

Minors as Miners

Modelling and Evaluating Ontological and Linguistic Learning

David M W Powers

AILab, School of Computer Science, Engineering and Mathematics
Flinders University of South Australia
PO Box 2100, Adelaide 5001, South Australia

David.Powers@flinders.edu.au

Abstract

Growing up is in large measure learning about the world and our social and linguistic environment. We might call this data mining, although it is far more multimodal and immersive than most applications. This paper describes computational research into how children learn, with a particular focus on evaluation in both supervised and unsupervised paradigms.

Conversely, we gain additional insight into association mining by considering psycholinguistic experiments that quantify the way human association by both adults and children relate to a variety of association measures. Learning and evaluation are not dealt with in isolation, but a program of formal and application-based evaluation is expounded and exemplified to show how to evaluate discovered patterns with and without a gold standard.

In this context, some serious issues with current evaluation techniques and accuracy measures are identified and the unbiased techniques identified.

Keywords: Natural Language Learning, Data Mining, Text Mining, Signal Processing, Speech Processing, AudioVisual Speech Recognition, Cognitive Linguistics, Computational Psycholinguistics, DeltaP, Receiver Operating Characteristics, Bookmaker Informedness and Markedness, Brain Computer Interface.

1 Introduction

Over the last 60 years, human-like performance by computers in tasks requiring broad cognitive and linguistic competence has remained elusive. In many specific areas, solid algorithms and useful methodologies have been developed and been hived off from Artificial Intelligence as fields in their own right, or have emerged independently from the seeds of AI. In the 70s and 80s Cognitive Science emerged as an interdisciplinary nexus that took over the traditional psychological modelling of AI in the 50s and 60s, leaving AI to become increasingly algorithm oriented and focussed on engineering goals. On the other hand, Computational Intelligence emerged to espouse the softer fuzzier aspects that the AI community seemed to be resistant to (leaving behind GOFAL, Good Old-Fashioned AI). These fuzzier aspects included Fuzzy Logic, Neural Networks, Ant Colony Optimization,

Genetic Programming and a variety of other approaches based on biological or physical metaphors, whilst Cognitive Science explored sometimes similar models based on stronger ideas of biological plausibility.

This paper is written in the context of a program of research undertaken by the author since the mid-70s, focussed on the idea of getting computers to learn to understand the world, and language, the way a baby does. However, this paper will not attempt a logical or chronological development of Computational Psycholinguistics, but will focus specifically on aspects of relevance to Data Mining, including in particular *evaluation*.

2 Evaluation

How do you evaluate patterns? Where we are doing *unsupervised learning* with no teacher to guide us with appropriate examples and no marker to grade our efforts, how do we know how useful our patterns or rules are? Where we are doing *supervised learning* and are aiming to achieve a specific objective, how do we rate our system and which of the many competing measures do we use? And how do children rate the patterns and rules they learn?

2.1 Evaluation in Applications

One answer to the problem of evaluation in unsupervised learning is to find and use an appropriate gold standard – which begs the questions of where this comes from, how reliable it is, and if it is reliable why we are bothering with trying to learn it. If the answer is that it is one person's theory, then it is inherently subjective and begs the circular question as to how that theory was evaluated.

Another answer is to turn it into a supervised problem with measurable outcomes. Thus phonologies, grammars, ontologies, etc. may be evaluated by embedding them in an application where there is inherent and objective performance evaluation – for example in web search, machine translation, speech recognition, lip reading, electroencephalographic computer interface, etc.

If the question is which paper is more relevant, or which gloss (translation) of a word is more appropriate, there is usually little doubt unless they are roughly equally good, and human raters are well qualified to make these judgements. On the other hand, if it comes to deciding between two competing grammars, it seems that there are more grammars than linguists, and none of them are likely to have much relationship with what goes on in our heads.

A third approach is to appeal to some concept of parsimony. The child's problem of learning about the

world is very similar to the scientist's problem of learning about the world, and good scientific method has specific biases to theories that are simple and testable. However parsimony and testability relate to theories that are already shown to be equally good on some objective evaluation.

Evaluation measures based on parsimony tend to have an information theoretic basis, using overall evaluation paradigms like Minimum Message Length, or local measures employing conditional entropy or mutual information. In many cases, such as log-likelihood models, this use is blind to the effectiveness of the outcome and more about significance. In other cases, such as in building decision trees, the usage is more like a heuristic and aimed at building a smaller model rather than a more correct model. In both cases, it recognizes that improved performance from overtuning is misleading.

2.2 Supervised Evaluation & Gold Standards

Although the focus in this paper is unsupervised learning and data mining, we commence by examining evaluation in the context of supervised learning, as well as association learning as investigated in children. We will however in the process relate this to unsupervised learning, clustering and association rule mining before considering these paradigms closely in the next section.

We will consider the value of rules that predict a Result **R** based on a Precondition **P**, where we assume **P** and **R** take the same labels representing the predicted class and the real class. In the binary or dichotomous case we have in evaluating a single rule **P**→**R**, **P** or **R** may take only the two values true (+) or false (-). Table 1 shows two standard notations for labeling the contingency table showing the 4 combinations possible. Both of these are used in upper case variants summing to N and lower case variants normalized to probabilities that sum to 1, with the first version being mnemonic (e.g.true or false positive, predicted or real negative).

Precision, known as Confidence in data mining, is a form of accuracy based on the proportion of positive predictions that have correct outcomes (true positive accuracy, $tpa = tp/pp = TP/PP$). Recall measures the rate of finding positives and is the proportion of positive outcomes that have correct predictions (true positive rate, $tpr = tp/rp = TP/RP$). Support measures $tp = TP/N$, which is proportional to Recall as RP and RN and N are assumed to be constants related by fixed Prevalence $rp = RP/N$ and Inverse Prevalence, $rn = RN/N$, being the set of real marginal statistics. One of the sources of problems in evaluation is that the prediction marginal statistics are not in general fixed and act as biases, Bias $pp = PP/N$ and Inverse Bias $pn = PN/N$.

Severe bias problems with Recall and Precision have been demonstrated by Powers (1997 with Entwisle, 2003, 2007 and 2008) from a theoretical and empirical statistical perspective (proposing Informedness and Markedness), Perruchet and Peereman (2004) from a theoretical and empirical psychological perspective (proposing DeltaP and DeltaP'), and Flach (2003 and 2005 with Fürnkranz) from a theoretical machine learning perspective (proposing the concept of skew and WRAcc). Similar issues with Confidence and Support in

association mining date back equally far with for example Brin et. al (1997) proposing Conviction and Interest (aka Lift). Note that Lift ($tp/[pp \cdot rn]$) is a ratio of actual frequency to expected frequency (joint probability to product of Prevalence and Bias) and Leverage is the difference between actual and expected frequency, being proposed by Piatetsky-Shapiro (1991) even before the classic advocacy of support in Apriori (Agrawal et al., 1993). Note further that pointwise Mutual Information uses $\log(\text{Lift})$ to assess individual rules. Conviction ($[pp \cdot rn]/fp$) is the *reciprocal* of Lift applied to the cell/rule **+P**→**-R**, reflecting a desire not only to see that tp is *high* relative to chance, but that fp is *low* relative to chance. However, recall that rp and rn are constants of the dataset that are assumed to apply to any random sample (from the dataset or collected in the future). This means that Lift is equivalent to Confidence or Precision apart from a linear scale factor (which may be useful if thresholds are employed). Similarly, Lift is equivalent to $1/[1-\text{Confidence}]$ and thus is equivalent to Confidence or Precision or Overgeneralization (see below) except for the non-linear scaling (which may be useful if thresholds are employed).

