Proceedings of the

# ESSLLI Workshop on Distributional Lexical Semantics

*Bridging the gap between semantic theory and computational simulations*

Edited by Marco Baroni, Stefan Evert and Alessandro Lenci

ESSLLI 2008, Hamburg, Germany
4–9 August 2008

## Workshop Organizers

Marco Baroni, *University of Trento*

Stefan Evert, *University of Osnabrück*

Alessandro Lenci, *University of Pisa*

## Programme Committee

Reinhard Blutner, *University of Amsterdam*

Gemma Boleda, *UPF, Barcelona*

Peter Bosch, *University of Osnabrück*

Paul Buitelaar, *DFKI, Saarbrücken*

John Bullinaria, *University of Birmingham*

Katrin Erk, *UT, Austin*

Patrick Hanks, *Masaryk University, Brno*

Anna Korhonen, *Cambridge University*

Michiel van Lambalgen, *University of Amsterdam*

Claudia Maienborn, *University of Tübingen*

Simonetta Montemagni, *ILC-CNR, Pisa*

Rainer Osswald, *University of Hagen*

Manfred Pinkal, *University of Saarland*

Massimo Poesio, *University of Trento*

Reinhard Rapp, *University of Mainz*

Magnus Sahlgren, *SICS, Kista*

Sabine Schulte im Walde, *University of Stuttgart*

Manfred Stede, *University of Potsdam*

Suzanne Stevenson, *University of Toronto*

Peter Turney, *NRC Canada, Ottawa*

Tim Van de Cruys, *University of Groningen*

Gabriella Vigliocco, *University College, London*

Chris Westbury, *University of Alberta*

# Contents

# Preface

Corpus-based distributional models (such as LSA or HAL) have been claimed to capture interesting aspects of word meaning and provide an explanation for the rapid acquisition of semantic knowledge by human language learners. Although these models have been proposed as plausible simulations of human semantic space organization, careful and extensive empirical tests of such claims are still lacking.

Systematic evaluations typically focus on large-scale quantitative tasks, often more oriented towards engineering applications than towards the challenges posed by linguistic theory, philosophy and cognitive science. Moreover, whereas human lexical semantic competence is obviously multi-faceted – ranging from free association to taxonomic judgments to relational effects – tests of distributional models tend to focus on a single aspect (most typically the detection of semantic similarity), and few if any models have been tuned to tackle different facets of semantics in an integrated manner.

The goal of this workshop was to fill such gaps by inviting researchers to test their computational models on a variety of small tasks that were carefully designed to bring out linguistically and cognitively interesting aspects of semantics. To this effect, annotated data sets were provided for participants on the workshop wiki, where they remain available to interested parties:

`http://wordspace.collocations.de/doku.php/esslli:start`

The proposed tasks were:

- *semantic categorization* – distinguishing natural kinds of concrete nouns, distinguishing between concrete and abstract nouns, verb categorization;
- *free association* – predicting human word association behaviour;
- *salient property generation* – predicting the most salient properties of concepts produced by humans.

The focus of these "shared tasks" was not on competition, but on understanding how different models highlight different semantic aspects, how far we are from an integrated model, and which aspects of semantics are beyond the reach of purely distributional approaches. Most papers in the proceedings report experiments with the proposed data sets, whereas some of the authors explore related tasks and issues.

We hope that this initiative – and the ESSLLI workshop – will foster collaboration among the nascent community of researchers interested in computational semantics from a theoretical and interdisciplinary rather than purely engineering-oriented point of view.

We would like to thank all the authors who submitted papers, as well as the members of the programme committee for the time and effort they contributed in reviewing the papers.

Marco Baroni, Stefan Evert, Alessandro Lenci

# Semantic Categorization Using Simple Word Co-occurrence Statistics

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

## Abstract

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

## 1 Introduction

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

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

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

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

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

## 2 Methodology

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

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

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

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

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

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

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

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

## 3 Results

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

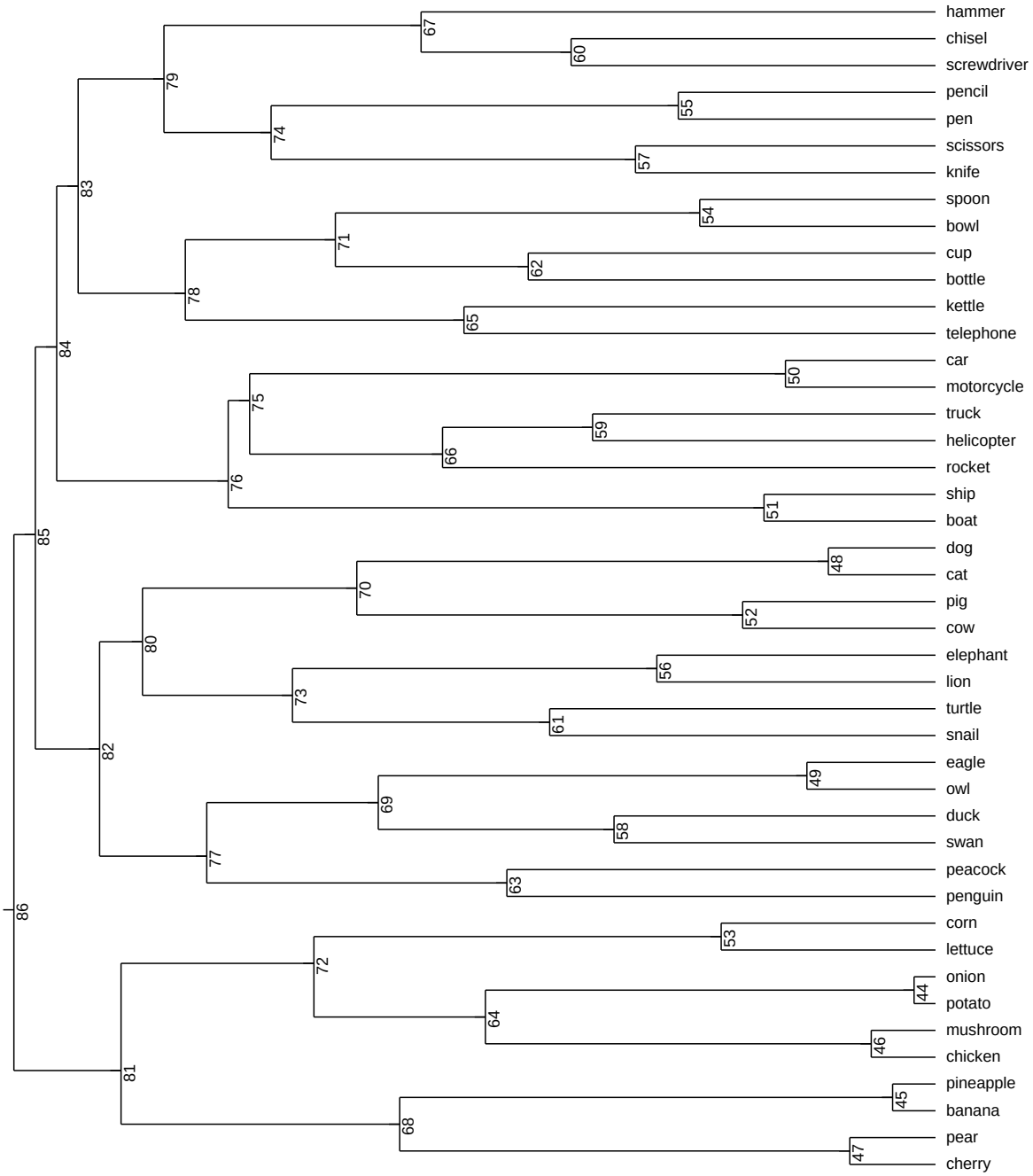



Figure 1: Noun categorization cluster diagram.

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

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

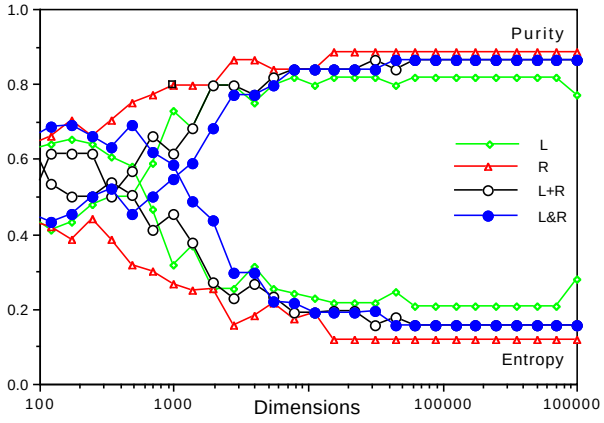



Figure 2: The effect of vector dimensionality on noun clustering quality.

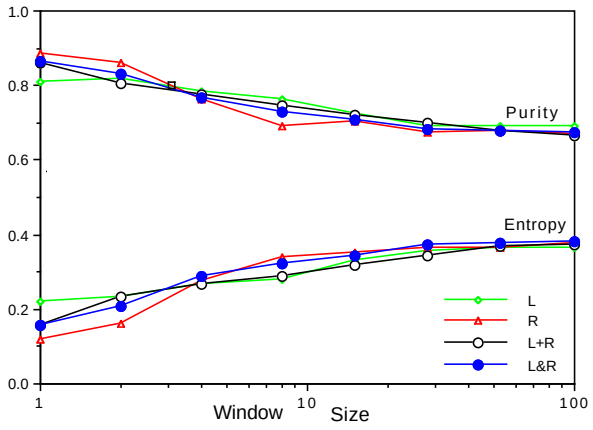



Figure 3: The effect of window size on noun clustering quality.

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

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

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

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

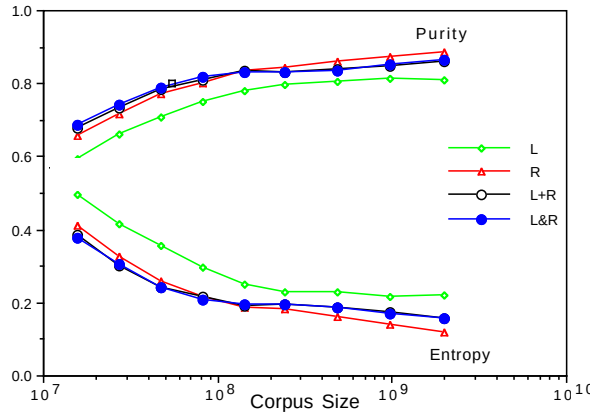



Figure 4: The effect of corpus size on noun clustering quality.

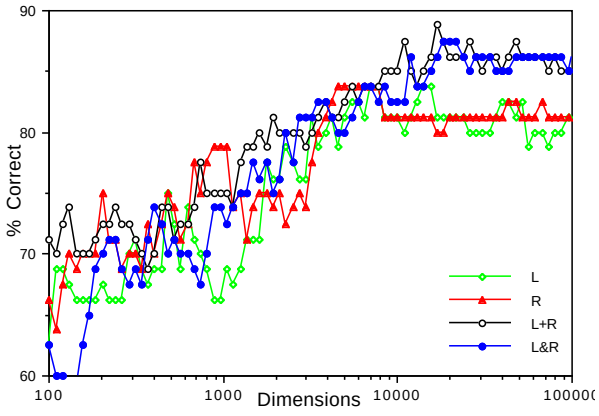



Figure 6: The effect of vector dimensionality on TOEFL performance.

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

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

## 4 Conclusions and Discussion

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

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

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

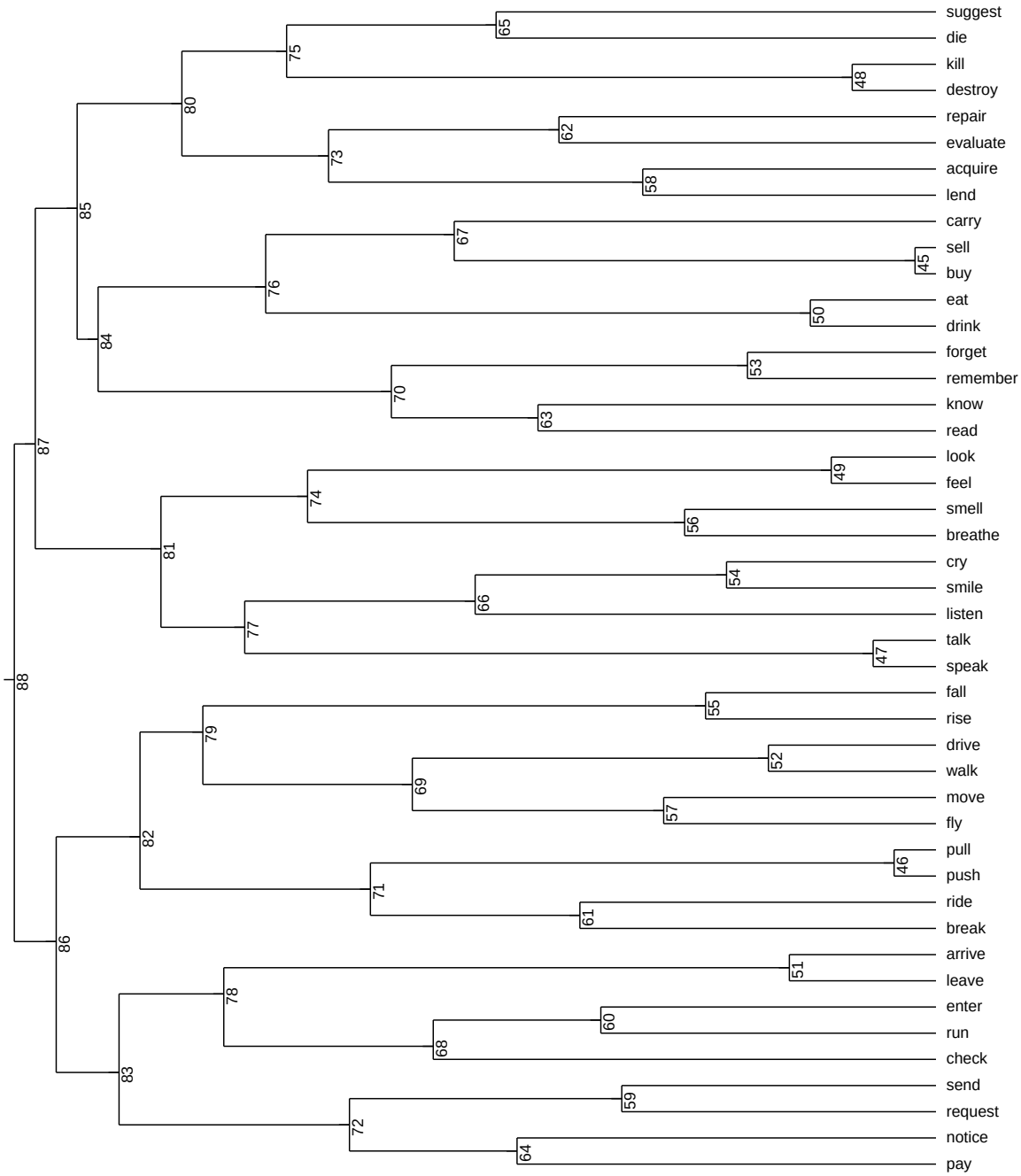



Figure 5: Verb categorization cluster diagram.

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

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

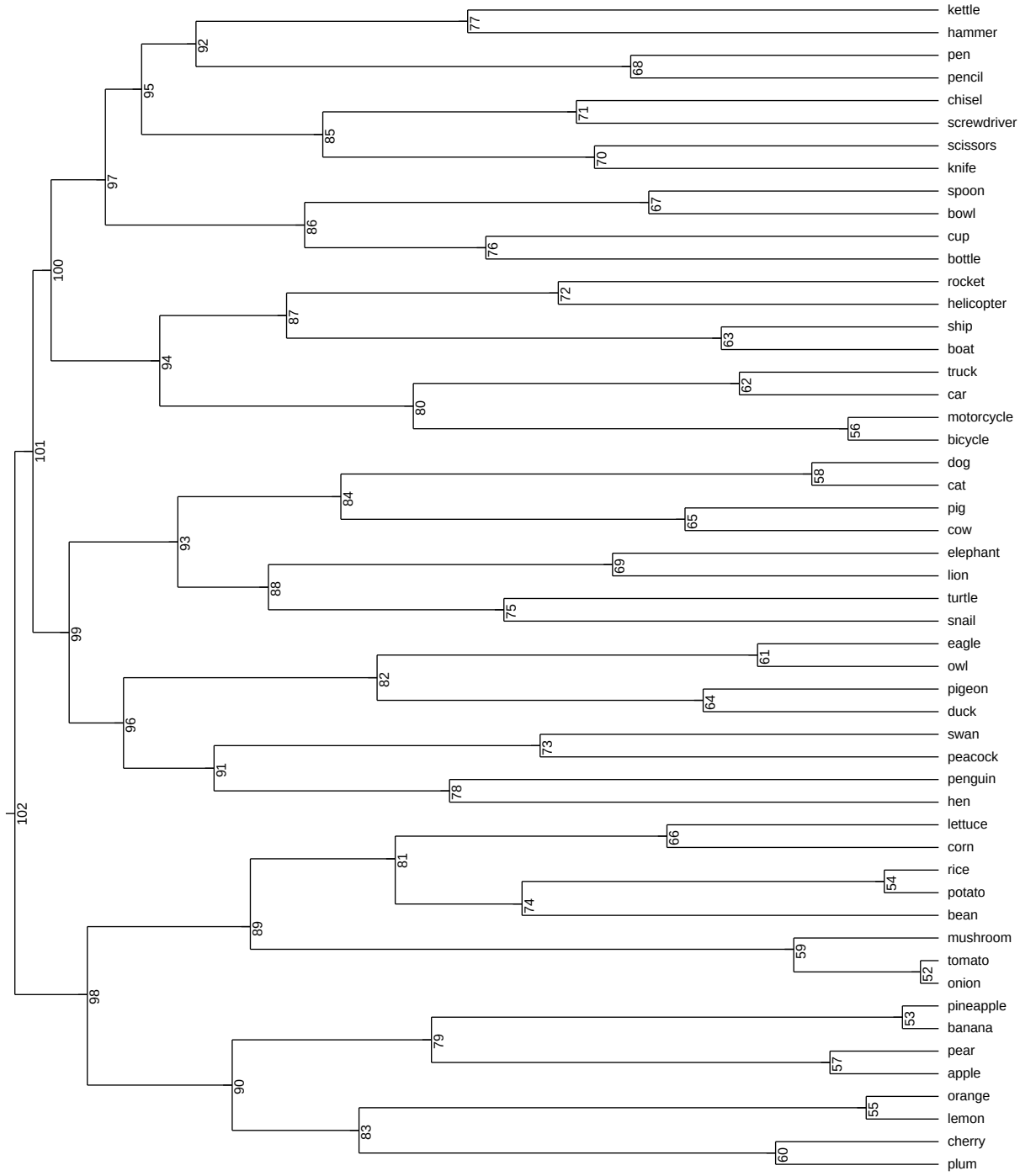



Figure 7: Extended noun categorization cluster diagram.

