers.

In contrast, FGREP's capacity is largely independent of its network's geometry. The structure of FGREP's networks also means that it is able to form its representations from a series of declarative sentences (rather than a vector of frequencies or hand-picked features), and that it can do so *incrementally* and using an arbitrary sentence structure.

3. METHODOLOGY

In this paper, we investigate the types of clusters that FGREP can produce given either perceptual data or information about an entity's usage and interactions, we compare them to the reference hierarchies and those found using LSA, and we examine how well the resulting clusters generalise. This is in contrast to the earlier studies which used FGREP or extended backpropagation as part of a larger natural language processing system, or to study the process of human concept formation.

The system used in our experiments extends that of Miikkulainen and Dyer [6]. To focus on developing the distributed representations, and to improve their stability, we

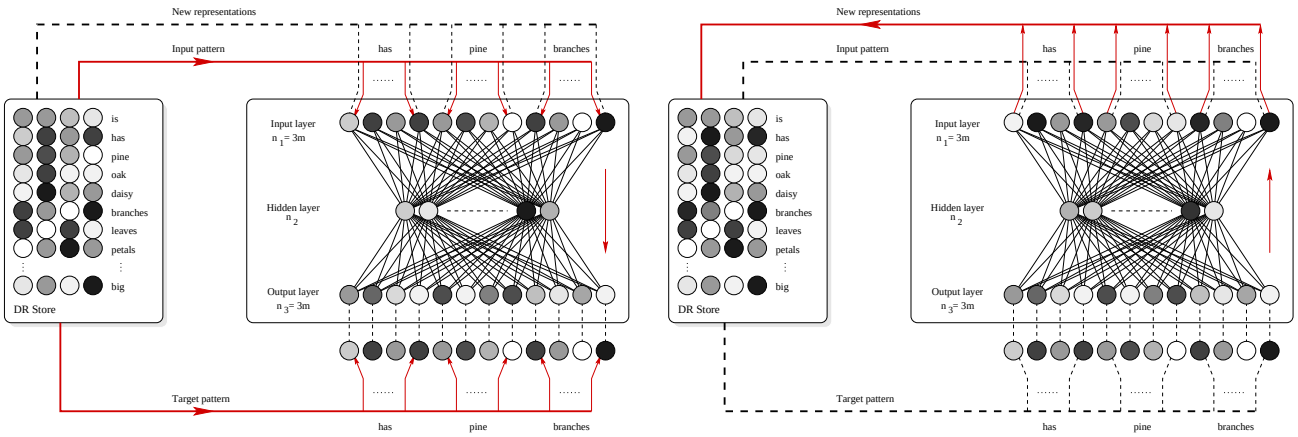


Figure 1: A generalised FGREP system, based upon Miikkulainen and Dyer [6]. The system consists of a symmetric, three-layer perceptron, a routing network, and a distributed representation store. In this example, the network is being taught to remember a three-term (or symbol) input sequence; at the same time, the representations used to construct the sequences are being updated by propagating the errors back to the input layer.

made several changes to the network's structure, input–target pairs and learning rule. Firstly, we used a *symmetric* three-layer perceptron, or auto-encoder, in place of the original model's asymmetric feed-forward network. Secondly, we taught our network to reproduce its input sentences where Miikkulainen's system was trained to identify the agents, patients, actions and modifiers given a sentence's syntactic representation. (See Figure 1 for a diagram of our generalised FGREP system.)

To improve the stability of the system, we implemented a batch learning algorithm alongside the existing sequential, or iterative, rule. (In sequential mode, the order of the training data varies from epoch to epoch, and the weights are changed after the presentation of each example; in batch mode, the weight changes are accumulated over the entire training set, and the differences are only applied at the end of each epoch.) Normally, the latter tends to yield better results and a faster rate of convergence; in our case, however, it also means that otherwise identical terms — i.e. those with the same initial position and usage — will develop divergent representations[1].

Finally, our system supports both dynamic and static representations. While the vast majority of representations continue to develop over time, this addition allows some terms, and the relations in particular, to be given a fixed representation. This use of static relations follows from Rogers and McClelland [7], and initial tests have shown that the judicious use of fixed representations can help to 'ground' the system and improve both the quality of the other representations and their rates of convergence.