Other measures that normalize fp are Fallout (false positive rate, $fpr = fp/rn = FP/RN$, the proportion of negative outcomes that incorrectly have positive predictions) and Imprecision = $1 - \text{Precision} = 1 \div \text{Overgeneralization}$ (false positive accuracy, $fpa = fp/pp = FP/PP$, the proportion of positive predictions that incorrectly have negative outcomes). It is also possible to normalize fn as Missrate or Inverse Fallout (false negative rate, $fnr = fn/rp = FN/RP$) and as Inverse Imprecision (false negative accuracy, $fna = fn/pn = FN/PN$, the proportion of negative predictions that incorrectly have positive outcomes). Similarly tn has normalizations corresponding to Inverse Precision and Inverse Recall reflecting the result of application of Precision and Recall to the inverse rule **-P**→**-R**, and all Inverse measures are interpretable this way and are also complements of other named measures.

This Inverse problem is technically a different (dual) problem as it uses the rule in the opposite way to its logical intent (it is abductive and equivalent to **P**←**R**). However under conditions of forced single choice it is effectively used this way by virtue of the closed world assumption (if we don't say it's positive it is negative and vice-versa, which is typical of a neural net or decision tree, but not of association rules, as discussed below).

	+R	-R					
+P	tp	fp	pp	+P	A	B	A+B
-P	fn	tn	pn	-P	C	D	C+D
	rp	rn	1		A+C	B+D	N

Table 1. Systematic and traditional notations in a binary contingency table. Colour coding indicates correct (green) and incorrect (pink) rates or counts in the contingency table. Left table is systematic terminology based on true/false/real/predicted positives and negatives and in lower case represents probabilities and in UPPER case counts. The right table is an common alternative notation.

2.2.1 Informedness, Markedness & Correlation

We now introduce Informedness (DeltaP' or skew-insensitive WRAcc) and Markedness (DeltaP). Informedness has been advocated by several authors under its various names as discussed previously, and shown to be unbiased, corresponding to the probability of making an informed decision versus a chance decision (Powers, 2003). Shanks (1995) calls DeltaP "the normative measure of contingency" in that it explains human association data so much better than direct unnormalized measures such as Precision or Confidence.

In the dichotomous case we have been discussing,

$$\text{Informedness} = \text{Recall} - \text{Fallout} = \text{Recall} + \text{InvRecall} - 1 \\ = [\text{Recall} - \text{Bias}] / \text{Inverse Prevalence}.$$

We can thus see that it takes into account Fallout (f_{pr} a constant scaled normalization of f_p) as well as Recall (t_{pr} a constant scaled normalization of t_p), that it reflects equally both the forward and inverse problems, that it is effectively a (constant scaled) renormalization after subtracting the Bias. Although directly based on Recall-like measures, the difference of t_{pr} and f_{pr} , Informedness is also qualitatively similar to Precision as the ratio of t_p and f_p . When Precision is 1, $t_p=p_p$ and $f_p=0$, so that $f_{pr}=0$, and Informedness = Recall. When Recall=1, Informedness = InverseBias/InversePrevalence and is thus only maximized when Bias=Prevalence. This matching of Bias to Prevalence is a common heuristic.

The dual of Informedness is Markedness or DeltaP:

$$\text{Markedness} = \text{Precision} + \text{Inverse Precision} - 1 \\ = [\text{Precision} - \text{Prevalence}] / \text{Inverse Bias}.$$

This is thus based on the Precision-like measures but has some similarity to Recall.

Both Informedness and Markedness are unbiased unlike other common averages of Precision and Recall, the F-Factor and Rand Accuracy. Flach's skew insensitive version of F-Factor and Precision remain similar, whilst the skew insensitive (skins) version of Accuracy and Weighted Relative Accuracy (WRAcc) are equivalent to Informedness (BMI) and the ROC area under the curve (AUC) with the relationship:

$$\text{skinsAcc} = \text{AUC} = [\text{BMI}+1]/2 = [\text{skinsWRAcc}+1]/2.$$

Informedness and Markedness are not in general independent as they are regression coefficients for dual problems, and thus by definition their geometric mean is Correlation (Perruchet & Peereman, 2004; Powers, 2007 & 2008). The correlation of Informedness and Markedness themselves tends to be about 0.5 in a study by Perruchet & Peereman (2004) that investigated the learning of phonological associations in children and adults and found that Frequency, Markedness, Informedness and Pearson/Matthews Correlation (their geometric mean) correlated significantly more strongly with children and adult performance than Precision and Recall, in the indicated order of increasing correlation ($p<0.005$ in all cases for adults, and $p<0.05$ in all cases for children, where the difference was less marked). This indicates that we learn associations based on both forward and backward predictability (corresponding to both classical and operant conditioning) but give slightly more weight to the forward direction (using well marked predictors to successfully predict well marked outcomes).

2.2.2 An Example of Need for Informedness

Precise formulae and vague statements about bias are neither of them enough to give a good feel for how serious the problem is with Precision and Recall, or Confidence and Support, or F-Factor and Accuracy. Thus it is appropriate to give some examples. In the discussion of the various experiments below we will see examples where Accuracy increases and Informedness decreases – and this has been mind-blowing for the students working on the project who were sceptical about technicalities.

However, here I will go into one example, which was one of those that originally inspired the development of Informedness and is discussed in Entwisle and Powers (1997): the problem is *when water is a noun or a verb*. The real problem is that several real-life parsers and taggers made the deliberate decision that their systems were so bad at deciding part of speech that they could increase their F-scores and/or Accuracies by saying water was always a noun, which it is 90% of the time. It is instructive to do the maths and see that chance level for guessing with Bias matching Prevalence gives Recall = Precision = Bias = Prevalence = 90%, Inverse Precision = Inverse Recall = Inverse Bias = Inverse Prevalence = 10% and F-Factor is thus 30% and Rand Accuracy 82%. However, by saying it is always a noun we set Recall = 100%, Precision = 90%, F-Factor = 95% and Rand Accuracy = 90%.

2.2.3 A Feel for Informedness & Markedness

For Informedness and Markedness any form of guessing, whether following Prevalence or some other random Bias, whether always setting positive or always setting negative, always gives the same long term result: an expected value of 0. For a perfect performance with no errors, Informedness and Markedness will both be 1. Informedness tells you the proportion of the time your predictor made an informed (correct) decision versus guessed (and averaged the expected value over time). Markedness tells you the proportion of the time the outcome actually marked the predictor correctly (gave rise to the symptom or indicator), as opposed to it taking a random value (viz. expected value over time).

2.2.4 The General Case (K>2)

In the above we considered a single rule **P→R** where each of the variables was restricted to take a Boolean or dichotomous value. In general, there may be more than one choice or more than one rule. To the extent that the rules are independent, this latter point need not concern us as we can calculate Informedness and Markedness separately for each rule. To the extent that we can build additional evidence for a particular decision we are moving beyond the paradigm of the contingency table. In fact, we can combine weightings for different rules in any way we like, but this introduces the concept of cost, where as the Bookmaker and ROC principles behind Informedness and Markedness are based on an unbiased model with skew or costs determined by relative prevalence – the more likely a horse is to win, the lower the odds the Bookmaker will give you. Adding different costs or penalties to specific outcomes, changes the biases that are appropriate to achieve an optimum payoff. So we will ignore this for the time being, and revisit the question

of multiple overlapping rules and their effect on cost/skew.

Here we will deal with only the fact that contingency tables may be any size, and for $K \times K$ tables with $K > 2$ the above formulae don't work. In fact the modification is fairly simple – we perform a weighted average of the Informedness or Markedness determined for a single label. For each label we effectively have a binary contingency table regarding whether that label was the prediction and whether it was the outcome. For Informedness we weight by Prevalence, and this is why Informedness tends to be of most practical value, and can be empirically significantly more predictive of human performance (Perruchet & Peereman, 2004). For Markedness we weight by Bias.