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

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

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

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

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

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

## References


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

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

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

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

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

Landauer, T.K. & Dumais, S.T. 1997. A Solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, **104**, 211-240.

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

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

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

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

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

# Combining methods to learn feature-norm-like concept descriptions

**Eduard Barbu**
Center for Mind/Brain Sciences
University of Trento
Rovereto, Italy
eduard.barbu@unitn.it

**Abstract**

Feature norms can be regarded as repositories of common sense knowledge. We acquire from very large corpora feature-norm-like concept descriptions using shallow methods. To accomplish this we classify the properties in the norms in a number of property classes. Then we use a combination of a weakly supervised method and an unsupervised method to learn each semantic property class. We report the success of our methods in identifying the specific properties listed in the feature norms as well as the success of the methods in acquiring the classes of properties present in the norms.

## 1 Introduction

In the NLP and Semantic Web communities there is a widespread interest for ontology learning. To build an ontology one needs to identify the main concepts and relations of a domain of interest. It is easier to identify the relevant relations for specialized domains like physics or marketing than it is for general domains like the domain of the common sense knowledge. To formalize a narrow domain we use comprehensive theories describing the respective domain: theories of physics, theories of marketing, etc. Unfortunately we do not have broad theories of the common sense knowledge and therefore we do not have a principled way to identify the properties of "every day" concepts.

In cognitive psychology there is a significant effort to understand the content of mental representation of concepts. A question asked in this discipline is: Which are, from a cognitive point of view, the most important properties of basic level concepts?

An answer to this question is given by feature norms. In a task called feature generation human subjects list what they believe the most important properties for a set of test concepts are. The experimenter processes the resulting conceptual descriptions and registers the final representation in the norm. Thus, a feature norm is a database containing a set of concepts and their most salient features (properties). Usually the properties listed in the norms are pieces of common sense knowledge. For example, in a norm one find statements like:

(1) An *apple* (concept) is a *fruit* (property).

(2) An *airplane* (concept) is used for people *transportation* (property).

In this paper we explore the possibility to learn feature-norm-like concept descriptions from corpora using minimally supervised methods. To achieve this we use a double classification of the properties in the norms. At the morphological level the properties are grouped according to the part of speech of the words used to express them (noun properties, adjective properties, verb properties). At the semantic level we group the properties in semantic classes (taxonomic properties, part properties, etc.).

The properties in certain semantic classes are learnt using a pattern-based approach, while other classes of properties are learnt using a novel method based on co-occurrence associations.

The main contribution of this paper is the devising of a method for learning feature-norm-like

conceptual structures from corpora. Another contribution is the benchmarking of four association measures at the task of finding good lexico-syntactic patterns for a group of four semantic relations.

The rest of the paper has the following organization. The second section briefly surveys other works that make use of shallow methods for relation extraction. The third section discusses the classification of properties in the feature norm we use for the experiments. The fourth section presents the procedure for property learning. In the fifth section we evaluate the accuracy of the procedure and discuss the results. We end the paper with the conclusions.

## 2 Related work

The idea of finding lexico-syntactic patterns expressing with high precision semantic relations was first proposed by Hearst (1992). For identifying the most accurate lexico-syntactic patterns she defined a bootstrapping procedure[1]. The procedure iterates between three phases called: pattern induction, pattern ranking and selection, and instance extraction.

**Pattern Induction**. In the pattern-induction phase one chooses a relation of interest (for example hyperonymy) and collects a list of instances of the relation. Subsequently all contexts containing these instances are gathered and their commonalities identified. These commonalities form the list of potential patterns.

**Pattern Ranking and Selection**. In this stage the most salient patterns expressing the semantic relation are identified. Hearst discovered the best patterns manually inspecting the list of the potential patterns.

**Instance Extraction**. Using the best patterns one gathers new instances of the semantic relation. The algorithm continues from the first step and it finishes either when no more patterns can be found or the number of found instances is sufficient.

The subsequent research tries to automate the most part of Hearst's framework. The strategy followed was to make some of the notions Hearst employed more precise and thus suitable for implementation.

The first clarification has to do with the meaning of the term *commonality*. Ravichandran and Hovy (2002) defined the commonality as being the maximum common substring that links the seeds in k distinct contexts (sentences).

The second improvement is the finding of a better procedure for pattern selection. For example, Ravichandran and Hovy (2002) rank the potential patterns according to their frequency and selects only the n most frequent patterns as candidate patterns. Afterwards they compute the precision of these patterns using the Web as a corpus and retain only the patterns that have the precision above a certain threshold.

Pantel and Pennachiotti (2006) innovated on the work of Ravichandran and Hovy proposing a new pattern ranking and instance selection method. A variant of their algorithm uses the Web for filtering incorrect instances and in this way they exploit generic patterns (those patterns with high recall but low precision).

The pattern-based learning of semantic relations was used in question answering (Ravichandran and Hovy, 2002), identification of the attributes of concepts (Poesio and Abdulrahman, 2005) or for acquiring qualia structures (Cimiano and Wenderoth, 2005).

## 3 Property classification

For our experiments we choose the feature norm obtained by McRae and colleagues (McRae et al., 2005). The norm lists conceptual descriptions for 541 basic level concepts representing living and non-living things. To produce this norm McRae and colleagues interviewed 725 participants.

We classify each property in the norm at two levels: a morphological level and a semantic level.

The morphological level contains the part of speech of the word representing the property. The semantic classification is inspired by a perceptually based taxonomy discussed later in this section. Table 1 shows a part of the conceptual description for the focal concept *axe* and the double classification of the concept properties.

A focal concept is a concept for which the human subjects should list properties in the feature production task.

[1] The terminology for labeling Hearst's procedure was introduced by Pantel and Pennacchiotti (2006).

| Property | Morphological Classification | Semantic classification |
|---|---|---|
| Tool | Noun | superordinate |
| Blade | Noun | part |
| Chop | Verb | action |

Table 1. The double classification of the properties of the concept *axe*

The semantic classification is based on Wu and Barsalou (WB) taxonomy (Wu and Barsalou, in press). This taxonomy gives a perceptually oriented classification of properties in the norms. WB taxonomy classifies the properties in 27 distinct classes. Some of these classes contain very few properties and therefore are of marginal interest. For example, the Affect Emotion class classifies only 11 properties. Our classification considers only the classes that classify at least 100 properties.

Unfortunately, we cannot directly use the WB taxonomy in the learning process because some of the distinctions it makes are too fine-grained. For example, the taxonomy distinguishes between external components of an object and its internal components. On this account the heart of an animal is an internal component whereas its legs are external components. Keeping these distinctions otherwise relevant from a psychological point of view will hinder the learning of feature norm concept descriptions[2]. Therefore we remap the WB initial property classes on a new set of property classes more adequate for our task. Table 2 presents the new set of property classes together with the morphological classification of the properties in each class.

| Semantic classification | Morphological classification |
|---|---|
| Superordinate | noun |
| Part | noun |
| Stuff | noun |
| Location | noun |
| Action | verb |
| Quality | adjective |

Table 2. The semantic and morphological classification of properties in McRae feature norm

[2]We mean learning using the methods introduced in this paper. It is possible that other learning approaches should be able to exploit the WB taxonomy successfully.

The meaning of each semantic class of properties is the following:

- *Superordinate*. The superordinate properties are those properties that classify a concept from a taxonomic point of view. For example, the dog (focal concept) is an animal (taxonomic property).
- *Part*. The category part includes the properties denoting external and internal components of an object. For example blade (part property) is a part of an axe (focal concept).
- *Stuff*. The properties in this semantic class denote the stuff an object is made of. For example, bottle (focal concept) is made of glass (stuff property).
- *Location*. The properties in this semantic class denote typical places where instances of the focal concepts are found. For example, airplanes (focal concept) are found in airports (location property).
- *Action*. This class of properties denotes the characteristic actions defining the behavior of an entity (the cat (focal concept) meow (action property)) or the function, instances of the focal concepts typically fulfill (the heart (focal concept) pumps blood (function property)).
- *Quality*. This class of properties denotes the qualities (color, taste, etc.) of the objects instances of the focal concepts. For example, the apple (focal concept) is red (quality property) or the apple is sweet (quality property).

The most relevant properties produced by the subjects in the feature production experiments are in the categories presented above. Thus, asked to list the defining properties of the concepts representing concrete objects the subjects will typically: classify the objects (Superordinate), list their parts and the stuff they are made from (Parts and Stuff), specify the location the objects are typically found in (Location) their intended functions and their typical behavior (Action) or name their perceptual qualities (Quality).

## 4 Property learning

To learn the property classes discussed in the preceding section we employ two different strategies. Superordinate, Part, Stuff and Location properties are learnt using a pattern-based approach. Quality and Action properties are learnt using a novel method that quantifies the strength of association between the nouns representing the focal concepts and the adjective and verbs co-occurring with them in a corpus. The learning decision is motivated by the following experiment. We took a set of concepts and their properties from McRae feature norm and extracted sentences from a corpus where a pair concept - property appears in the same sentence.

We noticed that, in general, the quality properties are expressed by the adjectives modifying the noun representing the focal concept. For example, for the concept property pair (apple, red) we find contexts like: "She took the red apple".

The action properties are expressed by verbs. The pair (dog, bark) is conveyed by contexts like: "The ugly dog is barking" where the verb expresses an action to which the dog (i.e. the noun representing the concept) is a participant.

The experiment suggests that to learn Quality and Action properties we should filter the adjectives and verbs co-occurring with the focal concepts.

For the rest of the property classes the extracted contexts suggest that the best learning strategy should be a pattern-based approach. Moreover with the exception of the Location relation, that, to our knowledge, has not been studied yet, for the relations Superordinate, Part and Stuff some patterns are already known.

The properties we try to find lexico-syntactic patterns for are classified at the morphological level as nouns (see Table 2). The rest of the properties are classified as either adjectives (Qualities) or verbs (Action).

To identify the best lexico-syntactic patterns we follow the framework introduced in section 2. The hypothesis we pursue is that the best lexico-syntactic patterns are those highly associated with the instances representing the relation of interest. The idea is not new and was used in the past by other researchers. However, they used only frequency (Ravichandran and Hovy, 2002) or pointwise mutual information (Pantel and Penacchiotti, 2006) to calculate the strength of association between patterns and instances. We improve previous work and employ two statistical association measures (Chi-squared and Log-Likelihood) for the same task. Further we benchmark all four-association measures (the two used in the past and the two tested in this paper) at the task of finding good lexico-syntactic patterns for Superordinate, Part, Stuff and Location relations.

The pattern induction phase starts with a set of seeds instantiating one of the four semantic relations. We collect sentences where the seeds appear together and replace every seed occurrence with their part of speech. The potential patterns are computed as suggested by Ravichandran and Hovy (see section 2) and the most general ones are eliminated from the list.

The remaining patterns are ranked using each of the above mentioned association measures.

We introduce the following notation:

- $I = \{i_1, i_2 ... i_m\}$. $I$ is the set of instances in the training set.
- $P = \{p_1, p_2 ... p_k\}$. $P$ is the set of patterns linking the seeds in the training set and inferred in the pattern induction phase.
- $S = \{\{i_1, p_1\}, \{i_1, p_2\}, ..., \{i_s, p_k\}\}$.) $p_i \in P$. $S$ is the set of all instance-pattern pairs in the corpus.

If we consider an instance $i$ and a pattern $p$, then, following (Evert, in press), we define:

- $O_{11}$ the number of occurrences the instance has with the pattern $p$.
- $O_{12}$ the number of occurrences the instance has with any other pattern except $p$.
- $O_{21}$ the number of occurrences any other instance except $i$ has with the pattern p.
- $O_{22}$ the number of occurrences any instance except $i$ has with any pattern excepts $p$..
- $R_1$ and $R_2$ are the sums of the table rows
- $C_1$ and $C_2$ are the sums of the table columns. All defined frequencies can be easily visualized in table 3.

- $N$ is the number of all instances with all the patterns (the cardinality of S).

| | $p$ | $\neg p$ | Row sum |
|---|---|---|---|
| $i$ | $O_{11}$ | $O_{12}$ | $R_1 = O_{11} + O_{12}$ |
| $\sim i$ | $O_{21}$ | $O_{22}$ | $R_2 = O_{21} + O_{22}$ |
| Column Sum | $C_1 = O_{11} + O_{21}$ | $C_2 = O_{12} + O_{22}$ | $N = R_1 + R_2$ |

Table 3. The contingency table

The tested association measures are:

**Simple Frequency**

The frequency $O_{11}$ gives the number of occurrences of a pattern with an instance.

(**Pointwise) mutual information** (Church and Hanks, 1990)

Because this measure is biased toward infrequent events, in practice a correction is used to counterbalance the bias effect.

$$MI^2 = \log_2 \frac{O_{11}^2}{\frac{R_1 \cdot C_1}{N}}$$

**Chi-squared (with Yates continuity correction) (**DeGroot and Schervish, 2002**)**

$$chi_{corr} = \frac{N(|O_{11} \cdot O_{22} - O_{12} \cdot O_{21}| - \frac{N}{2})^2}{R_1 \cdot R_2 \cdot C_1 \cdot C_2}$$

**Log-Likelihood**(Dunning ,1993) :

$$\log - likelihood = 2 \cdot \sum_{ij} \log \frac{O_{ij}}{\frac{R_i \cdot C_j}{N}}$$

Once the strength of association between the instances and patterns in $S$ is quantified, the best patterns are voted. The best patterns are the patterns having the higher association score with the instances in the training set. Therefore, for each pattern in P we compute the sum of the association scores of the pattern with all instances in the set $I$. In the pattern selection phase we manually evaluate the two best patterns selected using each association measure. In case the patterns have a good precision we used them for new property extraction otherwise, we use the intuition to devise new patterns. The precision of a pattern used to represent a certain semantic relation is evaluated in the following way. A set of 50 concept-feature pairs is selected from a corpus using the devised pattern. For example, to evaluate the precision of the pattern: "N made of N" for the Stuff relation we extract concept feature pairs like *hammer-wood*, *bottle-glass*, *car-cheese*, etc.. Then we label a pair as a hit if the semantic relation holds between the concept and the feature in the pair and a miss otherwise. The pattern precision is defined as the percent of hits. In the case of the three pairs in the example above we have two hits: *hammer-wood* and *bottle-glass* and one miss: *car-cheese*. Thus we have a pattern precision of 66 %.

The Quality and Action properties are learnt using an unsupervised approach. First the association strength between the nouns representing the focal concepts and the adjectives or verbs co-occurring with them in a corpus is computed. The co-occurring adjectives are those adjectives found one word at the left of the nouns representing the focal concepts. A co-occurring verb is a verb found one word at the right of the nouns representing the focal concepts or a verb separated from an auxiliary verb by the nouns representing the focal concepts.

The strongest 30 associated adjectives are selected as Quality properties and the strongest 30 associated verbs are selected as Action properties.

To quantify the new attraction strength between the concept and the potential properties of type adjective or verb the same association measures introduced before are used. The association measures are then benchmarked at the task of finding relevant properties for the focal concepts.

## 5 Results and discussion

The corpora used for learning feature-norm-like concept descriptions are British National Corpus (BNC) and ukWaC (Ferraresi et al., in press). The BNC is a balanced corpus containing 100 million words. UkWaC is a very large corpus of British English, containing more than 2 billion words, constructed by crawling the web. For evaluating the success of our method we have chosen a test set of 44[3] concepts from McRae feature norm. In the next two subsections we report and discuss the results obtained for Superordinate, Stuff, Location

[3] The test set is the same set of concepts used in the workshop task "generation of salient properties of concepts".

and Part properties and Quality and Action properties respectively. All our experiments were performed using the CWB (Christ, 1994) and UCS toolkits (http://www.collocations.de/software.html).

### 5.1 Results for Superordinate, Stuff, Location and Part properties

For the concepts in the test set we extract properties using the manual selected patterns reported in table 4.

We evaluate the success of each association measure in finding good patterns and the success of manually selected patterns in extracting good properties.

The input of the algorithm for automatic pattern selection consists of 200 seeds taken from the McRae database. None of the 44 test concepts nor their properties is among the input seeds. The pattern-learning algorithm is run on BNC using each association measure introduced in section 4. Therefore for each relation in the table 4 we have four runs of the algorithm, one for each association measure. We evaluate the precision of the top two voted patterns.

| Relation | Pattern |
|---|---|
| Superordinate | N [JJ]-such [IN]-as N;<br>N [CC]-and [JJ]-other N;<br>N [CC]-or [JJ]-other N; |
| Stuff | N [VVN]-make [IN]-of N |
| Location | N [IN]-from [DT]-the N |
| Part | N [VVP]-comprise N<br>N [VVP]-consists [IN]-of N |

Table 4. The manually selected patterns

The manually selected patterns for *Superordinate* relation are voted by any of the tested association measures. Therefore, to find patterns for the Superordinate relation one needs to supply the algorithm presented in section 4 with a set of seeds and the top patterns voted by any of the four association measures will be good lexico-syntactic patterns.