For evaluation, we used two distinct data sets. The first is based on the training corpus of Rogers and McClelland [7, Appendix B2], but incorporates two changes: (1) the data was converted from a set of binary feature vectors to a sequence of `relation object attribute` sentences; and (2) information about the entities' class hierarchies was removed to produce a purely descriptive data set. The second set consists of the sentence templates and noun categories of Miikkulainen [5, Tables 5.1 and 5.2]; to generate the training set proper, we went through the list of templates and enumerated every possible instance according to the list of nouns. These sets are shown in Tables 1a and 1b, respectively.

The data sets were chosen for two main reasons: (1) they allow us to verify the behaviour of our system using existing corpora, and (2) collectively, they allow us to see how FGREP performs under a variety of conditions. The first set is relatively small, contains only perceptual information, and represents knowledge using `relation object attribute` triples. In contrast, the second set is larger (containing nearly 700 sentences versus 60), consists of relational information, and uses five-part tuples to describe its entities and the roles they play.

The experiments consisted of 25 trials that ran for 100,000 epochs. We used default values for the distributed representation and hidden-layer sizes (12 and 18 neurons, respectively[2]), and sampled the representations periodically to track their development. Each trial within an experiment used the same declarations but seeded the random number generator with a different value; this meant that each trial had a unique set of distributed representations — both initial and learned. On a 2.8GHz Pentium IV system, each trial took approximately six minutes to execute for the first set of data, and 190 minutes for the second.

[1] Our choice of learning rule, and the way in which it affects the representations' sensitivity to perturbations in their initial patterns of activation, will be the subject of a later investigation, however initial tests suggest that batch learning leads, on average, to a better set of final representations.

[2] The effects of the distributed representation and hidden layer sizes are the focus of ongoing experiments; for these trials, however, we used the default values, as recommended by Miikkulainen [5, p. 54].

	pine	oak	rose	daisy	robin	canary	sunfish	salmon
is-pretty	0	0	1	1	0	0	0	0
is-big	1	1	0	0	0	0	0	0
is-living	1	1	1	1	1	1	1	1
is-green	1	0	0	0	0	0	0	0
is-red	0	0	1	0	1	0	0	1
is-yellow	0	0	0	1	0	1	1	0
can-grow	1	1	1	1	1	1	1	1
can-move	0	0	0	0	1	1	1	1
can-swim	0	0	0	0	0	0	1	1
can-fly	0	0	0	0	1	1	0	0
can-sing	0	0	0	0	0	1	0	0
has-skin	0	0	0	0	1	1	1	1
has-roots	1	1	1	1	0	0	0	0
has-leaves	0	1	1	1	0	0	0	0
has-bark	1	1	0	0	0	0	0	0
has-branch	1	1	0	0	0	0	0	0
has-petals	0	0	1	1	0	0	0	0
has-wings	0	0	0	0	1	1	0	0
has-feathers	0	0	0	0	1	1	0	0
has-gills	0	0	0	0	0	0	1	1
has-scales	0	0	0	0	0	0	1	1

(a)

`human` ate	
`human` ate `food`	
`human` ate `food` with `food`	
`human` ate `food` with `utensil`	
`animal` ate	
`predator` ate `prey`	
`human` broke `fragileobj`	
`human` broke `fragileobj` with `breaker`	
`breaker` broke `fragileobj`	
`animal` broke `fragileobj`	
`fragileobj` broke	
`human` hit `thing`	
`human` hit `human` with `possession`	
`human` hit `thing` with `hitter`	
`hitter` hit `thing`	
`human` moved	
`human` moved `object`	
`animal` moved	
`object` moved	

`human`	boy girl man woman
`animal`	bat chicken dog lion sheep wolf
`predator`	lion wolf
`prey`	chicken sheep
`food`	carrot cheese chicken pasta
`utensil`	fork spoon
`fragileobj`	plate vase window
`hitter`	ball bat hammer hatchet paperwt rock vase
`breaker`	ball bat hammer hatchet paperwt rock
`possession`	ball bat dog doll hammer hatchet vase
`object`	ball bat carrot cheese chicken curtain desk dog doll fork hammer hatchet paperwt pasta plate spoon vase window
`thing`	`animal human object`
`action`	ate break hit move

(b)

Table 1: (a) The entity definitions, after Rogers and McClelland [7, Appendix B2]; and (b) the sentence templates and noun categories of Miikkulainen [5, Tables 5.1 and 5.2].

4. EXPERIMENTS AND RESULTS

The representations resulting from training were analysed using the open-source statistical package, R [9]. Firstly, dendrograms depicting the relationships between each of the terms were generated using hierarchical cluster analysis and the Euclidean measure of distance. The distances between each of the terms' distributed representations were averaged over all 25 trials to produce the 'typical' cluster plots shown in this paper; term–passage frequency vectors were also calculated using LSA in R, with stemming disabled, and an intermediate dimensionality that was half the number of terms in the corpus.