The original Bookmaker Informedness derivation (Powers, 2003) was expressed in probabilistic notation in a form that weighted over a table of costs associated with the respective cells of the contingency table, with the costs being determined by Bookmaker bets and payoffs (for fair odds). Mutual Information is similarly a weighted average based on an information theoretic value equivalent to $\log(\text{Lift})$ as discussed above (Powers, 2003). Multiplying this Mutual Information by N in the general case (Powers, 2008), gives the χ^2 significance (log-likelihood or G^2), whereas multiplying Correlation by N in the binary case (Perruchet & Peereman, 2004) fits the χ^2 distribution. For the general case it is necessary to multiply the Correlation by $(K-1)N$ to approximate χ^2 and Powers (2003) derives corrections that allow greater accuracy. In addition χ^2 significance can be calculated separately in a similar way directly from Informedness and Markedness, and confidence intervals can also be estimated directly (Powers, 2007 and 2008).

The general Informedness and Markedness formulae can be elegantly expressed in ways which clearly show the simplified dichotomous form (the original formulation did not reveal the simple connections with Recall, Bias and Prevalence, or with WRAcc and DeltaP'). We define Bookmaker Informedness (BI) and Bookmaker Markedness (BM) as follows, and we refer to the secondary sums we average over as the dichotomous Informedness $B(l)$ and Markedness $M(c)$, and the weighted terms as $BI(l)$ and $M(c)$:

$$BI = \sum_{l \in P} \text{Bias}(l) \sum_{c \in R} \text{Recall}_l(c) \frac{\text{Prev}_c(c)}{\pm \text{Prev}_l(c)} \quad (1)$$

$$BM = \sum_{c \in R} \text{Prev}(c) \sum_{l \in P} \text{Precision}_c(l) \frac{\text{Bias}_l(l)}{\pm \text{Bias}_c(l)} \quad (2)$$

$$\pm \text{Prev}_l(c) = \text{Prevalence}(l) - (c \neq l) \quad (3)$$

$$\pm \text{Bias}_c(l) = \text{Bias}(c) - (l \neq c) \quad (4)$$

The cost factors, the reciprocal probabilities represented by the \pm terms in the denominators, will have a sign depending on whether the prediction was accurate (rewarded) or inaccurate (penalized). In the dichotomous case it will cancel with its numerator as the stake you lose is what the bookmaker wins and amount you win is what the bookmaker loses – notice that this cancelation says the expected win is +1 and the expected loss is -1 (sum of prevalence times a payoff is the weighted average).

However, in the multi-horse case, for Bookmaker Informedness for the classic “edge”, your loss if you lose is dependent only on the odds for the horse you bet on (l), not the horse that actually won (c). Your win (expected +1) comes at the expense of many losers, and your loss (expected to be better than -1) is not the whole of the winner's gain. For Informedness your risk is determined by your prediction (which horse you bet on), not your outcome (which horse won). Markedness reverses this as if the outcome was the bet and the prediction determined the payoff (in practice odds *are* set by bias in the bets).

Note that for the dichotomous case you have only one degree of freedom and are making only one binary decision, so $BI = B(l)$ and $BM = M(c)$ for both positive and negative cases.

2.3 Unsupervised Evaluation by Gold Standard

The previous examples assumed a Gold Standard, viz. that we knew what the correct answers were, implying a supervised training and test set. But in fact we can do unsupervised training and still test with a Gold Standard if one exists – often this will be a small hand tagged corpus, perhaps tagged by multiple annotators to allow testing for subjective interannotator differences. In such a case the Informedness and Markedness calculations can proceed as above, effectively discounting for the chance baseline. We can also calculate Informedness and Markedness between two annotators and choose the one with greater Informedness as our Gold Standard (this will be the Markedness for the other annotator). This provides a human baseline which is not discounted automatically, and in fact it is often treated incorrectly as an upper bound – some of the experiments reported here exceed human performance.

2.3.1 Hard Clustering

One particular form of unsupervised learning is clustering, and there are a wide range of techniques that can be used to compare clusterings, or clusterings with a Gold Standard (Pfitzner, Leibbrandt & Powers, 2008). Some of these are based on pair counting (how many of the possible pairs occur in the same cluster for both clusterings), and the pair counting results themselves can be put into a contingency table allowing use of all of the measures we have been discussing.

However, a small set of clear cases can be used to match up unsupervised clusters and Gold Standard classes, for purposes of evaluation (generally, the classes or clusters are wanted for some purpose and can be used directly without use of seeding by a Gold Standard, but sometimes specific classes are required and the best possible matching is required for our application – a good evaluative application will not have this semisupervised requirement).

The question is how to match these up. The original Bookmaker paper (Powers, 2003) dealt with this form of unsupervised learning, assuming a hard clustering (every case is a member of exactly one class with no fuzzy membership function) and determining that for each cluster it is allocated to the class which it had the highest probability of labelling correctly – that is the highest number of instances, Precision or Confidence in the row. For a number of classes $C > K$, we would expect to

sometimes get more than one cluster contributing to a class, and even with $C=K$ this will be the case if one class doesn't get assigned a cluster.

However, it is appropriate, in supervised or semisupervised approaches, to use Informedness directly as the measure to optimize, rather than some arbitrary heuristic. Equation 1 can be applied pointwise to each unsupervised rule or cluster u we are considering adding to the support for a particular class label l . At this point we are assuming hard clustering and hard classification – a given data point or situation is in exactly one cluster and exactly one gold class, and we will assign it exactly one label. Omitting a class effectively includes it with a chance level informedness of 0, so we multiply $B(l)$ terms, the dependent internal sum of (1), by the true bias, $Bias(l) = p(l)$, according to (1), where all probabilities are calculated relative to the total number of items, N .

The internal sum involves weighted recall, where $Recall(c) = p(l|c)$, and will need to be accumulated across all clusters assigned label l . This allocation has usually been done by some heuristic that may introduce a bias, e.g. Powers (2003) used the most popular label in each cluster (or weighted them if equal), which corresponds to maximizing precision for each cluster. Equation (1) seems to suggest we should maximize Recall, but it is not so simple because of the weighting by the Bias, Prevalence and Cost terms. Moreover, even maximizing the pointwise Bookmaker $B(l)$ terms is not sufficient due to the $Bias(l)$ weighting. Empirically, maximizing Precision works better than Recall. However to maximize unbiased cost benefit of the predictions it is recommended to optimize for BI. It is also convenient that the cost factor is independent of the labelling of u for BI (1), although the biases do depend on u and l for BM (2).

If we seek to optimize the cluster labelling iteratively or recursively, by exposing the probabilities underlying $Bias(l)$ and $B(l)$ it is clear that both factors, $p(l)$ and $p(l|c)$, will be incremented, by respectively $p(u)$ and $p(u|c)$, so we must not maximize $p(u) \cdot p(u|c)$, but rather the increase in $p(l) \cdot p(l|c)$ that would be achieved by making the $u=l$ assignment. Viz. Maximize $\Delta BI(l) = BI(l) - BI(l+u)$.

This is reminiscent of Ward's method, which empirically usually gives the best discrimination in clustering, where we effectively consider the effect of merging two clusters (in this case l and u) rather than using direct distance measures (Powers, 1997a).

2.3.2 Soft Associations

Soft clustering allows items to be in more than one class, and often associates a weight. Data-oriented methods can associate items according to relative distance from cluster centroids. Fuzzy classes or sets can have weights or membership functions that express degree of membership in a class or applicability of a label. Bags allow multiple instances of an item in the same class, which can also therefore be represented as a set with associated counts, also interpretable as a form of weighting. Association mining will in general allow items or itemsets to predict multiple distinct items, and conversely some items will never be predicted. Strengths associated with rules are also a form of weighting.