The pattern-learning algorithm run with any association measure except the simple frequency will rank higher the pattern manually selected to represent the *Stuff* relation. The simple frequency votes the following patterns as the strongest associated patterns with the instances in the test set: *N from the N* and *N be in N*. The first pattern does not express the Stuff relation whereas the second one expresses it very rarely.

In the case of *Location* relation all association measures select the pattern in the table except Chi-squared. The top two patterns (*N cold N* and *N freshly ground black N*) selected with the aid of the Chi-squared measure are very rare constructions that appear with the input instances.

The manually selected patterns for Part are not found by any association measure. Only one of the patterns voted by frequency and log-likelihood (*N have N*) sometimes expresses the Part relation, the rest of patterns voted are spurious constructions appearing with the instances in the input set.

Therefore the contest of association measures for a good pattern selection marginally favors pointwise mutual information with correction and log-likelihood.

Using the manually selected patterns presented in the above table we gather new properties for the concepts in the test set from UkWaC corpus.

The results of property extraction phase are reported in table 5. The columns of the table represent in order: the name of the class of semantic properties to be extracted, the recall of our procedure and the pattern precision. The recall tells how many properties in the test set are found using the patterns in table 4. The pattern precision states how precise the selected pattern is in finding the properties in a certain semantic class and it is computed as shown in section 4. In case more than one pattern have been selected, the pattern precision is the average precision for all selected patterns.

| Property class | Recall | Pattern Precision |
|---|---|---|
| Superordinate | 87% | 85 % |
| Stuff | 21% | 70 % |
| Location | 33% | 40 % |
| Part | 0 % | 51 % |

Table 5. The results for each property class

As one can see from table 5, the recall for the superordinate relation is very good and the precision of the patterns is not bad either (average precision 85 %). However, some of the extracted superordinate

properties are roles and not types. For example, banana, one of the concepts in the test set, has the superordinate property fruit (type). Using the patterns for superordinate relation we find that banana is a fruit (type) but also an ingredient and a product (roles). The lexico-syntactic patterns for the superordinate relation blur the type-role distinction.

The pattern used to represent the Stuff relation has a bad recall (21 %) and an estimated precision of 70 %. To be fair, the pattern expresses better than the estimated precision the substance an object is made of. The problem is that in many cases constructions of type "Noun made of Noun" are used in a metaphoric way as in: "car made of cheese". In the actual context the car was not made of cheese but the construction is used to show that the respective car was not resistant to impact.

The pattern for Location relation has bad precision and bad recall. The properties of type Location listed in the norm represent typical places where objects can be found. For example, in the norm it is stated that bananas are found in tropical climates (the tropical climate being the typical place where bananas grow). However what one can hope from a pattern-based approach is to find patterns representing with good precision the concept of Location in general. We founded a more precise Location pattern than the selected one: N is found in N. Unfortunately, this pattern has 0% recall for our test set.

The patterns for Part relation have 0% recall for the concepts in the test set and their precision for the general domain is not very good either. As others have shown (Girju et al. 2006) a pattern based approach is not enough to learn the part relation and one needs to use a supervised approach to achieve a relevant degree of success.

### 5.2 Results for Quality and Action properties

We computed the association strength between the concepts in the test set and the co-occurring verbs and adjectives using all four-association measures. The best recall for the test set was obtained by log-likelihood measure and the results are reported for this measure.

The results for Quality and Action properties are presented in table 6. The columns of the table represent in order: the name of the class of semantic properties, the Recall and the Property Precision. The Recall represents the percent of properties in the test set our procedure found. The Property Precision computes the precision with which our procedure finds properties in a semantic class. The property precision is the percent of quality and action properties found among the strongest 30 adjectives and verbs associated with the focal concepts.

| Property class | Recall | Property Precision |
|---|---|---|
| Quality | 60% | 60 % |
| Action | 70% | 83 % |

Table 6. The results for Quality and Action property classes

Because the number of potential properties is reasonable for hand checking, the validation for this procedure was performed manually.

The manual comparison between the corpus extracted properties and the norm properties confirm the hypothesis regarding the relation between the association strength of features of type adjective and verbs and their degree of relevance as properties of concepts.

For each concept in the test set roughly 18 adjectives and 25 verbs in the extracted set of potential properties represent qualities and action respectively (see Property Precision column in table 6). This can be explained by the fact that all concepts in the test set denote concrete objects. Many of the adjectives modifying nouns denoting concrete objects express the objects qualities, whereas the verbs usually denote actions different actors perform or to which various objects are subject.

There are cases in which the properties found using this method are excellent candidates for the semantic representation of focal concepts. For example, the semantic representation of the concept *turtle* has the following Quality properties listed in the norm {green, hard, small}. The strongest adjectives associated in the UkWaC corpus with the noun turtle ordered by the loglikelihood score are: {marine, green, giant}. The property *marine* carries a greater distinctiveness than any of similar feature listed in the norms.

The actions typically associated with the concept *turtle* in the McRae feature norm are {lays eggs, swims, walks slowly}. The strongest verbs associated in the UkWaC corpus with the noun turtle are: {dive, nest, hatch}. The *dive* action is

more specific and therefore more distinct than the *swim* action registered in the feature norm. The *hatch* property is characteristic to reptiles and birds and thus a good candidate for the representation of the concept *turtle*.

## 6 Conclusions

The presented method for learning feature norm concept description has been successful at learning the semantic property classes Superordinate, Quality and Action. All these properties can be learnt automatically. For Superordinate relation one starts with a set of seeds representing the Superordinate relation and then, as shown in section 4, computes the best pattern associated with the seeds using any of the discussed measures. Then (s)he extracts new properties for a test set of concepts using the voted pattern. For Quality and Action properties one needs to apply the method based on concurrence association presented in the same section 4.

To learn all the other property classes to other methods (probably a supervised approach) must be devised.

As in the case of ontology learning or qualia structure acquisition it seems that the best way to acquire feature-norm-like concept descriptions is a semiautomatic one. A human judge makes the best property selections based on the proposals made by an automatic method.

## Acknowledgments

I like to thank Stefan Evert for the discussion on association measures and to Verginica Barbu Mititelu and two anonymous reviewers for their suggestions and corrections.

## 7 References

Adriano Ferraresi, Eros Zanchetta, Marco Baroni and Silvia Bernardini. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008* (to appear).

Christ Oli. 1994. A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94*, Budapest.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. *Proceedings of ACL-2002*: 41-47.

Ken McRae, George S. Cree, Mark S. Seidenberg, Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37: 547-559.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22–29.

Ling-Ling Wu, Lawrence W. Barsalou. *Grounding Concepts in Perceptual Simulation: Evidence from Property Generation*. In press.

Marti Hearst 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING-92*, 539-545.

Massimo Poesio and Abdulrahman Almuhareb. 2005. Identifying Concept Attributes Using a Classifier. *Proceedings of ACL Workshop on Deep Lexical Acquisition*.

Morris H. DeGroot and Mark J. Schervish. 2002. *Probability and Statistics*. Addison Wesley, Boston, 3rd edition.

Patrick Pantel, Marco Pennacchiotti. 2006. Espresso: A Bootstrapping Algorithm for Automatically Harvesting Semantic Relations. *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*.

Philipp Cimiano and Johanna Wenderoth. 2005. Automatically Learning Qualia Structures from the Web. *Proceedings of the ACL Workshop on Deep Lexical Acquisition*, 28-37.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2006. Automatic Discovery of Part-Whole Relations. *Computational Linguistics*, 32(1): 83-135.

Stefan Evert. Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin. In press.

Ted E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61–74.

# Qualia Structures and their Impact on the Concrete Noun Categorization Task

**Sophia Katrenko**
Informatics Institute
University of Amsterdam
the Netherlands
katrenko@science.uva.nl

**Pieter Adriaans**
Informatics Institute
University of Amsterdam
the Netherlands
pietera@science.uva.nl

## Abstract

Automatic acquisition of qualia structures is one of the directions in information extraction that has received a great attention lately. We consider such information as a possible input for the word-space models and investigate its impact on the categorization task. We show that the results of the categorization are mostly influenced by the formal role while the other roles have not contributed discriminative features for this task. The best results on 3-way clustering are achieved by using the formal role alone (entropy 0.00, purity 1.00), the best performance on 6-way clustering is yielded by a combination of the formal and the agentive roles (entropy 0.09, purity 0.91).

## 1 Introduction

Computational models of semantic similarity have been used for some decades already with various modifications (Sahlgren, 2006). In this paper, we investigate qualia structures and their impact on the quality of the word-space models. Automatic acquisition of qualia structures has received a great attention lately resulting in several methods which use either existing corpora or the Web (Cimiano and Wenderoth, 2007; Yamada et al., 2007). We build on the work reported in the literature and aim to test how suitable the results of automatic qualia extraction are for the word-space models. We approach a seemingly simple task of the concrete noun categorization. Previous research has shown that when humans are asked to provide qualia elements per role for a list of nouns, concrete nouns lead to the highest agreement. The words with the lowest agreement are abstract notions (Cimiano and Wenderoth, 2007). Naturally, a question arises of what information would be captured by the word-space models if qualia elements are used.

This paper is organized as follows. Section II presents some relevant information on word-space models and their modifications. Section III gives a brief overview of the Generative Lexicon Theory. Then, we describe a method used for an automatic qualia structure acquisition. We proceed with an experimental part by discussing results and analyzing errors.

## 2 Word-Space Models

Underlying idea behind the word-space models lies in the semantic similarity of words. In particular, if two words are similar, they have to be close in the word space which led to so called geometric metaphor. In his dissertation, Sahlgren (2006) discusses different ways of constructing such word spaces. One possible solution is to take into account word co-occurrences, the other would be using a limited number of semantic features. While the former method may result in a high-dimensional space containing redundant information, the latter may be too restrictive. The main concern about a list of features is how they can be defined and what kind of features are sufficient for a given task. Sahlgren (2006) argues that the word-space models have to be considered together with a task they are used for. He highlights differences between the word-space models based on paradigmatic and syntagmatic notions and shows that both models can be effectively

used. On the task of human association norm, word spaces produced by using syntagmatic information seem to have a higher degree of corelation with a norm, while paradigmatic word-spaces yield better results on the synonymy test.

## 3 Generative Lexicon Theory

In the semantic theory of Generative Lexicon, Pustejovsky (2001) proposes to describe lexical expressions by using four representation levels, argument structure, event structure, qualia structure, and lexical inheritance structure. For the work presented here, qualia structure is of the most interest. Qualia structure use defined by the following roles:

- *formal* - information that allows to distinguish a given objects from others, such as superclass
- *constitutive* - an object's parts
- *telic* - a purpose of an object; what it is used for
- *agentive* - origin of an object, "how it came into being"

While discussing natural kinds and artifacts, Pustejovsky (2001) argues that a distinction between these two categories can be drawn by employing a notion of intentionality. In other words, it should be reflected in the telic and agentive roles. If no intentionality is involved, such words are natural types. On the contrary, artifacts are identified by the telic and agentive roles.

## 4 Automatic Acquisition of Qualia Structures

After the theory of Generative Lexicon has been proposed, various researchers put it in practice. For instance, Lenci (2000) considered it for designing ontologies on example of SIMPLE. Qualia structure is used here to formally represent a core of the lexicons. In a nutshell, the SIMPLE model is more complex and besides qualia structure includes such information as argument structure for semantic units, selectional restrictions of the arguments, collocations and other. The Generative Lexicon theory was also used for different languages. For instance, Zavaglia and Greghi (2003) employ it to analyze homonyms in Portuguese.

Another interesting and useful aspect of qualia structure acquisition is automatic qualia extraction. Recently, Yamada et al. (2007) presented a method on the telic role acquisition from corpus data. A motivation behind the telic role was that there are already approaches to capture formal or constitutive information, while there is less attention to the function extraction. A method of Yamada et al. (2007) is fully supervised and requires a human effort to annotate the data.

Contrary to the work reported in (Yamada et al., 2007), Cimiano and Wenderoth (2007) proposed several hand-written patterns to extract qualia information. Such patterns were constructed in the iterative process and only the best were retained. Further, the authors used various ranking measures to filter out the extracted terms.

We start with the qualia information acquisition by adopting Cimiano and Wenderoth's (2007) approach. For each role, there is a number of patterns which might be used to obtain qualia information. Table 1 contains a list of the patterns per role which have been proposed by Cimiano and Wenderoth (2007). All patterns are accompanied by the parts of speech tags.[1] The patterns for the *formal* role are well-known Hearst patterns (Hearst, 1992) and patterns for the other roles were acquired manually.

In Table 1 $x$ stands for a seed in singular (e.g., *lion*, *hammer*) and $p$ for a noun in plural (e.g., *lions*, *hammers*). For pluralia tantum nouns only a corresponding subset of patterns is used. In addition, we employ a wildcard which stands for one word (a verb, as it can be seen in agentive patterns).

## 5 Experiments

The data set used in our experiments consists of 44 words which fall in several categories, depending on the granularity. On the most general level, they can be divided in two groups, *natural kind* and *artifact*. Further, *natural* group includes such categories as *vegetable* and *animal*. On the most specific level, the data set represents the following 6 categories: *green*, *fruitTree*, *bird*, *groundAnimal*, *tool*, and *ve-*

[1] the following categories are used : nouns in singular ($NN$), nouns in plural $NNP$, conjunctions ($CC$), determiners ($DET$), adjectives ($JJ$), prepositions ($IN$)

| Role | Pattern |
|---|---|
| formal | x_NN is_VBZ (a_DT\|the_DT) kind_NN of_IN<br>x_NN is_VBZ<br>x_NN and_CC other_JJ<br>x_NN or_CC other_JJ<br>such_JJ as_IN p_NNP<br>*,*(* especially_RB p_NNP<br>*,*(* including_VVG p_NNP |
| telic | purpose_NN of_IN (a_DT)* x_NN is_VBZ<br>purpose_NN of_IN p_NNP is_VBZ<br>(a_DT\|the_DT)* x_NN is_VBZ used_VVN to_TO<br>p_NNP are_VBP used_VVN to_TO |
| constitutive | (a_DT\|the_DT)* x_NN is_VBZ made_VVN (up_RP )*of_IN<br>(a_DT\|the_DT)* x_NN comprises_VVZ<br>(a_DT\|the_DT)* x_NN consists_VVZ of_IN<br>p_NNP are_VBP made_VVN (up_RP )*of_IN<br>p_NNP comprise_VVP<br>p_NNP consist_VVP of_IN |
| agentive | to_TO * a_DT new_JJ x_NN<br>to_TO * a_DT complete_JJ x_NN<br>to_TO * new_JJ p_NNP<br>to_TO * complete_JJ p_NNP<br>a_DT new_JJ x_NN has_VHZ been_VBN<br>a_DT complete_JJ x_NN has_VHZ been_VBN |

Table 1: Patterns

*hicle*. Such division of the data set poses an interesting question whether these distinctions can be adequately captured by a method we employ. For extracting hyperonymy relation, there seems to be a consensus that very general information (like *John is a human*) is not likely to be found in the data. It is therefore unclear whether the 2-way clustering (*natural* vs. *artifact*) would provide accurate results. We hypothesize that a more granular distinction can be captured much better.

To conduct all experiments, we use the Web data, particularly, Google API to extract snippets. Similarly to the experiments by Cimiano and Wenderoth(2007), a number of extractions is set to 50. However, if enumerations and conjunctions are treated, a number of extractions per seed might be greater than this threshold. All snippets are tokenized and tagged by a PoS analyzer, which in our case is TreeTagger[2]. Further, the preprocessed data is matched against a given pattern. PoS information allows us to reduce a number of candidates for the qualia roles. Unlike Cimiano, we do not employ any ranking of the extracted elements but use them to build a word-space model. In such a model, rows correspond to the words provided by the organizers of the challenge and columns are the qualia elements for a selected role. As in most word-space models, the elements of a matrix contain frequency counts. CLUTO (Zhao and Karypic, 2002) toolkit is used to cluster the seeds given the information in matrices.

Table 2 presents the results (when *formal* role only is used) in terms of purity and entropy. Entropy of cluster $i$ is usually measured as

$$e_i = -\sum_{j=1}^{m} p_{ij} log(p_{ij}) \quad (1)$$

where $m$ is a number of classes and $p_{ij}$ is probability that an object in cluster $i$ belongs to cluster $j$, $p_{ij} = n_{ij}/n_i$. Purity $r$ is defined as $r = \max_j p_{ij}$. The total entropy (purity) is obtained by a weighted sum of the individual cluster entropies (purities).

[2]available from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

| clustering | entropy | purity |
|---|---|---|
| 2-way | 0.59 | 0.80 |
| 3-way | **0.00** | **1.00** |
| 6-way | 0.13 | 0.89 |
| 2-way$_{>1}$ | 0.70 | 0.77 |
| 3-way$_{>1}$ | 0.14 | 0.96 |
| 6-way$_{>1}$ | 0.23 | 0.82 |

Table 2: Performance using *formal* role only

The 2-way clustering resulted in the imperfect discrimination between natural and artifact categories. Errors are caused by classifying vegetables as artifacts while they belong to the category *natural kind*. The 3-way clustering was intended to exhibit differences among fruit and vegetables (1st category), birds and animals (2nd category) and tools and vehicles (3rd category). In contrast to the 2-way clustering, there have been no errors observed in the clustering solution. While conducting experiments with the 6-way clustering aiming at finding 6 clusters corresponding to the abovementioned categories, we have noticed that vehicles and tools are not properly discriminated.