Secondly, the average per-term reconstruction errors squared were calculated and plotted against time to visualise the relative rates of convergence. That is, the squared difference between each distributed representation and the corresponding signal at the network's output layer was calculated, and averaged over the 25 trials; these values were further averaged to determine the per-category and overall errors.

Finally, the classification accuracies of FGREP and LSA were compared. The statistic we chose for this analysis was *category intrusion*, defined as the number of *unrelated* terms whose representations fall within the hypersphere defined by the category's centroid and its furthest member (i.e. its radius). Not only does this metric provide us with an indication of the techniques' relative accuracy, but by calculating the intrusions for the super-classes as well as the categories, we can see the extent to which they overlap, and if the clusters, themselves, form meaningful class hierarchies.

4.1 Experiment 1: Identifying Concepts by their Properties

The first experiment focuses on the nature and quality of the clusters that FGREP can form given only information about an entity's *perceivable* attributes, and its results are summarised in Figures 2 and 3. From Figure 2a, we can see that LSA tends to emphasise co-occurrences over semantic similarities; for example, it contains groups such as `(robin (sing (fly feathers wings)))`, `(swim gills scales)`, `(petals pretty)` and `(move skin)`. It is, perhaps, for this reason that there are some inconsistencies within the entities' hierarchy: for example, there is a significant distance between the `canary` and the `robin`, and the `trees` are closer to the `fish` than the `flowers`.

In contrast, the clusters formed using FGREP capture both the terms' associations *and* type (see Figures 2b–d). Continuing with the above examples, after just 10,000 epochs FGREP was able to separate the entities, parts, attributes and actions into their own clusters, each of which could be further divided into plant and animal varieties:

```
(((salmon sunfish) (canary robin)) ((oak pine) (daisy rose)))
(((feathers wings) (skin (gills scales))) ((bark branch)
                                          (petals (leaves roots))))
((big green) (yellow (red (living pretty))))
((move (fly sing)) (grow swim))
```

Furthermore, in all of the 25 trials, this process of differentiation continued over time — i.e. the longer the training continued, the more distinct these clusters became.

Figure 3a shows the per-term and per-category errors-squared, averaged over the 25 trials. From this plot we can see not only how the reconstruction errors change over time, but how well the distributed representations capture the terms' nature. On the first point, we can see that, as a rule, the errors decrease for the first 40,000–50,000 epochs; after this point, however, they seem to plateau, or even oscillate. While determining the reason for this behaviour is beyond the scope of this paper, it might be a sign that the network is trying to over-fit the representations, in turn, to each of the sentences in which they appear.

On the second point, we can see that the representations corresponding to the *entities* — that is, the plants and an-