All of these paradigms take us away from the contingency table with its assumptions of mutual exclusion between categories, its binary yes/no nature, and its effective closed world assumption – anything not stated to be true is false and vice-versa. In referring to a conjunction of items rather than items, we move to a contingency table on the powerset of the items in which many items do not occur.

Under the constraint that there is a weight of 1 associated with each label and class, a contingency table can be produced by accumulating weighted membership information (e.g. predicted Coke or Pepsi \rightarrow Coke 0.6, Pepsi 0.4). Also for any labelling, such a normalization constraint can be achieved for a set of latent classes by finding the eigenvectors using singular valued decompositions or similar algorithms (Powers, 1997a).

The problem of unselected or underselected items, labels whose weights don't sum to one, can be solved by simply including an additional 'no-prediction' group used as a class for items that don't predict anything particular (predicts just about everything at near chance levels and hence not significantly), and a label on dummy rules for items that are never predicted (predicted at close to chance level and hence not significantly). These dummy classes will reduce Recall and Precision, Informedness and Markedness, to correct levels – without them they would be overinflated.

There are also some issues that can most easily be clarified using the notation of clausal logic – reversing the traditional form of an association rule to the traditional form of clausal logic:

Coke \leftarrow Frozen Fish \wedge Frozen Chips

Pepsi \leftarrow Frozen Fish \wedge Frozen Chips

do not mean quite the same thing as

Coke \vee Pepsi \leftarrow Frozen Fish \wedge Frozen Chips

which implies Coke and Pepsi are alternatives rather than being independent. This is the difference between non-Horn and Horn notations, with the Horn limitation to exactly one prediction per item requiring explicit statement of exclusions (e.g. \neg Pepsi). The assignment of weights can thus add even more precision, but has essentially the character of alternation when they are conditional probabilities that sum to 1:

Coke(0.6) \vee Pepsi(0.4) \leftarrow Frozen Fish \wedge Frozen Chips

In this case we seem to have confidence- or precision-like weights, indicating what proportion of the time we predict Coke or Pepsi, once this rule fires, but the true precision for Coke and Pepsi based on this rule involves multiplying by the precision of the rule, the probability that the rule is correct, irrespective of weights. If the weights sum to more than one, it indicates that some households buy both Coke and Pepsi and they are not totally mutually exclusive.

However, as we have been discussing, Precision is misleading because it reflects overall Prevalence – the fact that 80% of people buy Coke and only 20% by Pepsi would seem to mean that people who buy Fish and Chips are *less* likely to buy Coke! Nonetheless these are appropriate as the respective weights for the original pair of rules (and thus precision and prevalence) in reporting

results in a contingency table, and we can then calculate total or pointwise Informedness in the standard way.

Informedness tells you what proportion of sales you have predicted rather than guessed, while Markedness tells you what proportion of bought products are markers of other needs rather than chance associations.

3 Language Technology Applications

We now return to our primary focus on text mining and unsupervised learning of linguistic and ontological rules and categories. We will review the work in this area bottom up, starting from raw audio, video or character data, starting with textual input.

3.1 Structural Learning

Notice the emphasis on character data – that is the form in which text comes, and conventions about words and spaces are not universal and not reliable, so even for English there is some effort required to establish what the words are. This is the word segmentation problem and relates to other specialized problems such as named entity recognition (International Business Machines), other similar noun collocations that aren't entities as such (Object-Oriented Programming), separable and inseparable verbs involving particles (put up X, put X up, put up with), and composite content and function words (object-oriented, 'in your face', to day, vs to-day vs today, into vs out of). Note the convention of either hyphenating or quoting when a phrase is pressed into service as an adjective. Note that spaces and quotes tend to moderate to hyphenation and eventually disappear as a phrase becomes accepted as a word.

A similar problem occurs in Chinese where the characters are like English morphemes or syllables, and content words normally consist of multiple characters. Spoken English doesn't come nicely packaged into words either, and we are increasingly wanting to work with spoken language. As we aggregate units into bigger units, segmentation becomes the basis for a kind of structural learning that encompasses the phonological, morphological, grammatical and prosodic aspects of language – all without any semantic information.

Techniques based on conditional entropy (confidence- or precision-like measures) are fairly good for assessing how likely the next character or syllable is to be part of the word, and techniques based on mutual information (leverage-like) are good for determining boundaries of words or other higher level units (Magerman, 1991). The combination of the two techniques can be even more powerful (Huang & Powers, 2004). These techniques can also be used to detect affixes and clitics – functional words, prefixes and suffixes, which are important foundations for full syntactic analysis. With just this information, entire sentences can be parsed (Entwisle, 1997) without knowing the actual content words: cf. the slithy toves did gyre and gimble in the wabes (Lewis Carroll, *Alice in Wonderland*).

These functional elements (morphs, which include also the null morph or null inflection \emptyset) have also been used to achieve effective spelling correction (Powers, 1997b; Huang and Powers, 2001) and have similar applications in disambiguation of confusable words for

speech recognition, optical character recognition, sign language recognition and machine translation.

The approach taken here is pure text mining and reflects several of the issues discussed in section two. Each functional context (e.g. for the previous sentence from Alice: the $-y$ $-s$ did $-\emptyset$ and $-\emptyset$ in the $-s$) provides a solid grammatical basis for distinguishing confusable words or multiple meanings that can be disambiguated on the basis of part of speech – it deliberately avoids semantic information.

This reverse approach of clustering based on contextual information is also powerful and by allowing pairs or triples, implicit segmentation can be performed while categorizing characters or words into classes (Powers, 1983, 1991). By replacing segments by non-terminal symbols, a finite or context-free grammar can be induced, including left-, right- or centre-recursive rules by allowing the proposed non-terminal to be included in its own contexts (Powers, 1992), although generally the non-recursive grammars were found to be more stable.

Given the assumption of word segmentation, functional words can be reasonably well distinguished by frequency alone – the 150 most frequent words of English constitute about half of any text, and are mainly functional words with a primarily grammatical function, or placeholder words (like thing, person or place) that have a similar function. Again we can generalize and collect sequences of words (out of) that are very frequent, and define these as templates that connect to an adjacent word ('the X' or 'out of the X' are both templates that characterize nouns, although the second has stronger implications that X refers to some place). It is also possible to have contexts with two separated open class and/or two separated closed class words ('the X of Y').

In child speech and child-directed speech, many sentences have this character, whereas in adult speech sentences will tend to combine several phrases and/or clauses each of which have this character. The child already recognizes key aspects of their native language by the time they are born, and at a very early age English-exposed babies can be shown to be sensitive to these "closed class" functional words that characterize English. There are also intonational and voicing features that characterize these closed class words as functional in nature – they tend to be less stressed, and for English words that start with the voiced /dh/ sound of 'the' are *all* functional words and constitute around 20% of a typical text.

By using a corpus of conversations with children (CHILDES) it is possible to pick up frequent templates that are entire utterances and have forms like those illustrated above. It is hypothesized that children listen only to those template-bound portions of any sentences that are currently too complex for them. But as templates become units in their own right, more complex templates involving them can be learned, including recursive usages. Leibbrandt (2008) has successfully modelled this process.