We have not filtered the acquired formal role elements in any way. As there is noise in the data, we decided to conduct an additional experiment by removing all features with the frequency 1 (2-way$_{>1}$, 3-way$_{>1}$, 6-way$_{>1}$). We observe lower purity for all three categorization tasks which suggest that some of the removed elements were important. In general, seeds in such categories as *bird* or *animal* get many qualia elements with the high frequency for the *formal* role varying from very general such as *creature*, *mammal*, *species* to quite specific (*pheasant*, *vertebrate*). Some members of other categories such as *tool* or *vehicle* do not possess as many features and their frequency is low.

## 6 Discussion

To evaluate which features were important for each particular solution and shed light on problematic areas, we carried out some additional analysis. Table 3, Table 4 and Table 6 present descriptive and discriminative features for the $2-$way, $3-$way and $6-$ way clustering respectively. Descriptive features correspond to the features that describe a given cluster the best and the discriminative are those which highlight the differences between a given cluster and the rest. Each feature is provided with the percentage of its contribution to a given category. In each table $A$ stands for a descriptive part and $B$ denotes discriminative one.

The categories are not necessarily homogeneous and this can be observed given the descriptive features. For instance, the category *vegetables* includes the member *mushroom* the features of which are quite distinct from the features of the other members (such as *potato* or *onion*). This is reflected by the feature *fungi* in the descriptive part of *vegetables*. The category *bird* includes the false descriptive feature *story*. As mentioned, we did not rank extracted terms in any way and such titles as *owls and other stories* heavily contributed to the feature *story*. It turned out that such titles are frequent for the *bird* category and, fortunately, a presence of this feature did not result in any misclassifications.

**Telic role** Having hoped that additional information might be helpful for such categories as *tool* and *vehicle*, we added terms extracted for other qualia roles. Unfortunately, none of them drastically changed the overall performance. The fact that *telic* role does not have a considerable impact on the final performance can be explained if one looks at the extracted terms. As some examples in Table 5 suggest, patterns for the *telic* role provide useful cues to what a purpose of the given entity is. Nevertheless, accurate extractions are usually not shared by other members of the same category. For instance, various tools have a quite different purpose (e.g., *to cut*, *to hit*, *to serve*) and there is no common denominator which would serve as a descriptive feature for the entire category.

**Constitutive role** In contrast to the telic role, constitutive information is either too general or too specific and does not contribute to the overall performance either. For instance, it is known that mushrooms consist of mycelium, water and have a cap; telephones have transceivers and handsets; boats consist of cabins, engines, niches and hulls but this information is not really used to discriminate among categories.

**Agentive role** The last role to consider is an agen-

| | Cluster | Features |
|---|---|---|
| A | NATURAL | animal (43.3%), bird (23.0%), story (6.6%), pet (3.5%), waterfowl (2.4%) |
| | ARTIFACT | tool (19.7%), fruit (14.6%), vegetables (10.0%), vehicle (9.7%), crop (5.1%) |
| B | | animal (22.1%), bird (11.7%), tool (10.1%), fruit (7.4%), vegetables (5.1%) |

Table 3: 2−way clustering: descriptive vs. discriminative features

| | Cluster | Features |
|---|---|---|
| A | VEGETABLE | fruit (41.3%), vegetables (28.3%), crop (14.6%), food (3.4%), plant (2.5%) |
| | ANIMAL | animal (43.3%), bird (23.0%), story (6.6%), pet (3.5%), waterfowl (2.4%) |
| | ARTIFACT | tool (31.0%), vehicle (15.3%), weapon (5.4%), instrument (4.4%), container (3.9%) |
| B | VEGETABLE | fruit (21.0%), vegetables (14.3%), animal (11.6%), crop (7.4%), tool (2.5%) |
| | ANIMAL | animal (22.1%), bird (11.7%), tool (10.1%), fruit (7.4%), vegetables (5.1%) |
| | ARTIFACT | tool (15.8%), animal (14.8%), bird (7.9%), vehicle (7.8%), fruit (6.8%) |

Table 4: 3−way clustering: descriptive vs. discriminative features

| seed | extractions |
|---|---|
| helicopter | to rescue |
| rocket | to propel |
| chisel | to cut, to chop, to clean |
| hammer | to hit |
| kettle | to boil, to prepare |
| bowl | to serve |
| pencil | to draw, to create |
| spoon | to serve |
| bottle | to store, to pack |

Table 5: Some extractions for the *telic* role

tive one. As pointed out by Pustejovsky (2001), it might be helpful to distinguish between natural kinds and artifacts. Indeed, by adding results delivered by agentive patterns to those corresponding to the formal role, entropy decreases to **0.09** and purity increases to **0.91** on 6-way clustering. However, the results do not change for the 2-way clustering still classifying the members of the category *vegetables* as artifacts. An interesting observation is that such division reflects a degree of animacy. It is well known that personal pronouns have the highest animacy followed by humans, animals, plants, concrete things and abstract things (in this order). Clustering based on the formal role alone or any combination of other roles with it always distinguishes well animals (a *groundAnimal* and a *bird*) from plants but it seems to cut the animacy hierarchy right after animals by placing vegetables to artifacts.

A combination of the agentive and the formal roles (Figure 1) finally correctly classifies *rocket* as a vehicle. We can also observe that agentive features (*to develop*, *to invent*, etc.) mostly influence such categories as vehicles and tools.

All features (rows) presented in Figure 1 were selected on the basis of the union of the discriminative and descriptive features per cluster. Actual seeds (words) are given in columns.

### 6.1 Error Analysis

We also analyzed the clusters by means of features that frequently occur together. In CLUTO, there are two possibilities to perform such analysis, either by employing cliques or itemsets. Our aim in this case is to find possible subclusters withing resulting clusters. If there are any subclusters, they can be either attributed to the existing subcategories of a given category or by clustering errors so that two subclusters were merged even though they do not belong to the same category. Such information would be especially useful for the *tool* and *vehicle* classes. A cluster containing a mixture of vehicles and tools is indeed described by three frequent itemsets, *container-load*, *container-appliances* and *vehicle-aircraft*. Consequently, such words as *bowl*, *bottle* and *kettle* are all correctly clustered together

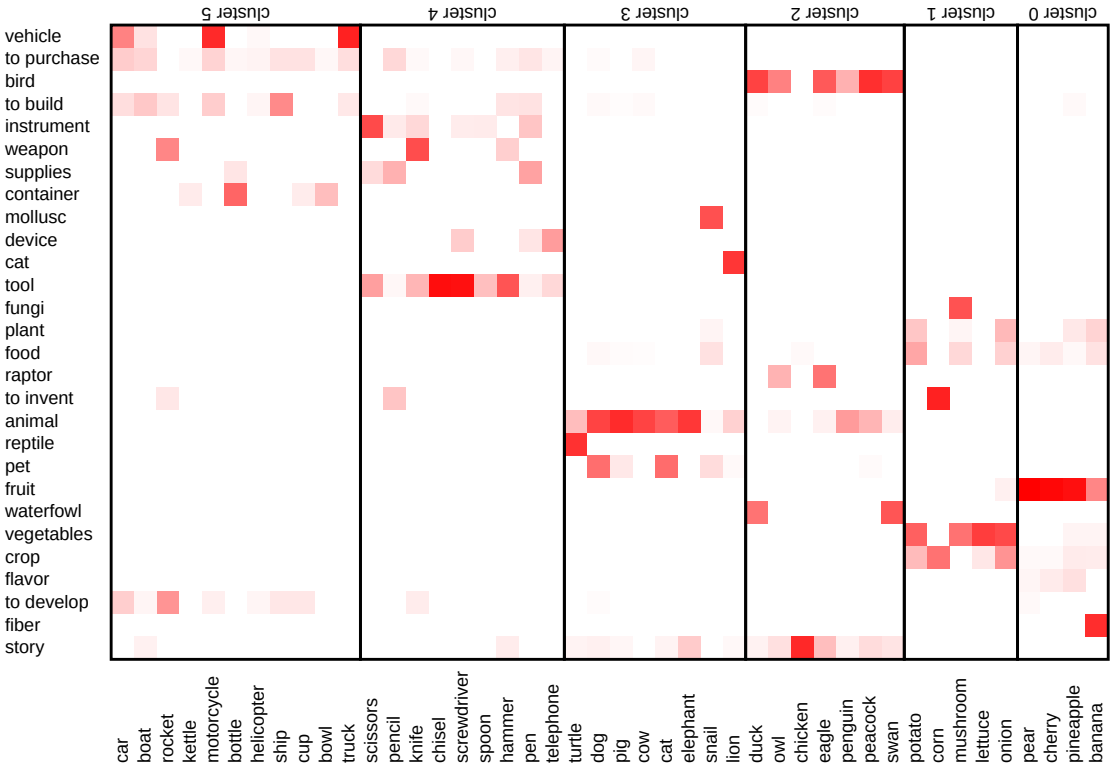



Figure 1: A combination of the formal and the agentive roles

but as a subcluster they occur in a wrong cluster which describes vehicles rather than tools.

CLUTO also provides a possibility to analyze how similar a given element is to other members in the same cluster. We expect a misclassified instance to be an outlier in a cluster. For this reason, we look at the $z$-scores of *rocket*, *bowl*, *cup*, *bottle* and *kettle*. There are two types of $z$-score, internal and external. Given an object $j$ which belongs to the cluster $l$, the internal $z$-score is computed as follows:

$$z_I = \frac{s_j^I - \mu_l^I}{\delta_l^I} \quad (2)$$

In Eq. 2 $s_j^I$ stands for the average similarity between the object $j$ and the rest objects in the same cluster, $\mu_l^I$ is the average of $s_j^I$ values over all objects in the $l$th cluster, and $\delta_l^I$ is the standard deviation of the similarities.

The external $z$-score is defined in a similar way with the only distinction that similarity is measured not with the object in the same cluster but in all other clusters.

The core of the cluster respresenting tools is formed by *chisel* followed by *knife* and *scissors* as they have the largest internal $z$-score. When the formal role only is used, the same cluster wrongly contains *rocket* but according to the internal $z$-score, it is an outlier (with the lowest $z$-score in the cluster). What concerns the "container" subcluster is the cluster of vehicles, *bowl*, *cup*, *bottle* and *kettle* all have the lowest internal $z$-scores. The core of the cluster of vehicles is a *truck* and *motorcycle*.

By examining features we found several types of errors:

1. Errors occurring at the extraction stage
2. Lexical ambiguity and erroneous statements in text

The first type of errors can be further classified as errors due to the incorrect tagging or due to the imperfect extraction mechanism. In general, seeds we have can be PoS ambiguous and to take this ambiguity into account, we put strict restrictions on the pattern matching. None of the seeds which were tagged

| | Cluster | Features |
|---|---|---|
| A | fruitTree | fruit (90.1%), fiber (5.2%), plant (0.6%), food (0.5%), flavor (0.5%) |
| | green | vegetables (56.4%), crop (25.2%), food (4%), fungi (3.6%), greens (2.8%) |
| | bird | bird (60.9%), story (10.7%), waterfowl (6.4%), animal (3.2%), raptor (3.1%) |
| | groundAnimal | animal (68.5%), pet (7.8%), reptile (2.5%), cat (2.4%), mollusc (1.8%) |
| | tool | tool (53.8%), weapon (9.4%), instrument (7.6%), supplies (5.4%), device (1.8%) |
| | vehicle | vehicle (42.8%), container (10.8%), aircraft (5.1%), appliances (4.9%), load (3.5%) |
| B | fruitTree | fruit (46.1%), animal (10.5%), tool (6.5%), bird (5.6%), vehicle (3.2%) |
| | green | vegetables (28.9%), crop (12.2%), animal (10.7%), tool (6.6%), bird (5.7%) |
| | bird | bird (34.7%), tool (8.4%), fruit (6.2%), vegetables (4.2%), vehicle (4.1%) |
| | groundAnimal | animal (31.4%), tool (8.7%), bird (7.4%), fruit (6.4%), vegetables (4.4%) |
| | tool | tool (28.1%), animal (12.4%), bird (6.6%), fruit (5.7%), weapon (4.9%) |
| | vehicle | vehicle (22.6%), animal (11.4%), tool (7.0%), bird (6.0%), container (5.7%) |

Table 6: $6-$way clustering: descriptive vs. discriminative features

as verbs are matched in text which narrows down a number of extracted terms. However, our analysis reveals that in most cases seeds tagged as verbs are not ambiguous and these are the PoS tagging errors. Besides, we mostly rely on the PoS information and some errors occur because of the imperfect extraction. For instance, if PP attachment is not handled, the extracted terms are most often incorrect.

The second type of errors is attributed to the lexical ambiguity such as the sentence below:

(3) in fact, scottish gardens are starting to see many more butterflies including peacocks, ...

From here, by using patterns for the formal role we get that *peacock* is a kind of *butterfly* which is not correct (according to the gold standard). However, we hope that such rare incorrect extractions will not play a significant role while clustering, if enough evidence is given. In our experiments *peacock* is always correctly categorized as a bird because of many other features which have been extracted. In particular, a feature *bird* has the highest frequency for this particular seed and, consequently, other rare features contribute to a lesser degree.

## 7 Conclusions

We presented a method for the concrete noun categorization which uses a notion of qualia structures. Our initial hypothesis of the difficulty of distinguishing between very general levels of categorization was supported by the empirical findings. While all tools and vehicles are always correctly identified as artifacts, vegetables are not classified as a natural category. The 3-way categorization provides the best performance by correctly identifying categories for all words. The 6-way categorization reveals difficulties in discriminating tools and vehicles. In particular, all containers are grouped together but incorrectly placed in the cluster *vehicle*. The most difficult element to classify is *rocket* which according to the gold standard is a vehicle. However, most features describing it are related to weapon and it is less surprising to find it in a category *tool* with such words as *knife* sharing this feature. When agentive role information is added, *rocket* is finally correctly classified as a vehicle.

Regarding qualia roles, the *formal* role is already sufficient to discriminate well between 3 categories. Adding extra information on other roles such as telic and constitutive does not improve results. By inspecting features which are extracted for these roles we can conclude that many of them are relevant (especially for the telic role) but there are hardly any members of the same category which would share them. This finding is in line with (Cimiano and Wenderoth, 2007) who mentioned the telic role as the one humans mostly disagree on.

As possible future directions, it would be interesting to experiment with other corpora. We have conducted all experiments with the Web data because of our intention to capture different qualia elements for each role. We noticed however that some pat-

terns proposed by (Cimiano and Wenderoth, 2007) are more suitable for the artifacts than natural kinds. More general patterns such as those by (Yamada et al., 2007) might be more helpful for the telic role but all candidates must be ranked to select the best suitable.

## References


Philipp Cimiano and Johanna Wenderoth. 2007. Automatic Acquisition of Ranked Qualia Structures from the Web. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007, pages 888-895*.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text data. *In Proceedings of COLING-92, pp. 539-545*.

Alessandro Lenci. 2000. Building an Ontology for the Lexicon: Semantic Types and Word Meaning. *In Ontology-Based Interpretation of Noun Phrases, Department of Business Communication and Information Science, University of Southern Denmark, Kolding:103-120* .

James Pustejovsky. 2001. Type Construction and the Logic of Concepts. *In The Syntax of Word Meaning, P. Bouillon and F.Busa (eds.) 2001, Cambridge University Press*.

James Pustejovsky. 2000. Syntagmatic Processes. *In Handbook of Lexicology and Lexicography, de Gruyter, 2000*.

Magnus Sahlgren. 2006. The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *PhD thesis, SICS Dissertation Series 44*.

Ichiro Yamada, Timothy Baldwin, Hideki Sumiyoshi, Masahiro Shibata and Nobuyuki Yagi. 2007. Automatic Acquisition of Qualia Structure from Corpus Data. *IEICE Transactions Inf. & Syst., vol. E90-D*.

ClaudiaZavaglia and Jualiana Galvani Greghi. 2003. Homonymy in Natural Language Processes: A Representation Using Pustejovsky's Qualia Structure and Ontological Information. *In Computational Processing of the Portuguese Language*.

Y. Zhao and G. Karypis. 2006. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. *In CIKM 2002*.

# Beyond Words:
# Semantic Representation of Text in Distributional Models of Language

**Kirill Kireyev**
Department of Computer Science
University of Colorado, Boulder
***kireyev@colorado.edu***

## Abstract

Empirical evaluations of distributional lexical semantic models (LSA, Topics, HAL) have largely focused on word-level tests (such as synonym and word association tests). This paper motivates the need to focus more on representation of larger units of text. A suite of evaluation metrics is proposed and used to compare the performance of two prominent distributional lexical models: LSA and Topics. Theoretical observations and broader implications related to text representation are discussed.

## 1 Introduction

Corpus-based distributional models are mathematical models of language that derive semantic representations of word meanings from patterns of occurrence of words in a large corpus of natural language. A prominent subset of these models are called vector space models, because they attempt to represent meanings of words and documents in a high-dimensional geometric space.

On the practical side, distributional models provide an efficient and robust way to represent semantics of words and text, which is useful in various Natural Language Processing applications, from information retrieval [2] to intelligent tutoring systems ([5], [11], [7]). On the theoretical side, such methods offer a way to model important processes in human cognition and language acquisition ([12], [8], [13]).

Several distributional models have gained prominence in the past few decades. Among them are Latent Semantic Analysis (LSA, [2]), Probabilistic Topics Model (Topics, [8]), Hyperspace Analogue to Language (HAL, [14]), Bound Encoding of the Aggregate Language Environment (BEAGLE, [9]) and others. With the emergence of several different models, it becomes natural to attempt to compare performance characteristics of these models.

Traditionally, tests of how well a particular model represents meaning, have largely revolved around word-level comparisons. The most common such metric is a synonym test, in which the model uses semantic similarity measurements between words to predict the synonym for a given cue word, among possible choices (akin to synonym questions presented to students on the TOEFL test; see [12]). On this task, models have shown performance equivalent to that of of college-admitted students. A related test is a word association task, in which the model attempts to imitate human performance on producing the first word that comes to mind in response to a particular cue word. Some other interesting word-level tasks are described in [8].