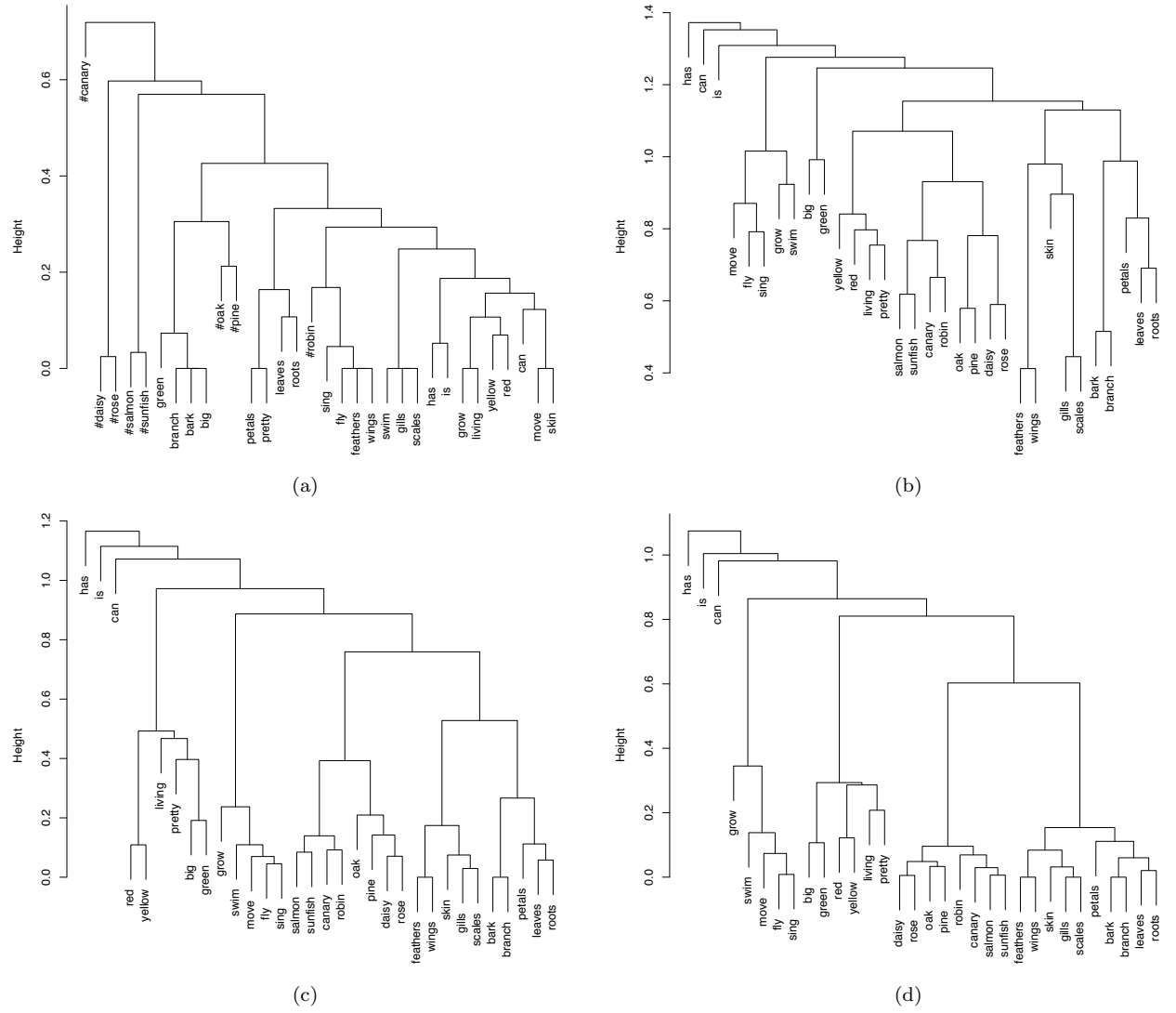
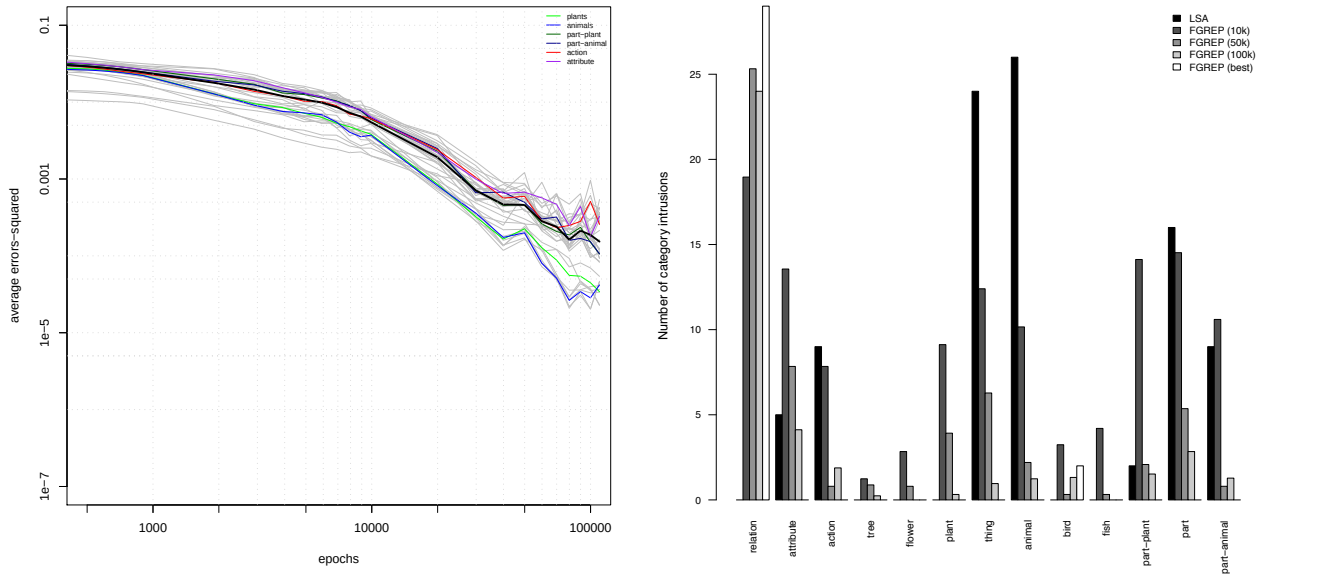