3.2 Semantics and Ontology

Powers (1983, 1989 with Turk, 1992) argued that structural learning can go only so far in learning language, and that it is necessary to learn about the world,

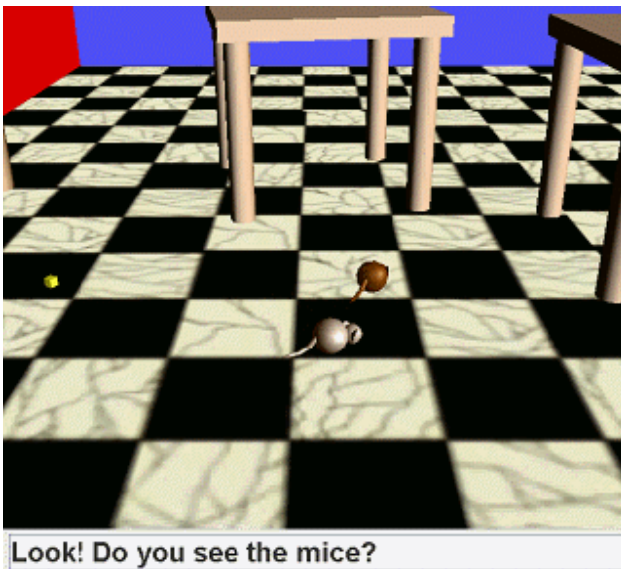


Figure 1. Example teaching scenario in MicroJaea robot world developed for teaching computer syntax and semantics, and also used by the Teaching Head.

Reproduced by permission of DMW Powers and R Leibbrandt

introducing the word ‘Ontology’ into Artificial Intelligence and Natural Language Processing from Philosophy where it denotes the study of what is and the development of models or theories of the world. Children are like scientists, finding patterns and testing theories, and this applies both to the structural aspects of language and to the way they structure and make sense of their world. The thesis that it is not possible to learn language fully without this kind of knowledge of the world later came to be called symbol grounding – as symbols are meaningless until grounded in reality (Harnad, 1990).

Powers and Turk (1989) also claim that this grounding contributes further to overcoming the so-called Poverty of the Stimulus problem touted in the 1980s by Chomskian linguists, but that there is no theoretical necessity to require such grounding to learn syntax, exposing a paradigm they called anticipated correction to explain how children could recognize that erroneous utterances didn’t sound right. Syntax by its nature is just rules that are slavishly obeyed by speakers and hearers in producing and interpreting language – if the meaning of the words is known in context, and hence the grammatical role of the word is clear, then syntax determines the grammatical rules of ordering the words as well as the cohesive connections between words (such as agreement and anaphora). Ambiguity in context is rare, and is usually corrected or repaired before or shortly after completion of the utterance, or observed with a wry “pun not intended”, or is a deliberate pun or a related form of humour.

Another important aspect that relates to semantics and ontology is metaphor. This is not just a term for tired phrases your English teacher explained to you, but is at the heart of how both language and learning work. Nothing is ever exactly the same, as time marches on, so does age, decay, dust, etc. It’s called entropy! So we are always classifying things as similar rather than dissimilar, and the classes we come up with have to do with the prevalence of the different exemplars and the need for particular features to contrast functionally different things

– two different fruit or two different people. If all apples or all oranges, or all Asians or all Caucasians, look the same, that’s because of lack of experience of naming them apart for some purpose.

Clustering is about grouping things together that are similar, and the density of clusters tends to relate to the density of items to be clustered – the more items in a region of attribute space, the more clusters as the smaller the thresholds on distance between member and non-member. This is why absolute thresholds are inappropriate, and it also brings into question nearest neighbour type algorithms. However, Powers (1991, 1992) is the only algorithm I know of based on cardinality rather than distance. On the other hand, self-organizing maps (e.g. Kohonen maps) do self-organize with a cluster area that is a direct function of density.

Powers (1983) introduced the idea that the same grammars and learning algorithms we use to learn language, we can use to learn about the world, specifying the development of a robot world simulation (Hume, 1984) using a grammar like notation that allowed learning meanings of nouns and verbs (Powers and Turk, 1989) as well as prepositions (Homes, 1994). The current version (Leibbrandt, 2008) is illustrated in Fig. 1. In this view, we have mechanisms that are designed to learn about the world and when we use them to interpret utterances we will do best with ones that exhibit the same biases we have in the world, the part-whole kind of structure, objects holding together rather than persisting in parts that move around independently, the various conservation laws. Thus we would expect grammar to derive from ontological learning rather than independently.

The field of Cognitive Linguistics is based on the centrality of the idea of metaphor, and distances itself from the Chomskian claim that language is a separate modality, claiming that it is integral with and inseparable from the rest of our cognitive processing. In particular, Deane (1992) extensively developed the idea of the part-whole nature of grammar deriving from the part-whole nature of the world.

But how we learn about the world and language should also be useful for teaching about the world (inc. maths, science, numeracy) and language (inc. literacy).

Recently however, the term Ontology is being used to describe taxonomies, thesauri, semantic networks, and text mark-up based on these. These are not truly grounded and don’t correspond to a true semantics, but to a pseudosemantics or logical semantics. Nonetheless they

Noun Similarity	r	r²
Resnik*	0.791	0.626
Jiang & Conrath	0.828	0.686
Lin*	0.834	0.696
<i>Average Human</i>	<i>0.902</i>	<i>0.814</i>
Yang & Powers*	0.921	0.848
Verb Similarity	r	r²
Yang & Powers	0.833	0.694
<i>Average Human</i>	<i>0.866</i>	<i>0.751</i>

Table 2 Comparison of results of published noun and verb similarity algorithms using Wordnet or Roget.

*Difference versus *human baseline* significant to $p < .05$.

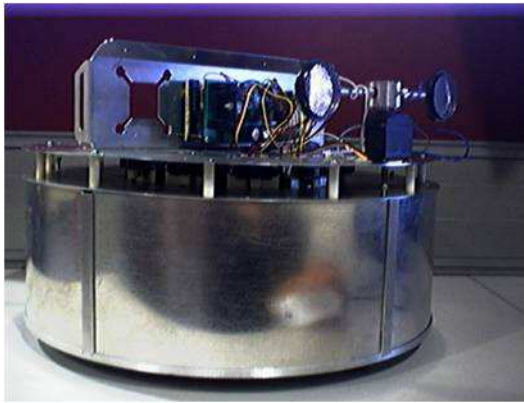


Figure 2. Can Robot with Sonar, Infrared and Webcam tracking options, plus can mount additional webcams and additional stages surmounted by a laptop – this has been used for Wizard of Oz building guide research.

can be useful, and we have explored algorithmic use of such WordNet to determine word similarity, as well as unsupervised self-organization of semantic networks and thesauri, achieving results that are comparable with average human subjects versus averages across subjects and published gold standards in fact significantly achieving significantly better than human performance for nouns (Yang and Powers, 2005) and breaking new ground for verbs (Yang and Powers, 2006b) – see Table 2. This has translated to extremely high accuracy in selecting the correct gloss for French to English Machine Translation, and is being explored as an automatic mechanism for disambiguation for the speech recognition, emotion recognition and topic selection components of the Thinking Head.

Unsupervised semantic network and automatic thesaurus construction (ATC) is difficult to evaluate, but for comparable size similarity classes relative overlap between each pair of Roget, WordNet and ATC are not significantly different. The ATC was developed using simple text mining techniques as described above, based on up to three templates for each part of speech (Yang and Powers, 2006a, 2008).



Figure 3. Mark 1 Robot Baby, 8 mikes and 8 touch sensors + 5 motors – crawl or look towards touch or sound.



Figure 4. Mark 2 Robot Baby, has 2 USB AV webcams mounted with additional eye convergence motor.

4 Heads Up!

4.1 The Talking Head

As well as working with simulated robots, it is worth trying to learning language and ontology in the real world, with real robots, sensors and actuators.