Although analyzing performance at the word level may provide some interesting insights into the behavior of the models, it does not capture important linguistic phenomena, as words are rarely used in isolation in natural language. Rather, we tend to combine words into larger structures, such as sentences and documents, to communicate meaning. Therefore, tasks focusing on isolated

words, such as synonym questions, are not a realistic reflection of use of language, and as such, constitute a questionable standard of performance for computational models.

Therefore, an especially important aspect of distributional models is their ability to represent semantic meaning of sentences, paragraphs and documents. This not only reflects natural linguistic phenomena, but also enables many useful practical applications such as information retrieval ([2]), intelligent tutoring systems ([5], [11]), conversation analysis ([4]), and others.

In this paper, we discuss some methods of evaluating semantic representations in distributional models on document-, paragraph- and sentence- length passages. We propose a suite of tests that we hope provide richer and more realistic evaluation metrics than traditional word-level tests. We demonstrate the use of these tests to evaluate performance of two prominent models (LSA and Topics). We then discuss some theoretical issues of semantic representation and its evaluation.

## 2 Models

LSA and Topics are two prominent and contrasting models of language. They both have the facility to represent and compare meanings of both words and text passages, though they each do so in very different ways. In this section we give very brief mathematical descriptions of these models.

### 2.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis ([2]) is an unsupervised methods of deriving vector space semantic representation from a large corpus of texts. LSA starts by representing a collection of documents by a term by document (*T x D*) matrix *A*, which in essence represents each word by a D-dimensional vector. It then performs singular value decomposition (SVD) on the matrix:

$$A = U \Sigma V^T \quad (1)$$

Subsequently, all but the first (largest) *k* values in the diagonal singular matrix Σ, are set to zero, resulting in a kind of principal component analysis. This effectively reduces the dimensionality of each word vector to *k*. (For more details, please consult [2]). The number of dimensions (*k*) is determined empirically. The dimensions have no intuitive interpretation; they simply serve to position word vectors in the high-dimensional space.

The measure of semantic similarity between two words in this model is typically[1] the cosine of the angle between their corresponding word vectors :

$$S(w1, w2) = \cos(v_{w1}, v_{w2}) = \frac{v_{w1} \cdot v_{w2}}{\|v_{w1}\| \|v_{w2}\|} \quad (2)$$

The simulated meaning of a new document (sometimes referred to as pseudo-document) can be represented in LSA using the following method:

$$v_d = q^T U_k \Sigma_k^{-1} \quad (3)$$

where *q* represents the array containing type frequencies for words in the document (weighted by *tf-idf*-derived entropy weights). Note that this is equivalent to (weighted) geometric addition of constituent word vectors corresponding to words in a document.. As a result, both words and documents are represented as vectors in *k*-dimentional space[2], allowing for straightforward word-word, word-document, and document-document comparisons., which reflect their semantic similarity according to the model:

$$S(w, d) = \cos(v_w \Sigma^{1/2}, v_d \Sigma^{1/2}) \quad (4)$$

$$S(d_1, d_2) = \cos(v_{d1} \Sigma, v_{d2} \Sigma) \quad (5)$$

The absolute values of cosines (which may range between -1 and 1 with larger values indicating greater similarity) have no strict interpretation; only comparisons of cosine values (e.g. between *pairs* of words) are meaningful.

### 2.2 Topics Model (LDA)

The Topics model ([8]) is a generative probabilistic model of language. It is sometimes referred to as LDA, because it is based on Latent Dirichlet Allocation (see [8]). At the heart of the model, is the assumptions each document may be represented by a mixture of topics (for example, a

[1]Other metrics like Eucledian distance and dot product are less commonly used

[2]Depending on the type of comparison, operands need to be multiplied by the singular matrix Σ (word-word) or its square root (word-doc). Please see LSA literature for more details.

news article about the use of steroids in baseball will likely have significant contributions from topics corresponding to "sports", "drugs" and "health"). Although not strictly a vector space model, it exhibits all of the relevant details of representation. Each word *w* in the Topics model is represented by a multinomial distribution over topics ($\varphi(w)$), which can be thought of as a multidimensional vector, where each dimension is the strength of association of the word with the particular topic.

The topics are derived in an unsupervised manner; only the number of topics (*T*) is set a-priori. Topics can be characterized by most probabilistically representative words, and are generally intuitively interpretable, *e.g.* $T_1$ = *{print, paper, ink, ...}*, $T_2$ = *{team, game, basketball, ...}*, etc.

A particular document is represented, as a mixture of topics ($\theta(d)$), a multinomial distribution derived from the particular topics assignments of its constituent word tokens. Note that while a *word type* is represented as a distribution ($\varphi(w)$), a particular *word token* in a document is assigned to a discrete topic. The assignment is based on the joint probability of its word type ($\varphi(w)$) and the probability of a particular topic manifesting in this document, which in turn is derived from topic assignments of other tokens.

Since both words and documents are represented as multinomial distributions ($\varphi(w)$, and $\theta(d)$ respectively), their semantic distance can be measured using the Kullback-Leibler (KL) divergence[3]:

$$KL(p,q)=\sum_{j=1}^{T} p_j \log_2\left(\frac{p_j}{q_j}\right) \quad (6)$$

which can be converted into a symmetrical similarity[4] measure (KLS):

$$KLS(p,q)=-\frac{1}{2}(KL(p,q)+KL(q,p)) \quad (7)$$

(this measure can range from negative infinity to zero). To compute the similarity between a word and a document, we measure the probability of a particular word occurring in a particular document:

$$P(w|d)=\sum_{z} P(w|z)P(z|d)=\phi(w)\cdot\theta(d) \quad (8)$$

Please refer to ([8], [16]) for more details.

### 2.3 Model Implementations

In our experiments we trained both models on the TASA corpus, containing roughly 44,000 reading passages for school children through college level. Each passage (document) contained around 300 words. The number of dimensions on LSA (*k*), as well as the number of topics in the Topics model (*T*) was set to 300. For the Topics model we used hyper-parameter values $\alpha$=50/T, $\beta$=0.01, N=500 (see [8] for more details). It should be noted that better performance on specific tasks may be accomplished by a more thorough analysis of the optimal parameter settings.

## 3 Evaluations

### 3.1 Primary Topic Detection

In this test we compute the representation of selected documents in each respective model, and ask the models to select the word that best represents the semantic content of a document, i.e. the word that is most semantically similar to the document. This corresponds to asking someone to describe in one word what a document is about. While on the surface, this resembles the problem of *keyword extraction*, it should be noted that this is not the ultimate goal. Rather we use the words to probe the models' representations of text.

To measure model's performance, we compare the model's answer with the document's actual (human-generated) title. More specifically, each model ranks all the words in the document by how well they represent the document (in decreasing order). The model's performance score is the rank of the actual title word in this sorted list, normalized by the total number of words in the list:

$$score_1(d) = rank(title_d) / \#unique\_words(d) \quad (9)$$

and falls between 0 (good; actual title is at the top of the list) and 1 (bad; actual title is at the bottom of the list). This allows the score to be independent both of the document size and the scaling of similarity metric.

[3]Some other metrics are possible, like Jensen-Shannon divergence. See [16] for more details.

[4]The *cosine* metric used in LSA is also a similarity measure; higher values means greater similarity.

We used 46 Wikipedia ([17]) articles from 6 different categories, as test documents:

| | | |
|---|---|---|
| Sports | Animals | Countries |
| Sciences | Religions | Diseases |

In the table below both the score for each article, as well as the word that each model picked as the most representative, are presented. Words that resemble the title (modulo stemming) are highlighted.

| **Original title** | **LSA** | | **Topics** | |
|---|---|---|---|---|
| | **top word** | **score$_1$** | **top word** | **score$_1$** |
| Baseball | player | 0.00135 | basketball | 0.0127 |
| Boxing | position | 0.02789 | championship | 0.0423 |
| Golf | player | 0.00480 | hockey | 0.0210 |
| Gymnastics | **gymnastics** | 0.00000 | athletes | 0.0409 |
| Tennis | players | 0.01308 | championship | 0.0111 |
| Hockey | players | 0.01146 | basketball | 0.0514 |
| Skiing | **skiing** | 0.00000 | **skis** | 0.0528 |
| Fencing | tournaments | 0.47745 | championship | 0.5255 |
| Zebra | species | 0.00404 | lion | 0.0319 |
| Giraffe | **giraffe** | 0.00000 | lion | 0.0140 |
| Bear | **bears** | 0.00302 | lion | 0.0070 |
| Deer | antelope | 0.00655 | zoo | 0.0112 |
| Wolf | **wolves** | 0.00370 | dog | 0.0051 |
| Fox | rodents | 0.05575 | dog | 0.0033 |
| Elephant | **elephants** | 0.00296 | zoo | 0.0028 |
| Tiger | elephants | 0.03197 | zoo | 0.0132 |
| Russia | **russian** | 0.00058 | soviet | 0.0089 |
| Germany | **germany** | 0.00000 | hitler | 0.0144 |
| Canada | **canadian** | 0.00161 | united | 0.0038 |
| Sweden | dominated | 0.02231 | asia | 0.0361 |
| Thailand | buddhism | 0.01402 | mali | 0.0692 |
| Kenya | uganda | 0.00115 | africans | 0.0175 |
| Australia | commonwealth | 0.00187 | latitude | 0.0044 |
| Brazil | **brazil** | 0.00000 | columbus | 0.0116 |
| Biology | **biologists** | 0.00816 | ecosystem | 0.0186 |
| Chemistry | **chemistry** | 0.00000 | hydroxide | 0.0783 |
| Physics | **physicists** | 0.00114 | theory | 0.0276 |
| Psychology | **psychology** | 0.00559 | **psychologist** | 0.0160 |
| Mathematics | **mathematical** | 0.00129 | hypotheses | 0.0148 |
| Sociology | **sociologists** | 0.00000 | emphasizes | 0.0601 |
| Economics | **economists** | 0.00142 | prices | 0.0381 |
| Geography | **geography** | 0.00000 | cultures | 0.0273 |
| Christianity | **christian** | 0.00155 | bishop | 0.0306 |
| Islam | prophet | 0.00000 | jesus | 0.0413 |
| Judaism | **judaism** | 0.00000 | egyptian | 0.0663 |
| Hinduism | **hinduism** | 0.00000 | rites | 0.0461 |
| Buddhism | philosophy | 0.00069 | thou | 0.1712 |
| Cancer | abnormal | 0.00259 | disease | 0.0241 |
| AIDS | incidence | 0.16098 | viruses | 0.2953 |
| Asthma | symptoms | 0.00453 | disease | 0.1066 |
| Syphilis | symptoms | 0.00514 | disease | 0.0722 |
| Flu | viral | 0.01984 | disease | 0.0359 |
| Pneumonia | infections | 0.00854 | disease | 0.0913 |
| **Mean (stdev)** | | 0.0211 (0.0758) | | 0.0529 (0.0900) |

**Table 1.** Performance on predicting the main theme (title)

In this task LSA outperforms the Topics model, having generally lower score, i.e. having the actual title word appear closer to the top of the list.

## 3.2 Text Categories

In this experiment we compare the semantic similarity judgments between articles within the same topic category and across categories. One would expect articles within the same category to be more similar, compared to articles across different categories. Hence, a model that adequately represents meaning of test should reflect greater similarity of within-category articles. We use the Wikipedia articles and categories described in section 3.1.

The scoring metric used in this test is the difference between average similarity μ (using metrics (5) and (7) for LSA and Topics, respectively) for within-category articles and across-category articles , normalized by standard deviation δ of across-category similarity:

$$score_2(C) = (\mu_{d1,d2 \in C} - \mu_{d1 \in C', d2 \in C''}) / \delta_{d1 \in C', d2 \in C''} \quad (10)$$

| **Categories** | **LSA score$_2$** | **Topics score$_2$** |
|---|---|---|
| **Sports** | 4.59 | 2.76 |
| **Animals** | 5.49 | 3.38 |
| **Countries** | 4.99 | 3.67 |
| **Sciences** | 3.11 | 1.82 |
| **Religions** | 6.30 | 3.72 |
| **Diseases** | 6.04 | 3.50 |

**Table 2.** Average semantic similarity of articles within the same category, compared to across-categories

Both models were correctly evaluate within-category articles to be more semantically related, as all scores are greater than zero. The pairwise t-test shows that LSA's scores within-category articles to be consistently higher.

We also used pairwise semantic distances between articles as input to agglomerative clustering, to see if the articles belonging to the same categories, will be clustered together

automatically. The clustering based on either of the models' calculated distances reconstructed the original groupings perfectly.

### 3.3 Text Coherence

Previous research ([4], [6]) employed distributional semantic models to measure coherence, or semantic connectedness of text. The assumption underlying it, is that adjacent paragraphs in expository texts will have higher semantic similarity, so as to ensure smooth, logical transitions between content segments. Furthermore, the difficulty of text should be inversely related to its cohesion; i.e. one would expect simpler texts (written for lower grade levels) to have higher cohesion.

In this experiment, we compute measures of coherence between paragraphs of text taken from chapters of science textbooks written for 3rd and 6th grades. We use text similarity metrics (equations (5) and (7)), to compare the mean semantic similarity between adjacent and non-adjacent paragraphs. The score is computed as the difference in the mean similarity ($\mu$) between adjacent and non-adjacent paragraphs in each text, normalized by the standard deviation ($\delta$) of non-adjacent paragraphs.

$$score_3(d) = (\mu_{adjacent} - \mu_{non\text{-}adjacent'}) / \delta_{non\text{-}adjacent'} \quad (11)$$

Another interesting task is to attempt to reconstruct the original order of the paragraphs, based on their pairwise semantic distances. This can be done using Multidimensional Scaling (MDS, [1]) to 1 dimension. This has the effect of attempting to fit all paragraphs along a line, in such a way as to maximally satisfy their pairwise distances (as reported by either of the models). One can then compare the resulting order of paragraphs to their original order in the text. This can be done by computing the Kendall tau rank correlation with the original order, essentially counting the number of indices that are "in order". We take the absolute value of the metric, since MDS may arrange results in "reverse" order:

$$score_3' = abs(\ Corr_{kendall}(\ mds(M), <1..P>)\ ) \quad (12)$$

In equation (12), M is the *PxP* pairwise semantic distance metric between *P* paragraphs. A higher correlation ($score_3'$), means that the original order of paragraphs can be reconstructed more accurately, using the output of the model.

| Grade | Text | $score_3$ | | $score_3'$ | |
|---|---|---|---|---|---|
| | | LSA | Topics | LSA | Topics |
| Grade 3 | 1 | 0.58 | 0.34 | 0.35 | 0.19 |
| | 2 | 0.65 | 0.52 | 0.52 | 0.26 |
| | 3 | 1.42 | 0.37 | 0.71 | 0.36 |
| | 4 | 0.35 | 0.25 | 0.30 | 0.33 |
| | 5 | 0.75 | 0.81 | 0.58 | 0.41 |
| **Mean (grade 3)** | | **0.75** | **0.46** | **0.49** | **0.31** |
| Grade 6 | 1 | -0.07 | -0.03 | 0.18 | 0.28 |
| | 2 | 0.34 | 0.35 | 0.30 | 0.00 |
| | 3 | 0.46 | 0.41 | 0.32 | 0.24 |
| | 4 | 0.74 | 0.87 | 0.27 | 0.08 |
| | 5 | 0.21 | 0.15 | 0.09 | 0.26 |
| **Mean (grade 6)** | | **0.34** | **0.35** | **0.23** | **0.17** |

**Table 3.** Semantic similarity of adjacent paragraphs compared to non-adjacent, and measure of accuracy in recreating the original paragraph score using MDS.

Both models show greater similarity between adjacent paragraphs (with one exception). Also, as expected, the models show that the (more difficult) sixth grade reading exhibits less coherence than the (easier) third grade reading. Attempting to restore original paragraph order was more effective using the output of LSA.

### 3.4 Sentence Coherence

In this test we use the data originally reported by McNamara et al ([15]). The authors used both distributional model (LSA) and human similarity judgments on a 6-point Likert scale to judge similarity between pairs of sentences that were (a) human-generated paraphrases, (b) adjacent sentences in expository text, (c) non-adjacent sentences in the same text, (d) random sentences from different texts. In general one would expect the sentence similarity to range from highest to lowest in the categories above, i.e.

$$(paraphrase) > (adjacent) > (jump) > (random) \quad (13)$$

In this experiment we computed sentence similarity scores with LSA and Topics (equations (5) and (7), respectively) and correlations with

human ratings reported by McNamara et al. We also show the item-by-item correlations between the models and human judgments.

| | Paraphrase | Adjacent | Jump | Random |
|---|---|---|---|---|
| **Human (Likert scale)** | 5.38 | 2.89 | 2.41 | 1.89 |
| **LSA (cosine sim)** | 0.53 | 0.35 | 0.27 | 0.00 |
| **Topics (KLS similarity)** | -0.35 | -0.38 | -0.37 | -0.50 |
| **Correlation (LSA, Human)** | 0.26 | 0.43 | 0.27 | 0.10 |
| **Correlation (Topics, Human)** | -0.48 | 0.08 | 0.06 | -0.33 |

**Table 4.** Semantic similarity scores and correlations with human judgments for sentences.

LSA measurements maintain the inequality (12), while Topics does not. LSA also maintains a more reliable correlation with human similarity judgments.

### 3.5 Sentence Paraphrases

In this experiment we used a collection of sentence pairs from Microsoft Research Paraphrase Corpus [3]. The sentence pairs were annotated by 2 human judges who were asked to give a binary judgment as to whether a pair of sentences could be considered "semantically equivalent" (1 – yes, 0 - no). We filtered out sentences containing rare words, resulting in a collection of 1323 pairs.

We computed the correlation between the human judgments, and the similarity scores between sentences in each pair, as reported by each model:

| | LSA | Topics |
|---|---|---|
| **Pearson Correlation** | 0.205 | 0.067 |

**Table 5.** Correlation of LSA and Topic similarity scores with human judgments.

LSA model produced higher correlation scores with human annotators.