Figure 2: Hierarchical cluster plots based upon: (a) the distances as estimated by LSA; and the average distances, as estimated by FGREP, after (b) 10,000, (c) 50,000 and (d) 100,000 epochs. The average distances were calculated over all of the 25 trials, while the join heights indicate the Euclidean distance between term representations and/or clusters.

imals — had the lowest reconstruction errors while the actions and attributes (which spanned these categories) had the highest errors. In particular, there appears to be an inverse relationship between a term's reconstruction error and the information gain it offers. Take, for example, the animals and the actions (represented by the blue and red lines, respectively): given an animal, we can completely predict which other terms are about to appear; most actions, however, are associated with several plants and/or animals and, as such, they have less predictive power.

Looking at the category intrusions of Figure 3b, and the errors associated with LSA, we can make several observations: (1) while LSA is able to pair the trees, flowers, birds and fish off into their own distinct clusters, it is unable to do the same for the parts, attributes and actions; (2) even though it is able to correctly group the entities, it has difficulty *generalising* these lower-level clusters into higher-level concepts such as animals, things and parts; and (3) it groups all of the relations together, even though they are used in completely different senses.

The subsequent four sets of errors in Figure 3b were calculated from the results of the FGREP trials. The first three were based on the average inter-term distances after 10,000, 50,000 and 100,000 epochs; the last shows the results of the 'best' trial (i.e. the one with the fewest classification errors). In contrast to the LSA results, we can see that: (1) while FGREP is, on average, able to correctly group the trees, flowers, birds and fish, it does occasionally misclassify some of the entities; (2) unlike LSA, however, it *is* able to separate the parts, attributes and actions into their own clusters; (3) while the number of intrusions increases as we move up the class hierarchy, it is actually able to generalise the lower-level clusters into higher-level concepts; and (4) the relations remain relatively distinct.



(a) (b)

Figure 3: (a) The average per-term and per-category reconstruction errors-squared, calculated over the 25 trials; and (b) The category intrusions when the terms were grouped according to their LSA and FGREP measures of similarity. The first three FGREP results are the average number of category intrusions after 10,000, 50,000 and 100,000 epochs, calculated over the 25 trials; the final column shows the number of intrusions, after 100,000 epochs, for the best individual trial.

Even with a relatively small data set, this experiment demonstrated that FGREP, and thus extended backpropagation, are able to form novel conceptual clusters from perceptual data alone. Like the network of Rogers and McClelland [7], our implementation was able to correctly identify the relationships between the birds, fish, trees and flowers. Unlike their system, our adaptation of Miikkulainen's feed-forward network was also able to develop representations for the other terms — representations that captured not only the terms' basic *types*, but their finer structure as well.

When compared to latent semantic analysis, FGREP seems to form more intuitive clusters, and is better able to arrange these groupings into meaningful hierarchies; in other words, while the average number of category intrusions increases as we move up the classification tree, it does so slowly enough for the resulting generalisations to still be useful. Of course, FGREP is a stochastic process and, thus, the quality of the representations it produces are affected by their initial values; this is in contrast to LSA, which is fundamentally deterministic.

4.2 Experiment 2: Identifying Concepts by their Usage

In the first experiment, we focused on FGREP's ability to form clusters from perceivable attributes; the aim of the second experiment is to see how it performs given only information about an entity's usage and interactions, and its results are shown in Figures 4 and 5. From Figure 4a, we can see that LSA tends to favour co-occurrences and shared 'neighbourhoods' over similarities in the terms' usage. Consider, for example, the groups `(ate (pasta carrot cheese))`, `(lion sheep wolf)` and `(paperwt vase)`. The first group contains both the unambiguous `foods` and the action `ate`. The second has placed the `predators` in with a `prey`, even though they play quite different roles in the training data. The final group consists of `paperwt` and `vase`; while both of these terms belong to the `thing`, `object` and `hitter` categories, there are some *semantically* significant differences too — `vase` is also a member of the `fragileobj` and `possession` groups while `paperwt` belongs to the `breaker` category.