We have worked both with baby-like robots (Fig. 3-4, Powers, 2001) and with garbage-can-like robots (Fig. 2) – not to mention micromice, lego robots and a variety of other physical critters of the mechatronic persuasion. But whilst this has produced nice demonstrations, most researchers revert sooner or later to simulation to tune their systems and develop their learning algorithms, and we are no exception. Our robot baby could crawl, turn its head to the sound of a voice or a touch, and that was about it. The micromice can zoom around a maze, and the can can guide a visitor round the building. But it is too much work and too much maintenance and much too irrelevant for everyday language learning research.

4.1.1 Sensors, Signals and Fusion

On the other hand the sensors can be deployed separately – the sensors for detecting faces, eyes and lips, and then gaze-tracking and lip-reading (Lewis and Powers, 2002), the sensors for detecting objects, detecting and calibrating motion (Matsumoto, Powers and Asgari, 2008), the sensors for detecting people coming and going, their identity and their expressions and emotions (Luerssen, Lewis, Leibbrandt and Powers, 2008) – these sensors are just simple cameras and microphones, mostly cheap webcams. The Informedness measures proved particularly important in the fusion of different sources of information with different numbers of classes and different biases and prevalences, enabling a clear unbiased evaluation.

Fusion of information from multiple sources becomes a major goal when we add multiple sensors – simply throwing everything in together (early fusion) tends to produce catastrophic results (sometimes worse than either source alone). Similarly analysing each separately and then combining can lead to catastrophic fusion with a result significantly worse than the best signal alone. We have therefore worked on developing techniques that guarantee that the fusion will not be significantly worse than the best signal, and will usually be better (Lewis and Powers, 2005). This is achieved by identifying orthogonal features and training them separately, and we also have investigated techniques to automatically assess error.

Researchers have tended to become too specialized – one only works on gaze-tracking, or dialogue, text mining, or grammar induction, or speech recognition, or speaker, or expression/emotion recognition. But common techniques are used in many of these areas, and what is noise for one researcher is the goal for another. We have also used many of the techniques we developed for language and learning in biomedical image processing and brain computer interface research using electroencephalography (EEG), including unsupervised techniques for signal separation, supervised techniques for optimal fusion, and it was in this context that we first go clear examples where conventional accuracy measures were increasing but the true utility, as measured by Bookmaker Informedness, was decreasing (Fitzgibbon,

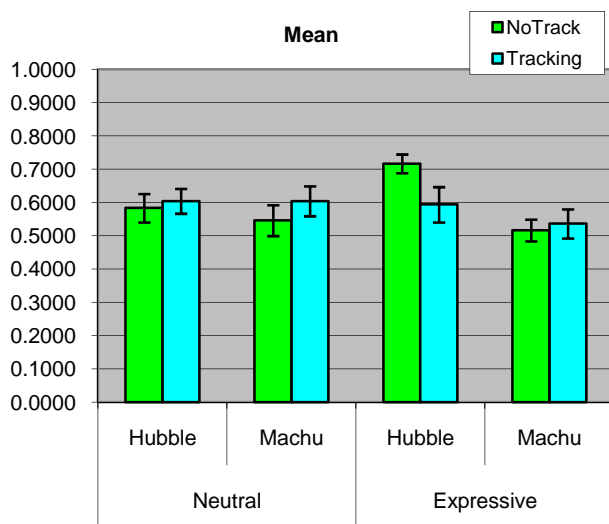


Figure 5. Up to 20% absolute gain in comprehension achieved by the Teaching Head’s students comparing the most appropriate and least appropriate gaze tracking and expression mark up. What is appropriate still needs further formal evaluation.

Powers, Pope and Clark, 2007), and much of the methodology developed in this context has been reapplied back in the speech and language area.

The original motivation for our work with EEG was to study the predictions about closed and open class words that emerged from our unsupervised learning research, as well as to understand some of the conscious and unconscious processing of speech – indeed we demonstrated a clear affect from inaudible subliminal audio (Powers, Dixon, Clark and Weber, 1996). We have also used EEG to determine where a subject is in their learning curve, investigating also how this relates to their attention and awareness available for other purposes – this is a theme that recurs in our human factors approach to user interface design. Currently we are looking at how to fuse biological signals and audio-visual signals for improved learning.

The Talking Head as a surrogate for a robot can be run on any old computer or laptop, can make use of any old webcam or even the built-in laptop camera and mike, and provide many of the benefits of a robot – being with its cameras and microphones embedded in the world, embodied notwithstanding its lack of a body, grounded although not crawling around on the ground.

4.2 The Thinking Head

Connected to a simple dialogue engine or “bot”, a Talking Head becomes a Thinking Head that can engage in conversations or act as a kiosk in a museum, describing and answering questions about its accompanying exhibit. With cameras and microphones, it immediately becomes more human-like if it can do speech recognition (in the noisy environment – using lip reading) and face recognition (recognizing people returning, or even just change of speaker and colour of shirt). By adding emotional expression to face or voice, and including appropriate eye tracking, it can become not only more human-like but achieve higher performance in getting information across (Powers, Leibbrandt, Pfitzner,

Luerssen, Lewis, Abrahamyan and Stevens, 2008; see Figure 5Figure 6).

But instead of operating in, understanding and being grounded in a real environment, like a museum, we can provide simulated grounding in a simulated environment doing simulated museum tours, or anything else we like.

4.3 The Teaching Head

What is becoming our major application for the Thinking Head at Flinders, is the Teaching Head. Our research has been focussed on teaching computers language – speech, syntax, semantics, ontology, etc. Our robot worlds and virtual environments were developed for this purpose and set up as teaching scenarios. The obvious application is to turn the scenarios around and make the computer, the Teaching Head, the teacher rather than the learner.

Our initial target for this is teaching English and German as a second or foreign language, with a particular focus on the German noun phrases, and the associated prepositional/declensional system (Leibbrandt, Luerssen, Matsumoto, Treharne, Lewis, Li Santi and Powers, 2008).

A key aspect of the Teaching Head is its hybrid environment (Figure 6) – user and environment are monitored by three cameras and the same props/toys are simulated in the virtual world so both teacher and student can illustrate sentences or obey commands. We have a 3D touch screen and are also exploring camera-based tracking, so there is no need for the user to use a keyboard and mouse – it thus doesn’t feel like you are using a computer at all!

A variety of additional teaching opportunities have emerged for the Teaching Head, including several related to health, and several related to specialist education...

4.3.1 VALIANT - Virtual Agent for Literacy and Numeracy Tutoring

The Thinking and Teaching Head have been displayed at many exhibitions and art-galleries and open days. The original head captivates with conversations about just about anything, and particularly attracts the attention of

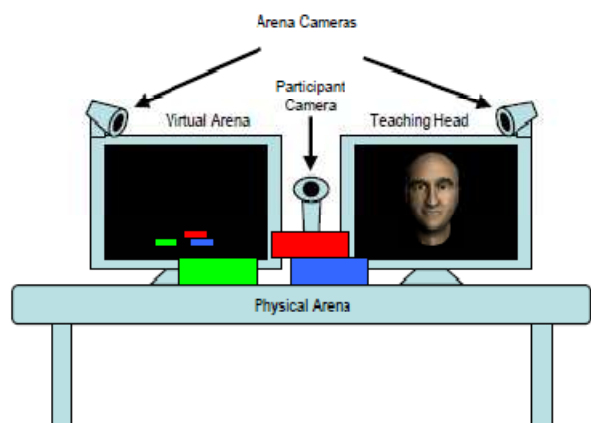


Figure 6. The Teaching Head set up classically has two screens angled at around 120° and three orthogonal webcams. The enclosed physical arena is reproduced in the virtual arena in a monitor or window – we can use a 3D scanner to scan in virtual objects or our 3D printer to print out virtual objects.

Reproduced by permission of DMW Powers and T Lewis

primary school kids, who queue up and chat for hours.