## 4 Discussion

Currently, no agreed standard for evaluating meaning of texts beyond the word level exists. The traditional approach to evaluating many Natural Language processing algorithms is to create a gold standard of human-annotated data, such as similarity judgments between sentences used in (3.4) and (3.5). However, human semantic similarity judgments beyond sentence level processing and simple binary judgments (same / not same), are a complex and subjective task. Furthermore they are not a natural task that people regularly perform with language. As such, these explicit judgments would be an unreliable metric.

We have proposed a few measures of estimating how well distributional models can represent text at the document (3.1 and 3.2), paragraph (3.3) and sentence (3.4 and 3.5) level. These measures are grounded in natural linguistic artifacts, such as expository texts. While no single one of these measures is the definitive test of semantic representation, together, the collection of tests can be used to paint a picture of strengths and weakness of various semantic models.

Even though applications that rely on semantic representation of text with distributional models have been used for some time, some important details of these models have been overlooked.

For example, consider the following three sentences and similarity scores between them, as calculated by LSA:

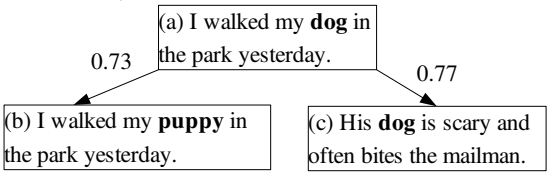



**Figure 1.** LSA semantic similarity scores between three sentences.

The result that (a) and (c) are deemed more similar than (a) and (b) should be counter-intuitive, given that (a) and (b) differ only by a single word, and the different words are roughly synonymous! The reason for this phenomenon can be found in the fact that the vector length for the word *dog*[5] is almost an order of magnitude greater than any of the other words in any of the sentences, and therefore largely dominates the meanings of sentences (a) and (c). Walter Kintsch writes in [10]:

> "Intuitively, the vector length tells us how much information LSA has about this vector. [...] Words that LSA knows a lot about (because they appear frequently in

[5]Word vectors are multiplied by entropy weights when being combined into documents.

the training corpus, in many different contexts) have greater vector lengths than words LSA does not know well. Function words that are used frequently in many different contexts have low vector lengths -- LSA knows nothing about them and cannot tell them apart since they appear in all contexts."

Roughly speaking, the vector length (and hence its level of semantic influence) in LSA is determined by frequency of its occurrence and its *concreteness* (the two factors may be in competition). Let's take a closer look as to whether this in fact corresponds to intuition.

It is true that a word that has only been encountered by the model once or twice, ought to have much uncertainty in its meaning, and, therefore, a small semantic contribution (i.e. short vector length). However, after a certain threshold number of occurrences, we would expect the word's meaning representation to stabilize and no longer be altered by subsequent encounters. For example, even though the word *canine* occurs almost two orders of magnitude less frequently than *dog* in language (in TASA corpus the word counts are 2962 and 17, respectively), it's difficult to argue that once its meaning is understood, *canine* should provide an equivalent semantic contribution to *dog.* Yet in LSA, the vector for *dog* is roughly 22 times greater. McNamara et al ([15]) have experimented with different word weighting schemes. One of the finding was that giving greater weight to *rare* words result results in better performance on some of the tasks involving semantic representations of sentences and passages. Along the same lines, the common practice of using a stop word lists to discard words that are too common (like *the, of*) is much more of a heuristic than a theoretically-motivated technique. The real solution would involve analysis of many factors, such as psycholinguistic characteristics, parts of speech, propositional attachment and so on.

In LSA, words that have more concretely defined meanings will tend to have longer vector lengths compared to more polysemous or abstract words, when controlled for the frequency of occurrence. For example, the vector for *cat* (occurring 1461 times in TASA) is 1.5 times longer than *bank* (occurring 1567 times). In other words, the vector for *bank* is shorter because of ambiguity of its meaning ("house for money" or "edge of river"). But while the meaning of *bank* is ambiguous in the abstract, it becomes clarified in most practical contexts. Once the word is disambiguated (i.e. in the context of a pseudo-document) to one concrete sense, it should no longer carry a "uncertainty penalty" by having a shorter vector length. By contrast, in the Topics model word tokens are automatically disambiguated when the pseudo-document representation is created. Kintsch has proposed a method of word sense disambiguation for LSA in [10], but it is far from systematic.

In Topics, all word tokens contribute equally to the the overall representation of a pseudo-document ($\theta(d)$); no weighting scheme is employed. As mentioned earlier, unlike LSA, the semantic contribution of each word may vary between documents, depending on the context, since the words are disambiguated. This may have some unpredictable effects, however. Consider for example the topic assignments generated by the Topics model[6] for words in the following two sentences, which differ by one word:

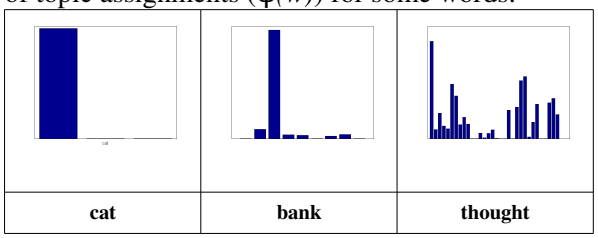
| A **boy** was holding a dog by a leash when the leash broke | | A **girl** was holding a dog by a leash when the leash broke | |
|---|---|---|---|
| Token | Topic | Token | Topic |
| boy | 61 | girl | 74 |
| **was** | **135** | **was** | 77 |
| **holding** | **144** | **holding** | **128** |
| dog | 148 | dog | 148 |
| **when** | **203** | **when** | **207** |
| leash | 148 | leash | 148 |
| **broke** | **116** | **broke** | **66** |

**Table 5.** Topic assignments for word tokens in the two sentences.

Note how most of the words change their assignment, depending on presence of *boy* or *girl* in the sentence. This behavior is particularly true of more abstract or polysemous words, since in the Topics model such words tend to have a wider spread of possible topics, in contrast to concrete words, which tend to be assigned to only a few

[6]Topics model trained with T = 308. Remember that it's not important here what individual topics mean; only whether they are the same or different.

possible topics. For example, consider histograms of topic assignments ($\varphi(w)$) for some words:



| cat | bank | thought |
|---|---|---|

**Table 6.** Histograms for topic assignments in the Topics model for some words.

This behavior has significant implications for computing semantic representation of texts in the Topics model, since many of the abstract words tend to have more unstable topic assignments and, thus, small variations may drastically affect the overall topic mixture of the document.

One interesting observation of this study is that LSA generally outperforms the Topics model at the text-level tasks, and therefore is likely to represent meaning of text passages better. This is despite showing somewhat lower performance on word-level tasks such as synonym tests and word priming. (see [8]). Aside from the possibility of obtaining better performance by using different model parameters or comparison functions, two explanations can be plausible for this discrepancy. One is that Topics, as a model, is simply not as well equipped to represent compositional meaning.

A related, and more interesting explanation, however, is that the word-level tasks are not well suited as theoretical insights into language and cognition. One can argue that human judgments of synonymy and word associations are not natural tasks, because words in language are never used or learned in isolation. Therefore, it might be entirely plausible, for example, that the kind of associations that occur when the brain represents meanings of words are very different from those that synonym judgments would suggest. In other words, explicit word similarity judgments may be a high-level cognitive task (e.g. based on explicitly enumerating and comparing attributes) rather than a direct insight into meaning representation in the brain.

## 5 Conclusions

In this paper we have started exploring issues of semantic representation of text passages and methods for their evaluation. Many questions remain open in this area. The answers will likely only come from combining advancements in computer science, neurophysiology, cognitive psychology and linguistics

## References


[1] Borg, I. and Groenen, P.: "Modern Multidimensional Scaling: theory and applications", Springer-Verlag New York, 1997

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407

[3] Dolan W. B., C. Quirk, and C. Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. COLING 2004, Geneva, Switzerland. (Corpus available from http://research.microsoft.com)

[4] Foltz, P. W., Kintsch, W.,& Landauer, T. K. (1998). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 285-307.

[5] Foltz, P., Laham, D., Landauer, T.: The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning (1999).

[6] Graesser, A., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36,* 193-202.

[7] Graesser, A.C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* . Mahwah, NJ: Erlbaum

[8] Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114,211-244.

[9] Jones, M, N., Mewhort Doulgas, J. K., "Representing Word Meaning and Order Information in a Composite Holographic Lexicon", *Psychological Review*, v114 n1 p1-37 Jan 2007

[10] Kintsch, W., Predication. *Journal of Cognitive Science,* 25 (2001).

[11] Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street ® : Computer-guided summary writing. In T. K. Landauer, D. M., McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis.*

[12] Landauer, T. K., On the Computation Basis of Learning and Cognition. In N. Ross (Ed.), *The Psychology of Learning and Motivation,* 41, 43-84.

[13] Landauer, T. K. and Dumais, S. T. (1997) A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2) , 211-240

[14] Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. Cognitive Science Proceedings (LEA), 660-665.

[15] McNamara, D. S., Cai, Z., Louwerse, M. M., Optimizing LSA Measures of Cohesion. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* . Mahwah, NJ: Erlbaum

[16] M. Steyvers, T. Griffiths, Probabilistic Topic Models. In T. K. Landauer, D. M., McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis*

[17] http://wikipedia.org

# Size Matters.
# Tight and Loose Context Definitions in English Word Space Models

**Yves Peirsman, Kris Heylen and Dirk Geeraerts**
Quantitative Lexicology and Variational Linguistics (QLVL)
University of Leuven
Leuven, Belgium
{yves.peirsman|kris.heylen|dirk.geeraerts}@arts.kuleuven.be

## Abstract

Word Space Models use distributional similarity between two words as a measure of their semantic similarity or relatedness. This distributional similarity, however, is influenced by the type of context the models take into account. Context definitions range on a continuum from tight to loose, depending on the size of the context window around the target or the order of the context words that are considered. This paper investigates whether two general ways of loosening the context definition — by extending the context size from one to ten words, and by taking into account second-order context words — produce equivalent results. In particular, we will evaluate the performance of the models in terms of their ability (1) to discover semantic word classes and (2) to mirror human associations.

## 1 Introduction

In recent years, Word Space Models (Landauer and Dumais, 1997; Schütze, 1998; Padó and Lapata, 2007; Baroni et al., 2007) have become the standard NLP answer to any question concerning lexical semantics. Be it query expansion, automated essay rating, thesaurus extraction, word sense disambiguation or question answering, Word Space Models are readily applied to the task at hand. Their success almost makes us forget that the word space approach itself presents us with a number of questions. For instance: what kind of semantic relations are captured by these models? Is it semantic similarity — as between *car* and *truck* — or more topical relatedness — as between *car* and *road*? Moreover, what is the influence of all parameters involved — from the definition of context to the similarity measure used to compare the context vectors of two words? In this paper, we will focus on the precise definition of context that the models use and investigate its effect on the semantic relations that they find.

### 1.1 Word Space Models

In order to get at the semantic relatedness between two words, word space approaches model their use. They do so by recording in a so-called *context vector* the contextual features that each word co-occurs with in a corpus. For instance, first-order bag-of-word models simply keep track of the context words that appear within a context window of $n$ words around the target (Gale et al., 1994; Levy and Bullinaria, 2001; Bullinaria and Levy, 2007). This implies that two words are similar when they often co-occur with the same context words. The tightest definition of context for bag-of-word models restricts itself to one word to the left and right of the target. Because this restriction may lead to data sparseness, it is often loosened in one of two ways: either the context window is stretched to a higher number of words around the target (Sahlgren, 2006), or the models take into account not the direct context words of the target, but the context words of these context words (Schütze, 1998). In this paper, we will investigate whether these two ways of loosening the context definition have the same influence on the results of the Word Space Models.

Without any doubt, enlarging the context window will change the type of features that the models are based on. With just one word to the left and the

right of the target, an English noun will tend to have mostly adjectives and verbs as contextual features, for instance. Most of these context words will moreover be syntactically related to the target. If we extend the window size to five words, say, the noun's context vector will look very different. Not only are other nouns more likely to appear; the majority of words will not be in a direct syntactic relation to the target, but will merely be topically linked to it. We can expect this to have an influence on the type of semantic relatedness that the Word Space Models distinguish.

This effect of context has obviously been noted before. Sahlgren (2006) in particular observes that in the literature, all sorts of context sizes can be found, from fifty words to the left and right of the target (Gale et al., 1994) via fifty words in total (Schütze, 1998) to a mere three words (Dagan et al., 1993). Through a series of experiments, Sahlgren was able to confirm his hypothesis that large context windows tend to model syntagmatic — or topical — relations better, while small context windows are better geared towards paradigmatic — similarity or antonymy — relations. In a similar vein, we investigated the influence of several context definitions on the semantic characteristics of a wide variety of Word Space Models for Dutch (Peirsman et al., 2007; Peirsman, 2008). We found that syntactic models worked best for similarity relations, while first-order bag-of-word approaches modelled human associations better, among other things.

### 1.2 Research hypothesis

In line with Sahlgren (2006), our research hypothesis is that tight context windows will give better results for semantic similarity, while looser context windows will score higher with respect to more general topical relatedness. 'Loose' here refers to the use of a larger context window or of second-order context words.

We will test this hypothesis through a number of experimental tasks that have been released for the ESSLLI 2008 Lexical Semantics Workshop. First, section 2 will present the setup of our experiments. Section 3 will then discuss three word clustering tasks, in which the Word Space Models are required to discover semantic word classes. In section 4, we will investigate if the models are equally suited to model free associations. Finally, section 5 will wrap up with conclusions and an outlook for future research.

## 2 Experimental setup

The data for our experiments was the British National Corpus, a 100 million word corpus of British English, drawn from across a wide variety of genres, spoken as well as written. On the basis of this corpus, we constructed fourteen Word Space Models, seven first-order and seven second-order ones. Context size varied from 1 via 2, 3, 4, 5 and 7 to 10 words on either side of the target.

We reduced the dimensionality of the context vectors by treating only the 5,000 most frequent words in the BNC as possible features — a simple, yet popular way of dimensionality reduction (Padó and Lapata, 2007). Although working with all features could still improve performance (Bullinaria and Levy, 2007), we feel confident that cutting off at 5,000 dimensions has no direct influence on the relationships between the models, and the semantic relations they prefer. Semantically empty words in our stop list were ignored, and all words were lemmatized and tagged by their part of speech. In addition, we also used a cut-off that linearly increased with context size. For context size $n$, with $n$ words on either side of the target word, we only took into account a feature if it occurred at least $n$ times together with the target word. This variable cut-off keeps the number of non-zero cells in the word by feature matrices from exploding for the larger contexts.

The context vectors did not contain the frequency of the features, but rather their point-wise mutual information (PMI) with the target. This measure indicates whether the feature occurs together with the target more or less often than we would expect on the basis of their individual frequencies. Finally, the similarity between two context vectors was operationalized as the cosine of the angle they describe.

## 3 Task 1: Word Clustering

In Task 1, we tested the ability of our models to discover semantic classes for three types of words: concrete nouns, verbs, and a mixture of concrete and abstract nouns. The data sets and their sources are

described on the website of the ESSLLI workshop.[1] The set of concrete nouns consisted of words like *hammer*, *pear* and *owl*, which our models had to cluster into groups corresponding to a number of semantic classes. The output was evaluated at three levels. The most fine-grained class distinctions were those between tools, fruit, birds, etc. — six clusters in total. Next, we checked the models' ability to recognize the differences between artifacts, vegetables and animals. Finally, animals and vegetables had to be combined into one natural category.

The second test set consisted of a mixture of concrete and abstract nouns — *truth* and *temptation* versus *hammer* and *eagle*, for instance. Here, the models were simply required to make the distinction between concrete and abstract — a task they were well capable of, as we will see.

The final test set contained only verbs. Again the models were evaluated several times. At the first stage, with nine clusters, we checked for the distinction between verb classes like *communication* (e.g., *speak*), *mental state* (e.g., *know*) and *body action* (e.g., *eat*). At the second stage, with five clusters, the categories were reduced to the likes of *cognition* and *motion*.

The vectors output by the models were clustered with the *repeated bisections* algorithm implemented in CLUTO (Karypis, 2003). This is a so-called partitional algorithm, which starts with one large cluster that contains all instances, and repeatedly divides one of its clusters in two until the requested number of clusters is reached. The resulting clusters are then evaluated against two measures: entropy and purity.

The entropy of cluster $S_r$ of size $n_r$ is defined as follows:

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \qquad (1)$$

Here, $q$ is the number of word classes in the data set, and $n_r^i$ the number of words of class $i$ in cluster $r$. As always, entropy expresses the *uncertainty* of a cluster — the degree to which it mixes up several categories. The lower the entropy, the better the cluster.

Purity, next, is the portion of the cluster taken up by the largest class in that cluster:

[1] http://www.wordspace.collocations.de/doku.php/esslli:start

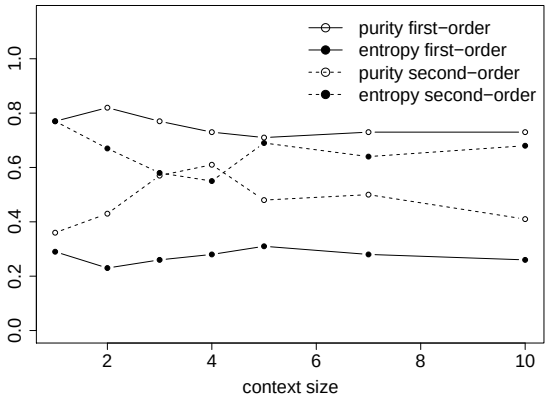



Figure 1: Performance of the Word Space Models in the 6-way concrete noun clustering task.

$$P(S_r) = \frac{1}{n_r} \max_i(n_r^i) \qquad (2)$$

The higher the purity of a given cluster, the better. The entropy and purity values of the total solution are simply the sums of the individual cluster scores, weighted according to cluster size.