In contrast, from Figure 4b–d, we can see that FGREP appears to do a better job of capturing the terms' types and usage — unlike LSA, the `foods` have been separated from the term `ate`; the `predators` are now distinct from their `prey`; and the `paperwt` is now more closely associated with the `hitters` and `breakers` than it is with the `vase`. A related observation is that, while the ambiguous terms' relationships are somewhat unclear, FGREP was still able to separate them from amongst the other `animals`, `hitters`, `fragileobjs` and `possessions`. Unlike the previous experiment, the terms in this data set cannot be arranged into any meaningful hierarchies; rather the intrusions were calculated for each of the ten basic noun categories, the actions, and the catch-all category, 'thing'.

As with the first experiment, we can make two observations from the average reconstruction errors shown in Figure 5a: (1) the reductions were only consistent for the first 40,000–50,000 epochs — after that they became much more erratic; and (2) a *category's* average error seems to reflect the diver-

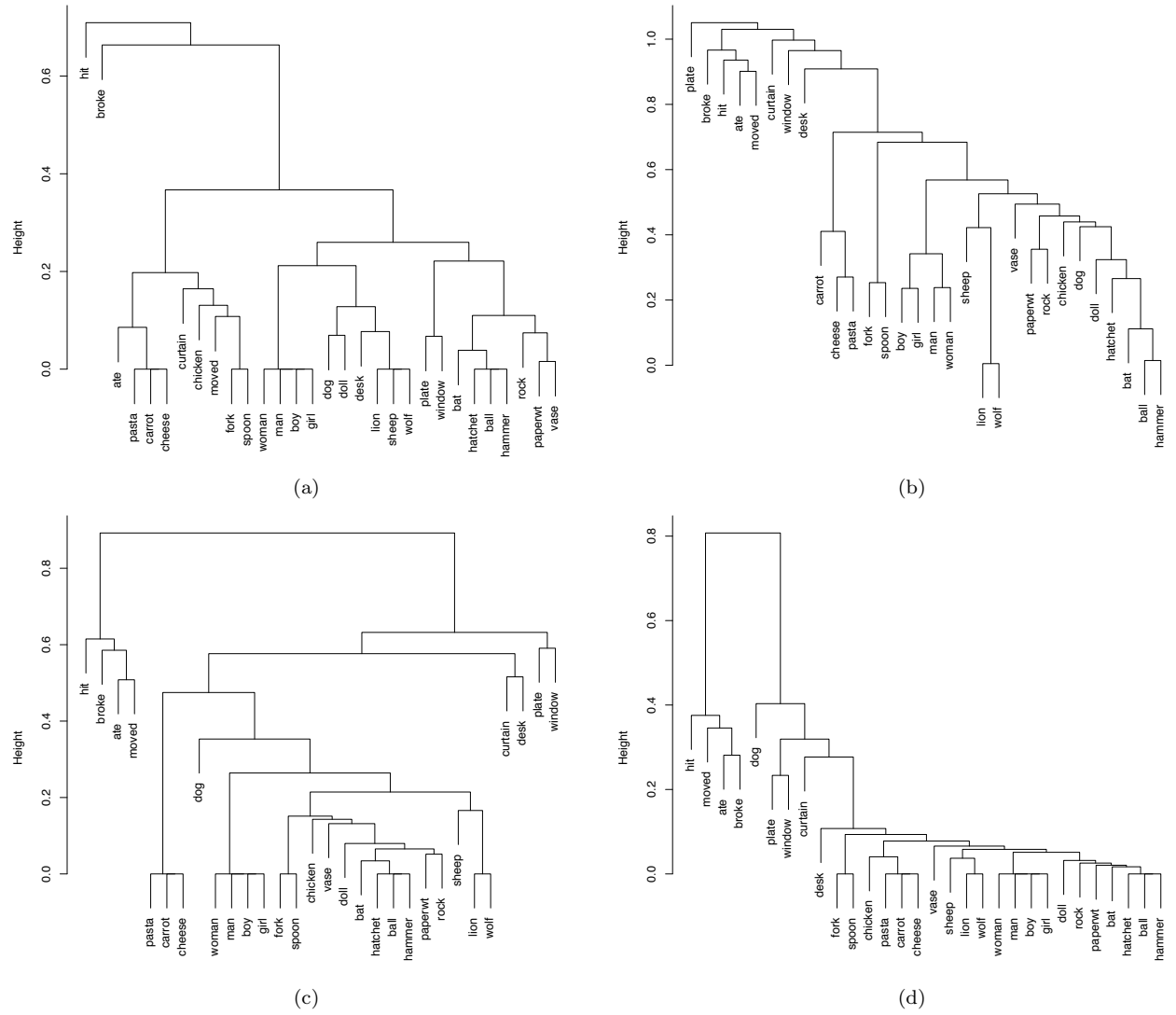


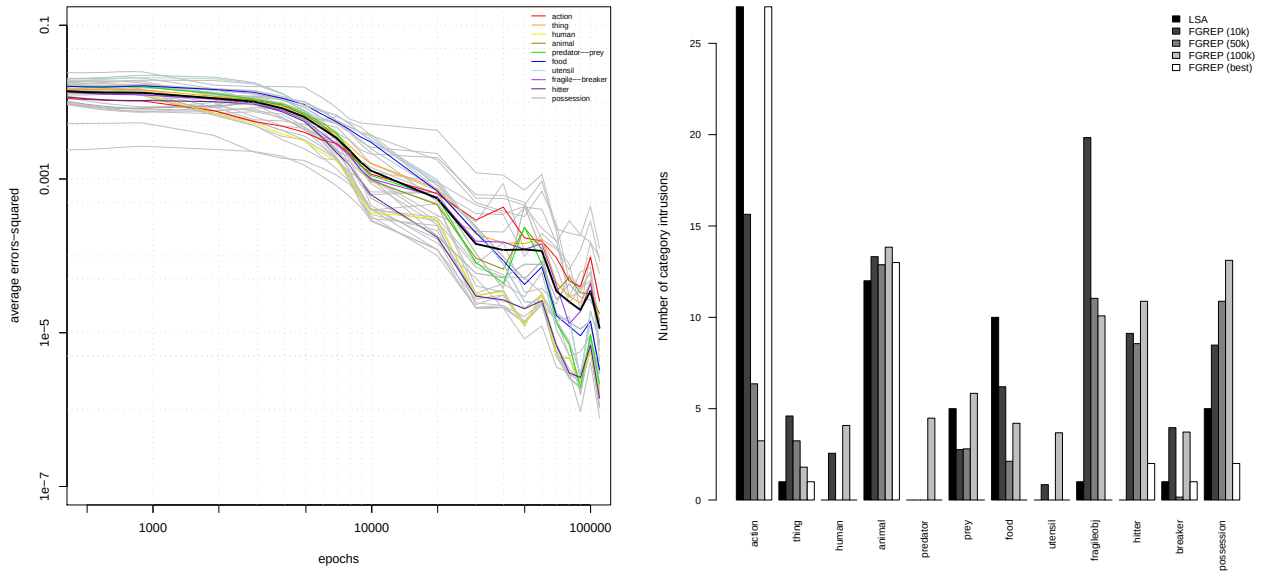
Figure 4: Hierarchical cluster plots based upon: (a) the distances as estimated by LSA; and the average distances, as estimated by FGREP, after (b) 10,000, (c) 50,000 and (d) 100,000 epochs, calculated over the 25 trials.

sity of its members' behaviour. Compare, for example, the categories `human` and `thing`: the former's members, which have amongst the lowest reconstruction errors, all exhibit the same behaviour, while the latter category consists of terms that assume ten distinct roles. The effects of these fluctuations can also be seen in Figure 5b; while the number of intrusions decreases as we go from 10,000 to 50,000 epochs, the numbers after 100,000 epochs are actually higher for nine of the twelve categories.

Looking at the category intrusions of Figure 5b and the errors associated with LSA, we can see that: (1) LSA was able to form distinct and disjoint clusters for each of the homogeneous categories — that is, the `human`, `utensil` and `hitter` categories whose members' behaviours are essentially identical; and (2) the categories with five or more intrusions each contained terms that played several different roles — for example, the concept `animal` includes terms that also belong to the `hitter`, `breaker`, `possession`, `predator`, `prey` and `food` categories.

FGREP's accuracy was less impressive in this experiment than it was for the first set of data. After 50,000 epochs, FGREP produced, on average, fewer intrusions than LSA for only three of the 12 categories, it was comparable for three, and actually yielded *more* intrusions for the six other categories; after 100,000 epochs, FGREP's average accuracy was worse than LSA's for 11 of the 12 categories. However, its *best* case performance after 100,000 epochs was more encouraging, yielding fewer intrusions for four categories, and producing comparable results for five other categories.

While the results of this experiment were less conclusive than those of the first, they demonstrate that FGREP is able to identify conceptual clusters from information about the entities' usage alone; furthermore, they show that, in the best case, FGREP is able to capture most of an entity's behaviour, even when it assumes several distinct roles.