By including some very simple teaching scenarios – where the computer knows the correct answer and what to look for or listen for, it is very easy to adapt this to teaching literacy (helping with correct reading, pronunciation and spelling) and the Head gives us an advantage over other educational software, in providing a teacher-like or peer-like focus that children like to interact with, in being able to monitor what the child is doing without relying on keyboard skills, in being able to sensitively use expression and emphasis to point out gently what is wrong and how to fix it. Numeracy is even easier! A number of demonstration videos are available from the authors for both literacy and numeracy applications.

At the moment, the main interest is in relation to helping indigenous children and remote communities, including extending the system to training re basic health and hygiene issues, as well as training health workers.

4.3.2 MANA – Memory, Appointment and Navigation Assistant

We now move to the other end of the age scale, as people retire and want to retain their ability to live independently. The Thinking Head is being developed as a memory aid and calendar service, given an ability to help in emergencies, and a mobile phone version will help people find their way around town on public transport. Some of these goals are reflected in the Memories for Life (M4L) Grand Challenge of the British Computer Society (<http://www.memoriesforlife.org>), but we also have goals relating to teaching them to retain their memory skills, who their grandchildren are, enhancing their cognitive skills and maintaining their interest in sports and current affairs.

To what extent it will become the “companion” the news reports picked up on remains to be seen!

4.4 The Social Head

Many of the companion and teaching functions have particular applicability to those with disabilities, many of which have an impact on people’s ability to function socially, to learn effectively in standard classes, and to feel a useful part of society. Analogous applications are being defined here – mainly encouraging conversation and good social practice.

But it should also be noted that what is best in this respect may differ from child to child, and these differences may in themselves have diagnostic value. For example, deaf children need to focus on your face more than the conventional norms allow, but ADHD children should not be allowed to be distracted by your lip movements as this has a negative effect on their learning.

4.5 The Instructive Head

The Teaching Head is quite unique in that the linguistic and conversational aspects appeal to students interested in language rich subjects and the social and people aspects of life. The robot world simulation provides opportunity for those interested in mathematics and physics, or computer games, or multimedia and creative arts, to explore their interests by designing worlds that reflect what they want to learn or do. Then there’s the

engineering and biological sides of audio and video, speech and vision processing – there’s something for everyone...

We are now running regular workshops for schools (typically one or two a week, with versions from year 5 to year 11) focussing on one or more aspects of the project – and indeed have developed a full 10 week course for year 10 students based around these topics, allowing students the flexibility to spend more time on the aspects that interest them. Children are finding their conception of science and the idea of interdisciplinary collaboration extremely broadening.

Our focus in this aspect of the project is to address the decline in numbers of students taking mathematics and science subjects, or seeing them as relevant to their futures. We believe we are getting the point across!

5 Human Factors and Flingle Search

We have already discussed experiments and results from experiments where we have compared human performance with computer performance (Table 2) or compared human performance with specific variations in experimental conditions relating to a specific aspect of a user interface (Figure 5).

One of the main applications that has emerged as a key one both in terms of commercial interest and evaluation of language technology, as well as for human factors research, is search. We have touched on other applications including speech recognition and synthesis, emotion recognition and expression evaluation, spelling correction and machine translation. We have touched on biometric approaches to study aspects of language and learning. But so far we have not discussed search and it seems appropriate as a case study illustrating further the multimodal aspects of our research.

Our approach to Human Factors is the same as our approach to Language Learning – we don’t want to rely on introspection or theories, and we specifically decry computing researchers and software manufacturers that inflict their own ideas on ideas on others whilst ignoring decades, sometimes centuries, of psychological research. We don’t want to assume we know the answer in discovering grammar rules or syntactic categories, or the best measures or attributes to use in an algorithm or interface, so we can’t use this kind of information for training or evaluation, although we inform ourselves of relevant work from both computer and cognitive science.

In relation to search our human factors/user interface research focuses in two areas, the visual interface and the text interface.

5.1 The Textual Search Interface

Here we are seeking to answer questions about the way people use particular words, how many words they use, how these words relate to the statistics of the documents they see as relevant, and to the commonly used ranking methods, and then of course, whether these characteristics differ for search versus description, or based on experience. The short answer to these last two questions is yes, but to the others the answer is that there is not a good match between human choices and rankings, and those provided by standard algorithms (Pfitzner, 2008).

This research, along with the previously discussed Teaching Head evaluation and the visually oriented work that follows, is being undertaken in computer-delivered experiments that present search, comprehension, tracking or other tasks, and automatically collate the results. Our lean simple system is available in a heavily trafficked lab, and many experiments are also able to be delivered over the web and thus receive additional exposure (Treharne, Pfitzner, Leibbrandt and Powers, 2008).

5.2 The Visual Search Interface

In the visual interface we are seeking to improve websearch by allowing navigation of the web in a very physical way – navigating hyperspace like the Enterprise! Each word or phrase or topic in a document or corpus is a potential dimension. Generally we can reduce the dimensionality by using semantic information such as WordNet or a thesaurus, or by using dimension reduction techniques like Singular Valued Decomposition (or Latent Semantic Indexing as it is known in this context). The choice of such reduction techniques belongs in the text part of the project. But how we display these thousands of dimensions on a 2D or 3D computer screen is another question.

Yes 3D – our human factors system is deployed on a Philips WOW! lenticular screen that shows 9 different views of each pixel, giving a 3D effect. Most of the current attempts at search visualization spend a lot of effort using just the right shades and shadows and reflections and perspective to give a great 3D effect that is totally useless as an interface, whilst many interfaces also waste a lot of screen real estate (Pfitzner, Hobs and Powers 2002).

Our project is controlling, and experimenting on dimension by dimension, each attribute of the domain (words, phrases, sizes, clusters, metadata) versus each attribute of the screen, physical dimension including real stereoscopic depth, versus shading, colour and animation effects, as well as leaving size available as size. This illustrates another point – we want the best mapping possible between screen attributes and domain attributes, as well as the best choices in each case.

The results are not surprising – but they are damning of most current interfaces (Treharne et al., 2008).

6 Acknowledgements

The Thinking Head project is funded under ARC SRI TS0669874, in the context of the Australian Research Council and National Health and Medical Research Council joint Special Research Initiative in Thinking Systems. Australian Partners include the University of Western Sydney, Flinders University, Macquarie University, University of Canberra, Industry Partners include Seeing Machines Ltd, International Partners include Carnegie Mellon University, Berlin University of Technology and Technical University of Denmark.

Development of the German-language Head is supported by the Deutsches Forschungs Gemeinschaft (DFG). Some of our search research has been commercialized by YourAmigo Pty Ltd or undertaken under contract with them, and some is also being pursued under contracts with EducationAU Pty Ltd and the Australian Defence

Science and Research Organization (DSTO). Some of our EEG Learning research was undertaken under contract with DSTO, whilst other BCI/EEG research is being commercialized by BioX Pty Ltd. Our speech control technology was developed in association with I2Net Pty Ltd and the Clipsal Homespeak version of this system is being distributed internationally by Clipsal.

We also acknowledge and appreciate the assistance of the Goethe Institute and numerous schools, teachers and students.