### 3.1 Results

By way of example, Figure 1 shows the performance of the models in the 6-way concrete noun clustering task. A number of observations we can make here apply to all results in this section. First, the purity and entropy of the models are almost perfect mirror images of one another. Second, the performance of the first-order models is clearly superior to that of the second-order ones. Purity lies considerably higher; entropy much lower. Third, our expectation that performance would decrease with larger contexts is not fully borne out. For the first-order models, the ideal context size seems to be two words on either side of the target. For the second-order models, it is four. This best second-order approach, however, gives results far lower than the least successful first-order wordspace. In the rest of this section we will therefore focus on the performance of the first-order models only. The results of the second-order approaches were invariably inferior and, because of this lack of quality, often hard to interpret.

Table 1 gives the performance of the first-order

| | concrete nouns | | | | | | concrete – abstract | | verbs | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 6 | | 3 | | 2 | | 2 | | 9 | | 5 | |
| | $E$ | $P$ | $E$ | $P$ | $E$ | $P$ | $E$ | $P$ | $E$ | $P$ | $E$ | $P$ |
| 10 | .26 | .73 | .54 | .71 | .97 | .59 | .18 | .97 | .44 | .53 | .41 | .64 |
| 7 | .28 | .73 | **.27** | **.86** | .97 | .57 | **.00** | **1.0** | **.41** | **.56** | **.39** | **.69** |
| 5 | .31 | .71 | .35 | .82 | .95 | .61 | **.00** | **1.0** | **.41** | **.56** | **.39** | **.69** |
| 4 | .28 | .73 | .54 | .71 | .96 | .61 | **.00** | **1.0** | .44 | .51 | **.39** | **.69** |
| 3 | .26 | .77 | .54 | .71 | .97 | .59 | **.00** | **1.0** | .42 | **.56** | .54 | .56 |
| 2 | **.23** | **.82** | .34 | .84 | **.55** | **.86** | **.00** | **1.0** | .48 | .47 | .63 | .56 |
| 1 | .29 | .77 | .50 | .75 | .98 | .57 | **.00** | **1.0** | .42 | .53 | .51 | .60 |

Table 1: Performance of the first-order Word Space Models in the word clustering tasks.

Word Space Models on the three clustering tasks, for each of the pre-specified numbers of clusters. It is hard to pin down an overall best context size: only the smallest and biggest windows under investigation never gave the best results. Let us first discuss the concrete noun clustering task. Here the systems were evaluated at three steps of their output. Their performance clearly deteriorates with each step. With six clusters, the most successful model is that with context size 2. It gives an average entropy of .23 and an average purity of .82. For the three-way clustering task, however, context size 7 unexpectedly gives the best results. We will see why this happens below. At the final evaluation stage, context size 2 is again distinctly in first position, as the only model that manages to come up with a decent clustering.

The division between concrete and abstract nouns, by contrast, is made much more easily. In fact, six out of seven first-order models are able to perfectly retrieve the two classes in the Gold Standard. The model with context size 10 makes a few mistakes here and there, but still finds a reasonable clustering. The verb clustering task, finally, seems to be of average difficulty. In general, intermediate context windows perform best.

### 3.2 Error analysis

Let us now take a closer look at the results. Again we start with the concrete noun subtask. At a first level, the models were required to distinguish between six possible classes. Broadly speaking all models here have the same three difficulties: (1) they are often not able to distinguish between vegetables and fruit, (2) they confuse some of the ground animals with birds, and (3) the tools are scattered among several clusters. Context size 1 makes a separate category for *screwdriver*, *chisel*, *knife* and *scissors*, for instance. The larger context sizes tend to put *spoon*, *bowl*, *cup* and *bottle* in a separate cluster, sometimes together with a number of animals or kinds of fruit. At the later stages, a hard core of artifacts seems to be easily grouped together, but the natural kinds (animals and fruit or vegetables) are still much harder to identify. Here and there a *kitchen* cluster that combines several types of tools, fruit and vegetables might be discerned instead of the Gold Standard grouping, but this is obviously open to interpretation.

The good performance of context size 2 in semantic similarity tasks has been observed before (Sahlgren, 2006). This is no doubt due to the fact that it combines useful information from a number of sources: a noun's adjectives, verbs of which it is the subject, and those of which it fills the object position. This last source of information is often absent from context size 1, at least when the noun is preceded by an article.

With three clusters, we observed that context size 7 suddenly outperforms this seemingly ideal configuration. This actually appears to be a question of chance. The main reason is that with six clusters, the model with context size 7 splits the ground animals and the birds evenly over two clusters. Because of their similarity, these are merged correctly afterwards. Context size 2 gives a far better classification early on, but at the next stage, it recovers less well from its few mistakes than context 7 does. It thus

looks like the high performance of context 7 may partly be an artifact of the data set. Overall, context size 2 still seems the best choice for a classification task of concrete nouns.

Let us now turn to the verb clustering task. At the lowest level, the models were asked to produce nine clusters. The models with intermediate context sizes performed best, although the differences are small. This might be due to the fact that verb clustering benefits from information from a large number of arguments to the verb: subjects and objects as well as prepositional and adverbial phrases. Note that verb classification seems harder than the noun clustering tasks. The boundaries between the classes are indeed more subtle and fuzzy here. Differences, for instance, between *change location* (as in *move*), *motion direction* (as in *arrive*) or *motion manner* (as in *run*) are often too small to discover on a distributional basis.

In this analysis, we regularly mentioned syntactically related words as interesting sources of semantic information. We can therefore expect a model that takes into account these syntactic relations (Padó and Lapata, 2007; Peirsman, 2008) to outperform the simple bag-of-word approaches in these tasks. For the time being, such a model is outside the scope of our investigation, however.

## 4 Task 2: Free Associations

Of course, two words can also be related across semantic classes. *Doctor* is linked to *hospital*, for instance, even though the former refers to a human being and the latter to a building. Similarly, *car* and *drive* are associated, despite the fact they belong to different parts of speech. In this second task, we will try and investigate the degree to which our models are able to capture this type of semantic relatedness, by comparing their nearest neighbours for a target word with the results from a psycholinguistic experiment in which people were asked to give an association for each cue word they were presented with.

Both training and test sets consist of a number of *cue word – association* pairs. All words occurred in at least fifty BNC documents. It was now the general task of our Word Space Models to automatically find the associate for each cue word. This differs considerably from the previous task: whereas word clustering requires the Word Space Models only to consider the words in the test set, now they have to compare the targets with a far larger set of words. We chose to use the 10,000 most frequent words in the BNC as potential associates, including semantically empty words, plus those associates in the test set that did not survive the cut-off at 10,000 words. Even though the words in the training and test set were not tagged for part of speech, our Word Space Models did take these tags into account. Each cue word therefore automatically received its most frequent part of speech in the BNC.

For each of the cue words in the test set, we had the Word Space Models recover the 100 nearest neighbours, in the same way as described in section 2. Since this is an unsupervised approach, we ignored the training set and worked on the test set only. The performance of the models was expressed by the average rank of the association in the list of 100 nearest neighbours to the respective cue word. If the association did not appear in this list, it was automatically given rank 101. Obviously, the lower the score of the model, the better it is able to capture the type of semantic relatedness this task represents.

We also added a different type of algorithm to the experiment. Since we expected syntagmatic relations to play an important role in human associations, we investigated if simple co-occurrence statistics allow us to model the data better than the more advanced Word Space Models. We therefore computed the log-likelihood statistic between each cue word and all potential associates, within a context window of $n$ words to the left and right of the cue. We then simply selected the 100 words with the highest log-likelihood scores.

### 4.1 Results

The results for the investigated models are presented in Figure 2. Because of the cut-off values, the coverage of our models was not always 100%. Context size 10, for instance, fails to come up with nearest neighbours for 7% of the words in the experiment. This is due to a slight inconsistency between our data and the Gold Standard. While we used a lemmatized version of the BNC, the words in the Gold Standard were not always lemmatized to the same base. A good example is *prepared*: in the lemmatized BNC, this is generally reduced to *prepare/VV*,

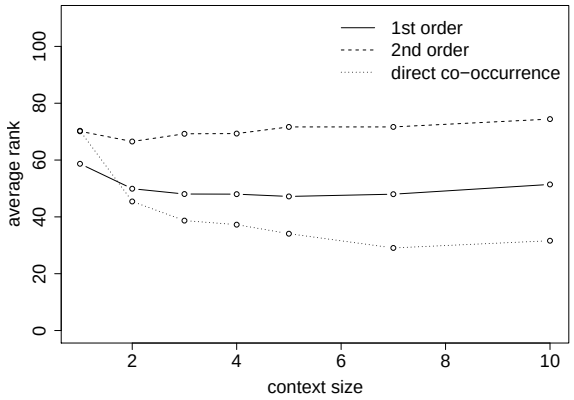



Figure 2: Performance of the Word Space Models in the free association task: average rank of association.

so that *prepared* as an adjective occurs very infrequently. If a cue word was not covered, the example automatically received rank 101.

A Friedman test confirms that there is indeed a statistical influence of the type of model on performance. Interestingly, the direct co-occurrence statistics clearly outperform the Word Space Models. When they take into account seven words to the left and right of the cue, they find the desired association at rank 30, on average. By contrast, the best first-order model (context size 5) only gives this association at a mean rank of 47, and the best second-order model performs even worse, with an average rank of 66.5 for context size 2. Moreover, the performance of the different context sizes seems to contradict our initial research hypothesis, which claimed that tight contexts should score better in the clustering task, while looser context windows should compare more favourably to free association norms. Tests for multiple comparisons after a Friedman test showed significant differences between the three types of models in the association task, but hardly any significant differences between the several context sizes. A detailed error analysis, however, adds some subtlety to this first impression.

### 4.2 Error analysis

To fully appreciate the outcome of the present task, we need to look at the results of the models in more detail. After all, semantically associated words come in many flavours. Some words may be associated to their cue because they are semantically similar, others because they are part of the same conceptual *frame*, still others because they represent typical collocations. This may explain the relatively low average ranks in Figure 2: each model could have its own preference for a specific type of association. It is therefore interesting to have a closer look at the precise associations that are recovered successfully by the different models.

Table 2 compares the results of the first-order model with context size 1 to those of the first-order model with context size 10. For both these models, it shows the twenty cue–association pairs with the highest gain in ranks, as compared to the other model. For instance, with a context size of 1, the associate of *hard* (*soft*) shows up 78 ranks higher than with a context size of 10. This last model, however, was able to recover the associate of *wave* (*sea*) at rank four — the first does not find it.

Interestingly, the nature of the associations for which the models display the highest difference in ranks, varies from one model to the other. The model with context size 1 tends to score comparably well on associations that are semantically similar to their target word. Many are (near-)synonyms, like *rapidly* and *quickly* or *astonishment* and *surprise*, others are antonyms, like *hard* and *soft* or *new* and *old*, while still others are in a IS-A relationship, like *cormorant* and *bird*. The associations for which the larger context window scores far better are generally of a completely different type. Here semantic similarity forms the exception. Most associations are topically related to their target word, either because they belong to the same conceptual *frame*, as with *reflection* and *mirror* or *spend* and *money*, or because they are typical collocates of their target word, like *twentieth* and *century* or *damsel* and *distress*. Of course, no clear line exists between the two categories, since frame-related words will often be collocates of each other.

This contrast is even more outspoken when we compare the first-order model with context size 1 to the best direct co-occurrence model. Among the association pairs recovered by the latter but not by the former are *wizard–oz*, *salvation–army* and *trafalgar–square*. This type of syntagmatic relatedness is indeed seldom modelled by the word spaces.

| strengths of context size 1 | | | | | | strengths of context size 10 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cue | asso | diff | cue | asso | diff | cue | asso | diff | cue | asso | diff |
| melancholy | sad | 100 | glucose | sugar | 63 | sill | window | 100 | damsel | distress | 97 |
| rapidly | quickly | 98 | fund | money | 61 | riding | horse | 100 | leash | dog | 96 |
| plasma | blood | 95 | suspend | hang | 61 | reflection | mirror | 100 | consultant | doctor | 95 |
| astonishment | surprise | 91 | adequate | enough | 54 | nigger | black | 100 | pram | baby | 94 |
| joyful | happy | 83 | levi | jeans | 49 | hoof | horse | 100 | barrel | beer | 94 |
| hard | soft | 78 | sugar | sweet | 46 | holster | gun | 100 | twentieth | century | 91 |
| cormorant | bird | 76 | din | noise | 44 | dump | rubbish | 100 | handler | dog | 90 |
| new | old | 70 | no | yes | 42 | spend | money | 98 | scissors | cut | 80 |
| combat | fight | 69 | tumour | brain | 39 | bidder | auction | 98 | deck | ship | 75 |
| wrath | anger | 64 | weary | tired | 33 | wave | sea | 97 | suicide | death | 72 |

Table 2: Top twenty cue words and associations for which either the first-order model with context size 1 or that with context size 10 scored better than the other.

Finally, when we put the first-order and second-order models with context size 1 side to side, it becomes more difficult to discern a clear pattern. Despite the fact that second-order context words are another way of loosening the definition of context, the second-order model with context size 1 still appears to have a preference for semantic similarity. In fact, word pairs like *companion–friend* and *chat–talk* are better covered here. As Figure 2 suggested, second-order models thus seem to follow the behaviour of the first-order approaches, even though they are consistently less successful.

Our findings so far are confirmed when we look at the parts of speech of the words that are recovered as nearest neighbours to a given cue word. Table 2 showed that for the smallest context window, these nearest neighbours tend to belong to the same part of speech as their cues. This does not hold for the models with larger context sizes. In fact, the table suggests that these sometimes even find nearest neighbours that typically appear as an argument of their cue. Nice examples are *dump–rubbish* or *spend–money*. We therefore calculated for each model the proportion of single nearest neighbours with the same part of speech as their cue. The results are given in Figure 3. It can clearly be seen that, as the context grows larger, the Word Space Models tend to find more neighbours with different parts of speech. For the first-order model with context size 1, 83% of the nearest neighbours have the same part of speech as their cue; for the model with context size 10, this figure has dropped to 58%. The second-order Word Space Models follow the behaviour of the first-order ones here. Not surprisingly, the algorithm that chooses associations on the basis of their log-likelihood score with the target shows the reverse pattern. The larger the co-occurrence span, the higher the chance of finding a word with the same part of speech.

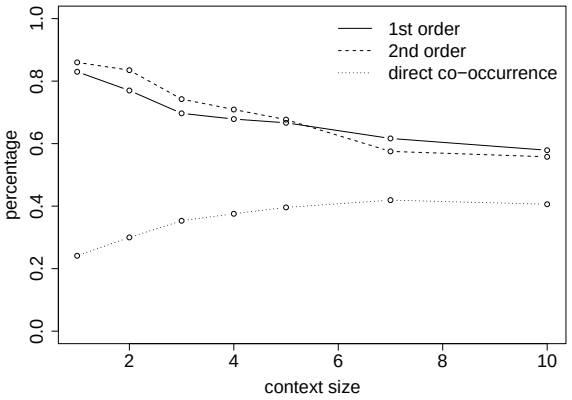



Figure 3: Percentage of nearest neighbours with same tag as their cue word in the free association task.

Overall, our findings demonstrate that human associations are a mixed bag of semantic similarity and topical relatedness. Models with small contexts better recover the former, those with large contexts have a preference for the latter.

## 5 Conclusions and future work

Word Space Models have become an indispensible tool in many computational-linguistic applications. Yet, the NLP community is only slowly gaining insight in the type of semantic information that gets

modelled by these approaches, and how this information is influenced by the way the models operationalize the vague notion of context. While it has previously been shown that first-order bag-of-word models with small context sizes tend to best capture semantic similarity (Sahlgren, 2006), this paper is the first to compare two ways of loosening this context definition. In particular, we contrasted larger first-order context windows with second-order context models, which model the meaning of a word in terms of the context words of its context words, and evaluated them through two series of experiments.

Our findings can now be summarized as follows. (1) Overall, second-order bag-of-word models are inferior to their first-order competitors. Switching to second-order co-occurrence moreover does not lead to an increased preference for syntagmatic relations. (2) With respect to semantic similarity, a context window of size 2 gives the best results for noun clustering. For verbs, the context is better stretched to 4-7 words to the left and right of the target word. (3) Even though there is only a minor impact of context size on the overall performance in the free association task, small contexts display a preference for semantic similarity, while large contexts model syntagmatic relations better. However, the Word Space Models here are clearly outperformed by direct co-occurrence statistics.

Obviously, the Word Space Models under investigation allow for much more variation than we have been able to explore here. Syntactic models, for instance, certainly deserve further investigation, as in our papers on Dutch (Peirsman, 2008). Moreover, the question still remains why the second-order contexts, despite their poor performance generally, did score extremely well on a number of examples. Is this coincidental, or could there be a pattern to this set of cases? Either way, the intriguing variation across the results from the different Word Space Models justifies further research in the precise relationship between distributional and semantic relatedness.

## References


Marco Baroni, Alessandro Lenci, and Luca Onnis. 2007. ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the ACL 2007 Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 49–56, East Stroudsburg, PA.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st International Conference on Association for Computational Linguistics (ACL-1993)*, pages 164–171, Columbus, OH.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1994. Discrimination decisions for 100,000-dimensional spaces. In A. Zampoli, N. Calzolari, and M. Palmer, editors, *Current Issues in Computational Linguistics: In Honour of Don Walker*, pages 429–450. Kluwer Academic Publishers.

George Karypis. 2003. CLUTO. A clustering toolkit. Technical Report 02-017, University of Minnesota, Department of Computer Science, Minneapolis, MN.

Thomas. K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Joseph P. Levy and John A. Bullinaria. 2001. Learning lexical properties from word usage patterns: Which context words should be used. In R.F. French and J.P. Sougne, editors, *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 273–282. London: Springer.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *Proceedings of the CoSMO workshop, held in conjunction with CONTEXT-07*, pages 9–16, Roskilde, Denmark.