(a) (b)

Figure 5: (a) The average per-term and per-category reconstruction errors-squared, calculated over the 25 trials; and (b) The category intrusions when the terms were grouped according to their LSA and FGREP measures of similarity. The first three FGREP results are the average number of category intrusions after 10,000, 50,000 and 100,000 epochs, calculated over the 25 trials; the final column shows the number of intrusions, after 100,000 epochs, for the best individual trial.

This experiment, and the differences between the average and best-case results, also highlighted two of FGREP's less desirable qualities: (1) that the quality of its clusters can be quite sensitive to the distributed representations' initial positions; and (2) that this dependence is exacerbated for terms that exhibit multiple types of behaviour.

Where FGREP *did* perform better than LSA, it seemed to do so by capturing terms' roles as well as their co-occurrences. For example, LSA was unable to differentiate between the `predators` and the `sheep` because it cannot take term order or syntax into account; FGREP, on the other hand, was able to distinguish between these two categories because they assume different roles, and thus occupy different slots, in Miikkulainen's data (the `predators` are agents while the `prey` are patients). Similar results were also observed for the `food` and `breaker` categories.

5. CONCLUSIONS AND FURTHER WORK

In this paper, we saw that FGREP is able to place terms into meaningful and intuitive clusters given either perceptual data or information about an entity's usage and interactions. While the quality of these clusters varied across trials and experiments, it is consistently able to: (1) arrange them into distinct and intuitive class hierarchies; (2) capture relationships and hierarchies that LSA cannot; and (3) distinguish between terms based upon the roles they assume.

These abilities, and at least some of FGREP's novelty as a technique, come from the fact that its fundamental unit of knowledge is the sentence, and that the order of terms *within* these sentences is both meaningful and well defined. Interestingly, the quality of its clusters seems to depend upon the number of distinct roles its members assume — i.e. the more complicated or ambiguous a term's usage, the harder it is to represent as a single point.

In the future, we will focus on characterising FGREP's typical behaviour, studying the stability of its learning rule, providing better support for terms that assume multiple roles, and extending FGREP to support discrete episodes and causal relationships; we will also use our analyses, both statistical and dynamical, to try to improve the average quality of its representations. Specific questions to be addressed include:

- How does the *initial* distribution of representations affect their later development, can we draw any conclusions about the 'typical' behaviour of the system, and can we develop any heuristics that improve the *average* quality of the representations?
- How stable is FGREP's learning rule? Treating the network as a dynamical system and the representations as particles, we can ask how stable are the representations to perturbations in their initial positions, how do the representations interact (and do the developed ones form concept-specific 'attractors'), and how does the trajectory of an ambiguous term's representation differ from those that participate in only one kind of relationship?
- Can we improve the system's performance, particularly for the ambiguous terms, by using *more* than one

particle per term? I.e. instead of forcing an ambiguous term to lie somewhere *between* each of its natural clusters, can we use multiple particles to develop unambiguous representations for each of its possible roles or use-cases?

- How is FGREP related to the auto-encoder, on the one hand, and a single-objective, multiple-particle form of gradient descent, on the other? Are there any optimisations that we can apply from these techniques to FGREP?
- What *other* kinds of information can it learn from? Thus far we have focused on declarative knowledge and case–role descriptions — can it be applied to other types of information, such as n-gram representations or the description of paths within a graph or semantic network?
- Can FGREP be extended to capture discrete episodes, and even causal relationships? Can this be achieved through simple extensions to the input vector, or will we need to make changes to the network's structure? and
- By relating sentence weights to the levels of activation within a semantic network, can we partition the network's sentence base into a set of smaller training sets? Put another way, can we implement a local, or context-specific, form of learning that increases the number of terms we can acquire by skipping over those entities and concepts that never co-occur?

6. REFERENCES

[1] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[2] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2nd edition, 2006.

[3] Thomas K Landauer and Susan Dumais. Latent Semantic Analysis. *Scholarpedia*, 3(11):4356, 2008.

[4] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

[5] Risto Miikkulainen. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. MIT Press, 1993.

[6] Risto Miikkulainen and Michael G. Dyer. Forming Global Representations with Extended Backpropagation. In *IEEE International Conference on Neural Networks*, volume 1, pages 285–292, 1988.

[7] Timothy T. Rogers and James L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press, 2004.

[8] E. Rosch, C. B. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic Objects in Natural Categories. *Cognitive Psychology*, 8:382–439, 1976.

[9] W. N. Venables, D. M. Smith, and the R Development Core Team. *An Introduction to R (Version 2.9.0)*. Available from `http://cran.r-project.org/doc/manuals/R-intro.pdf`, 2009.