7 References

- Agrawal, R., Imielinski, T. and Swami, A. (1993): Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD International Conference on Mgt of Data*, **22**:207-216, ACM Press.
- Brin, Sergey, Motwani, Rajeev, Ullman, Jeffrey D. and Tsur Shalom (1997). Dynamic itemset counting and implication rules for market basket data, *Proc. ACM SIGMOD Int'l Conf. on Mgt of Data*, pp 265-276.
- Deane, Paul D. (1992). *Grammar in Mind and Brain. Explorations in Cognitive Syntax*. Walter de Gruyter.
- Entwisle, Jim (1997) *A constraint parser*, Dept of Computer Science, Flinders University, Adelaide
- Entwisle, Jim and Powers, David M. W. (1997) The Present Use of Statistics in the Evaluation of NLP Parsers, pp215-224, *NeMLaP3/CoNLL98 Joint Conf. on New Methods in Language Processing and Conference on Natural Language Learning*.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996): From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*. 1-34. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. And Uthurusamy, R. (eds). AAAI.
- Flach, PA. (2003). The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics, *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003, pp. 226-233.
- Fitzgibbon, Sean, Powers, David M W, Pope, Kenneth and Clark, C. Richard (2007). *Removal of EEG noise and artefact using blind source separation*. *Journal of Clinical Neurophysiology* **24**(3):232-243
- Fürnkranz Johannes & Peter A. Flach (2005). ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms, *Machine Learning* **58**(1):39-77.
- Harnad, S. (1990) *The Symbol Grounding Problem*. *Physica D* 42:335-346
- Homes, David (1997) *Perceptually Grounded Language Learning*, B.Sc. Honours Thesis, Dept of Computer Science, Flinders University, Adelaide.
- Huang, JinHu and Powers, David M W (2004), Adaptive Compression-based Approach for Chinese Pinyin Input, *ACL SIGHAN Workshop*, pp. 24-27
- Huang, Jin Hu and David M W Powers (2001), Large scale experiments on correction of confused words, *Proc. Australian Computer Science Conference (ACSC01)*, pp77-82

- Hume, David (1984) *Creating Interactive Worlds with Multiple Actors*, Computer Science Honours Thesis, EECS, Uni. of NSW, Sydney, AUSTRALIA
- Leibbrandt, Richard, Luerssen, Martin, Matsumoto, Takeshi, Treharne, Kenneth, Lewis, Trent, Li Santi, Martin and Powers, David M W (2008), An Immersive Game-Like Teaching Environment with Simulated Teacher and Hybrid World, *Computer Games and Allied Technology*, pp217-225.
- Leibbrandt, Richard & Powers, David M. W. (2008), Grammatical category induction using lexically-based templates. *Supplement, Boston Univ. Conference on Language Development* 32, Nov 2-4, 2007, (8pp).
- Leibbrandt, Richard (2008), *Part-of-speech bootstrapping using lexically-specific frames*, PhD thesis, Computer Science, Engineering & Mathematics, Flinders Univ.
- Lewis, T. W. and D. M. W. Powers (2002). Audio-Visual Speech Recognition using Red Exclusion and Neural Networks. *Australian Computer Science Conference*
- Martin H. Luerssen, Lewis, T.W., Leibbrandt, R. and Powers, David M.W. (2008), Adaptive Multimodal Perception for a Virtual Museum Guide. *3rd Wkshp on Artificial Intelligence Techniques for Ambient Intelligence*
- Magerman, David (1991) Distituent Parsing and Grammar Induction, Powers, David M. W. and Larry Reeker, eds. (1991), *Proceedings of the AAI Spring Symposium on Machine Learning of Natural Language and Ontology*, pp122-125.
- Takeshi Matsumoto, David Powers and Nasser Asgari (2008), Webcam Configurations for Ground Texture Visual Servo, *IEEE International Conference on Cybernetics & Intelligent Systems, Robotics, Automation and Mechatronics (CIS-RAM 2008)*
- Perruchet, P. and Peereman, R. (2004). The exploitation of distributional information in syllable processing, *J. Neurolinguistics* 17:97–119.
- Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pp. 229-248.
- Pfitzner, Darius M (2008), *An Investigation into User Text Query and Text Descriptor Construction*, PhD thesis, CSEM, Flinders University.
- Pfitzner, Darius M, Leibbrandt, Richard E, and Powers, David MW (2008) Characterization and Evaluation of Similarity Measures for Pairs of Clusterings, *Knowledge and Information Systems: An Int'l Journal*.
- Darius Pfitzner, Vaughan Hobbs and David Powers (2002), A unified taxonomic framework for information visualization. *Proc. Australian Symposium on Information Visualization*, Adelaide, pp.57-66,.
- Powers, David. M. W., Richard Leibbrandt, Darius Pfitzner, Martin Luerssen, Trent Lewis, Arman Abrahamyan and Kate Stevens (2008), Language Teaching in a Mixed Reality Games Environment, *Proc. 1st International Conference on Pervasive Technologies Related to Assistive Environments (PETRA) Wkshp "Gaming Design and Experience: Design for Engaging Experience and Social Interaction"*
- Powers, David M. W. (2008), Evaluation Evaluation, *Proc. 18th European Conference on Artificial Intelligence (ECAI'08)*, July 21-25, 2008, Patras (2pp).
- Powers, David M. W. (2007), *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*, School of Informatics and Engineering, Flinders University • Adelaide • Australia, Tech. Report SIE-07-001, December 2007.
- Powers, David M. W. (2003). Recall and Precision versus the Bookmaker. *International Conference on Cognitive Science*, University of New South Wales, pp.529-534. See <http://david.wardpowers.info/BM/>
- Powers, David M. W. (2001), The Robot Baby meets the Intelligent Room, *AAAI Spring Symposium on Learning Grounded Representations*, pp59-62.
- Powers, David M W (1997a) Learning and Application of Differential Grammars, *CoNLL97: ACL Workshop on Computational Natural Language Learning*, pp88-96.
- Powers, David M. W. (1997b), Unsupervised learning of linguistic structure: an empirical evaluation, *Int'l Journal of Corpus Linguistics* 2#1:91-131
- Powers, David M W, Clark, C R, Dixon, S E and Weber, D L, Cocktails and Brainwaves: Experiments with Complex and Subliminal Auditory Stimuli, pp68-71, *Proc. of the Australian and New Zealand Conference on Intelligent Information Systems*, IEEE 96TH8234
- Powers, David M W (1991), How far can self-organization go? Results in unsupervised language learning. *Proc. AAI Spring Symposium on Machine Learning of Natural Language and Ontology*, pp131-136.
- Powers, David M. W. and Larry Reeker, eds. (1991), *Proceedings of the AAI Spring Symposium on Machine Learning of Natural Language and Ontology*, Document D-91-09 (205pp), DFKI, Univ. Kaiserslautern FRG.
- Powers, David M. W. (1983), Neurolinguistics and Psycholinguistics as a Basis for Computer Acquisition of Natural Language" *SIGART* 84, pp. 29-34.
- Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, 48A, pp257-279.
- Treharne, Kenneth, Pfitzner, Darius, Leibbrandt, Richard & David M. W. Powers (2008), A Lean Online Approach to Human Factors Research, *Proc. 1st International Conference on Pervasive Technologies Related to Assistive Environments (PETRA) workshop on "Pervasive Technologies in e/m-Learning and Internet based Experiments"* (PTLIE)
- Yang, Dongqiang and Powers (2008), Automatic Thesaurus Construction, *Australasian Computer Science Conference (ACSC2008)*, pp147-156.
- Yang, Dongqiang and Powers, David M W (2006a), Word sense disambiguation using lexical cohesion in the context. *Proc. Joint Conf. of the Inter'al Committee on Comp. Ling. and the Assn for Comp. Ling. (COLING-ACL 2006)*, Sydney Aust, pp929-936.
- Yang, Dongqiang and Powers, David M W (2006b), Verb similarity on the taxonomy of WordNet, *Proc. Third International WordNet Conference (GWC-06)*, Jeju Island, Korea. pp121-128
- Yang, Dongqiang and Powers, David M W (2005), Measuring Semantic Similarity in the Taxonomy of WordNet, *ACSC'05 Australasian Computer Science Conference*. pp315-322