Yves Peirsman. 2008. Word space models of semantic similarity and relatedness. In *Proceedings of the 13th ESSLLI Student Session*, pages 143–152, Hamburg, Germany.

Magnus Sahlgren. 2006. *The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

# Performance of HAL-like word space models on semantic clustering.

**Cyrus Shaoul**
Department of Psychology
University of Alberta
Edmonton, AB, Canada, T6G2E9
cyrus.shaoul@ualberta.ca

**Chris Westbury**
Department of Psychology
University of Alberta
Edmonton, AB, Canada, T6G2E9
chrisw@ualberta.ca

## Abstract

A recent implementation of a HAL-like word space model called HiDEx was used to create vector representations of nouns and verbs. As proposed by the organizers of the Lexical Semantics Workshop (part of ESSLLI 2008), these vectors were analyzed to see if they could predict behavioral data from lexical categorization experiments and feature generation experiments. HiDEx performed well on the Abstract/Concrete Noun discrimination task, but poorly on the other tasks. There is much work to be done before HiDEx can accurately categorize nouns and verbs or generate lexical features.

## 1 Introduction

HAL (Hyperspace Analog of Language) is a computational model of semantic memory (Lund & Burgess, 1996). HAL-like models have been shown to be capable of discovering semantic categories of lexical items (Burgess & Lund, 1997). For our contribution to the Lexical Semantics Workshop at ESSLLI 2008, we used an implementation of HAL (Shaoul & Westbury, 2006) to see if the vectors in a HAL-like word space model could accurately categorize words into different semantic categories.

HAL collects global co-occurrence statistics for words used in a corpus, and then analyzes the geometric relationships between words in this co-occurrence space by measuring the Euclidean distance between words. This type of model has been shown to accurately categorize English words by parts of speech, concrete noun categories and abstract noun categories when clustering was done using multi-dimensional scaling (Burgess & Lund, 2000). In this work we will investigate the capabilities of HAL-like models to categorize nouns and verbs in three tasks: Concrete Noun Clustering, Abstract/Concrete Noun Discrimination, and Verb Clustering. We also looked at the ability of HAL-like models to generate lexical features. Finally, we will see if the parameter settings used by HAL-like models contribute to the performance of the model on these tasks.

## 2 Model Performance

### 2.1 Methods

We used HiDEx (Shaoul & Westbury, 2006) to create a global co-occurrence matrix for 48,000 English words. The corpus used in this experiment was a 1 billion-word subset of the freely available Westbury Lab USENET corpus (Shaoul & Westbury, 2007). Two sets of parameters were used, the original HAL parameters (Lund & Burgess, 1998), and the parameters proposed by Shaoul & Westbury (in preparation) that had greater correlations with lexical and semantic decision reaction times (RT). The original HAL parameters were the following: the local co-occurrence window was 10 words ahead and 10 words behind, and the weighting scheme was a triangular one known as the linear ramp. This weighting scheme gives the most weight to words that co-occur near the target word.

The optimized parameters proposed by Shaoul & Westbury (in preparation) include a local co-occurrence window of 5 words ahead and 10 words behind, and a weighting scheme that is the inverse of the original HAL scheme, an inverse linear ramp. This weighting scheme gives the most weight to words that co-occur farther from the target word. Bullinaria and Levy (2007) and others have found optimized parameter sets for HAL-like models. Their optimizations differ from those used here as the tasks used to measure the model's fitness were not RT, but rather TOEFL scores and other measures.

During the creation of the global co-occurrence matrix, the vectors from the 10,000

most frequent words were retained from the full lexicon of 47,000 words. Rohde et al (submitted) and Bullinaria and Levy (2007), found that matrices of this size or greater produced good results. With vectors for the both the forward and backward windows, the actual number of columns in the global co-occurrence matrix is doubled, to 20,000.

## 2.2 Results

All cluster analysis was performed using the CLUTO clustering package (Zhao & Karypis, 2002) with the "Repeated Bisections by k-way refinement" clustering method. In all the analyses, the cluster *entropy* and *purity* values that are reported were calculated by CLUTO. Zhao and Karypis (2001) define cluster *entropy* as a measure of how much uncertainty there was in the clustering solution. A value of 0 indicates a no uncertainty, and larger values indicates increasing uncertainty. Cluster *purity* is defined as the degree to which the clusters contain items that are not in the maximally included target category. A value of 1 indicates a perfect match, with lower scores indicating more intrusions from other categories.

### 2.2.1 Concrete Noun Clustering

In this task, 44 concrete nouns from the nouns listed in McRae et al. (2005) were clustered into 6 categories. The same words were then clustered into more general 2 and 3 category arrangements. We used the vectors produced by HiDEx to cluster the words, and then compared the quality of our clustering solutions for both the Original HAL and Optimized HAL parameter sets.

The cluster measure results are shown in Table 1. The cluster entropy produced by our of our word space models shrank as the number of clusters increased, and the purity decreased in parallel. Since the best clustering solution has a high purity and low entropy, there appear to be no general trends towards better fits in this data. The two different parameter sets produced highly similar, poor results.

The confusion matrix for the six-way solution is shown in Table 2. Although most (4/6) noun categories were placed in two or fewer clusters, this is largely because two 'super-clusters' (Clusters 4 and 5) contained most (37/44) of the words.

| Original HAL | Cluster Entropy | Cluster Purity |
|---|---|---|
| 2-way | 0.931 | 0.545 |
| 3-way | 0.844 | 0.523 |
| 6-way | 0.770 | 0.409 |
| | | |
| **Optimized HAL** | **Cluster Entropy** | **Cluster Purity** |
| 2-way | 0.981 | 0.545 |
| 3-way | 0.869 | 0.523 |
| 6-way | 0.719 | 0.386 |

Table 1: Results for the 2-, 3- and 6-way clustering solutions for two HAL parameter sets in the Concrete Noun Clustering task.

| Cluster | Bird | Tree | Green veg | Ground animal | Tool | Vehicle |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 3 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4 | 2 | 3 | 5 | 3 | 6 | 6 |
| 5 | 2 | 0 | 0 | 5 | 5 | 0 |

Table 2: Confusion Matrix for the 6-way concrete noun clustering by the Optimized HAL model.

One pattern of errors became clear from observation: in the 2-, 3- and 6-way clustering solutions, the words EAGLE, OWL and PENGUIN were consistently clustered together in a separate cluster. It is unclear why the other "bird" items did not join these three in their cluster.

See the Appendix for a full report of the clustering solutions.

### 2.2.2 Abstract/Concrete Noun Discrimination

In this task, 40 nouns were clustered into two groups. One third of the nouns were rated as highly imageable using the MRC imageability norms (Wilson, 1988), one third had very low imageability ratings, and one third had intermediate imageability ratings.

The cluster measure results are shown in Table 3. The clusters produced from the HiDEx measures closely matched the imageability groups that were produced from human ratings. The Optimized HAL parameters produced a 2-way clustering that was lower in entropy and greater in purity than the Original HAL parameters.

| Original HAL | Cluster Entropy | Cluster Purity |
|---|---|---|
| 2-way | 0.84 | 0.6 |
| **Optimized HAL** | **Cluster Entropy** | **Cluster Purity** |
| 2-way | 0.647 | 0.725 |
| *Difference between parameter sets* | *0.193* | *-0.125* |

Table 3: The 2-way clustering solution for the 40 words in the Abstract/Concrete Discrimination data set.

Table 4 shows the categorizations of the words with intermediary imageability in the 2-way clustering solution, with category labels added for the clusters that clearly contained abstract or concrete words. There was no discernable pattern to this categorization of intermediate imageability words.

| Intermediate Imageability Word | Categorization |
|---|---|
| CEREMONY | HI |
| EMPIRE | HI |
| FIGHT | LO |
| FOUNDATION | HI |
| INVITATION | LO |
| POLLUTION | LO |
| SHAPE | HI |
| SMELL | HI |
| WEATHER | HI |

Table 4: Categorizations of words with intermediate imageability.

### 2.2.3 Verb Clustering

In this task, the challenge was to cluster 44 verbs into either 5 or 9 different semantic categories.

Table 4 shows the results for the 5- and 9-way clustering solutions. Neither the Original HAL nor the Optimized HAL produced clustering solutions of very good purity or entropy.

There was no obvious qualitative explanation for the cluster pattern of the verbs by our model in this task. One speculative is that there was an effect of verb polysemy on their representation in the HAL word space. The clusters that our model created three single-word clusters that are qualitatively suggestive of this: the words LEND, PAY and EVALUATE made up their own clusters in the 5-way categorization, with the rest of the verbs filling up the remaining two clusters.

The full confusion matrix for the 5-way solution is shown in Table 6. As with the noun clusters, most (40/44) of the words were categorized into two super-clusters that included words from all (Cluster 3) or almost all (Cluster 4) of the verb types.

| Original HAL | Cluster Entropy | Cluster Purity |
|---|---|---|
| 5-way | 0.816 | 0.4 |
| 9-way | 0.572 | 0.422 |
| **Optimized HAL** | **Cluster Entropy** | **Cluster Purity** |
| 5-way | 0.715 | 0.511 |
| 9-way | 0.709 | 0.333 |

Table 5: Scores for the 5- and 9-way clustering solutions for verb categories.

In the 9-way solution, there were 6 one-word clusters (the previous three words, plus BREATHE, ACQUIRE, CRY) and one two-word clusters (SMELL and SMILE). There is a possibility that the clustering algorithms used in CLUTO may not be well suited to this data, as all the available clustering methods produced approximately the same results. It is more likely that the clusters produced by our model reflect a non-intuitive vector similarity that is being used by the clustering method to create these clusters.

| Cluster | Body | State | Cog | Exch | Motion |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 | 1 |
| 3 | 4 | 5 | 4 | 3 | 9 |
| 4 | 5 | 0 | 5 | 2 | 4 |

Table 6: Confusion Matrix for the 5-way verb clustering by the Optimized HAL model.

### 2.2.4 Property Generation

The property generation task required the word space model to predict which properties human subjects would produce with the greatest frequency when asked to give properties of a

word. This task was built on the feature norms collected by McRea et. al (2005). For a word such as DUCK, subjects provided features such HAS FEET, SWIMS, and LIVES ON WATER. For this task, the target properties were converted into an expanded set of terms using the synset information from WordNet (Fellbaum, 1998). DUCK should produce the words FEET, SWIMS, and WATER, among others.

To apply our model to this task, we used HiDEx to generate the 200 closest neighbors to each of the words in the word list in our word space (the number 200 was chosen arbitrarily). We then used the evaluation script provided by the workshop organizers to calculate the precision of the match between our neighbors and the properties generated by participants. The results of this precision calculation are listed in Table 7.

Both the Original and Optimized HAL parameter sets produced very low precision scores (below 2%) in this task. The low precision was due to the fact that HAL neighbors are more often words with very similar contexts to the target word. Feature names are not often found in the same contexts as the words that they describe.

For example, the word DUCK has the features FLIES and WEB-FOOTED. These features have low "substitutability" for the word DUCK, and therefore would likely not appear in the list of neighbors in a HAL-like word space.

| **Original HAL** | **Average Precision** | **Std. Dev of Precision** |
|---|---|---|
| 10 best | 0.018 | 0.039 |
| 20 best | 0.013 | 0.022 |
| 30 best | 0.014 | 0.019 |

| **Optimized HAL** | **Average Precision** | **Std. Dev of Precision** |
|---|---|---|
| 10 best | 0.018 | 0.039 |
| 20 best | 0.016 | 0.023 |
| 30 best | 0.012 | 0.019 |

Table 7: Precision of lexical feature generation for 44 concrete nouns.

## 3 Conclusion

Using a HAL-like word space model, we attempted to algorithmically cluster English nouns and verbs, given pre-defined semantic categories. Our model performed well on the Abstract/Concrete Noun Discrimination task, but poorly on the Noun Clustering and Verb Clustering tasks. It is unclear to us why there was large variability in the performance of our model on these tasks, and why it performs relatively well on the Abstract/Concrete noun discrimination task.

We also used out word space model to on the more open-ended tasks of generating features for concrete nouns, but found that our model could did not produce many feature names in the list of word space neighbors. This is not surprising given the nature of the model, since features are not often substitutable for the concept with which they are associated.

These results show the strengths and weaknesses our HAL-like models. They also illuminate the pathway to future improvements of this type of model. We see two potential avenues for progress: changes to the core model and change to the usage of the model.

In terms of the core model, we used model parameters that were optimized to produce high correlations with particular behavioral measures: lexical decision reaction time and semantic decision reaction time. There might be a very different set of model parameters that would be the optimal ones for semantic clustering. One straightforward way of testing this would be to test many model parameter combinations and find which one produced the highest scores on the current clustering task set. There is even a possibility that the parameter settings that create the best clustering for nouns would be very different from those that perform best for verbs, and these settings my lower the score on the Abstract/Concrete Discrimination task. A true lack of generalization of the model may be an indicator of the underlying psychological complexity of lexical semantic organization.

Other changes could be made to the core model, including the algorithms used in HAL, such as the distance metric, and methods of vector normalization. Rohde et al (submitted) propose changes to normalization methods and metric calculations that greatly improve performance of a HAL-like model on many different semantic tasks. These types of modifications to the HAL model can influence the content of the word space vectors, but they do not change the fact that the HAL model has limitations in what kind of measures it can produce. The core capability of the HAL model is the construction of co-occurrence vectors, the calculation of inter-word distance measures, the calculation of word space neighborhoods, and the measurement of

the density of these neighborhoods (Shaoul & Westbury, 2006). To be able to perform the feature generation task, the HAL model will need to be used in a novel way.

How to change the usage of the HAL model? One proposal is to look at the neighbors of a word's neighbors. A "small world" network model might help a graph-based clustering algorithm find greater degree of similarity for the neighbors of a word like DUCK when compared to the neighbors of a word like OWL. With the addition of the neighbor's neighbors, the categorization might converge more quickly and accurately that with the raw vectors alone.

Another possibility is to use the category names as anchors, and find the distance between the targets and these anchors to test for category membership. These types of alternative uses of the information in the HAL model should be informed by current psychological theory where possible. They may also provide new perspectives on current debates about language and lexical memory. We hope to continue this kind of research in the future and make contributions to the understanding of lexical semantics and word space models.

## Appendix

All the vectors that were used in our clustering analyses and the clustering solutions are available for download at:

www.ualberta.ca/~westburylab/ESSLLI2008

## References


Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39, 510–52

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. Language & Cognitive Processes, 12, 177–21

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117–156). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. MIT Press. MA, USA

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28*, 203–20

McRae, K., Cree, G.S., Seidenberg, M.S., and McNorgan C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers, 37,* 547–559.

Rohde, D. L. T., Gonnerman, L., and Plaut, D. C. (submitted). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods, 38*, 190-19

Shaoul, C., & Westbury, C. (2007). *A usenet corpus* (2005-2008) Edmonton, AB: University of Alberta. Downloaded from http://www.psych.ualberta.ca/westbury-lab/downloads/usenetcorpus.download.html.

Shaoul, C., & Westbury, C. (In prep). Using a high dimensional model to predict Lexical Decision and Semantic Decision Reaction Time.

Wilson, M.D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20(1)*, 6-11.

Zhao, Y. & Karypis, G (2001). Criterion functions for document clustering: Experiments and analysis. Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN. Available on the WWW at http://cs.umn.edu/~karypis/publications.

Zhao, Y. & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In Proceedings of the eleventh international Conference on Information and Knowledge Management. ACM, NY, NY, USA.

# A Comparison of Bag of Words and Syntax-based Approaches for Word Categorization

**Tim Van de Cruys**
Humanities Computing
University of Groningen
t.van.de.cruys@rug.nl

## Abstract

This paper will examine the aptness of various word space models for the task of word categorization, as defined by the lexical semantics workshop at ESSLLI 2008. Three word clustering tasks will be examined: concrete noun categorization, concrete/abstract noun discrimination, and verb categorization. The main focus will be on the difference between bag of words models and syntax-based models. Both approaches will be evaluated with regard to the three tasks, and differences between the clustering solutions will be pointed out.

## 1 Introduction

For quite some years now, word space models are a popular tool for the automatic acquisition of semantics from text. In word space models, a particular word is defined by the context surrounding it. By defining a particular word (i.e. its context features) in a vector space, the word can be compared to other words, and similarity can be calculated.

With regard to the context used, two basic approaches exist. One approach makes use of 'bag of words' co-occurrence data; in this approach, a certain window around a word is used for gathering co-occurrence information. The window may either be a fixed number of words, or the paragraph or document that a word appears in. Thus, words are considered similar if they appear in similar windows (documents). One of the dominant methods using this method is LATENT SEMANTIC ANALYSIS (LSA).

The second approach uses a more fine grained distributional model, focusing on the syntactic relations that words appear with. Typically, a large text corpus is parsed, and dependency triples are extracted.[1] Words are considered similar if they appear with similar syntactic relations. Note that the former approach does not need any kind of linguistic annotation, whereas for the latter, some form of syntactic annotation is needed.

The results yielded by both approaches are typically quite different in nature: the former approach typically puts its finger on a broad, thematic kind of similarity, while the latter approach typically grasps a tighter, synonym-like similarity. Example (1) shows the difference between both approaches; for each approach, the top ten most similar nouns to the noun *muziek* 'music' are given. In (a), the window-based approach is used, while (b) uses the syntax-based approach. (a) shows indeed more thematic similarity, whereas (b) shows tighter similarity.

(1) a. **muziek** 'music': *gitaar* 'guitar', *jazz* 'jazz', *cd* 'cd', *rock* 'rock', *bas* 'bass', *song* 'song', *muzikant* 'musician', *musicus* 'musician', *drum* 'drum', *slagwerker* 'drummer'

b. **muziek** 'music': *dans* 'dance', *kunst* 'art', *klank* 'sound', *liedje* 'song', *geluid* 'sound', *poëzie* 'poetry', *literatuur* 'literature', *popmuziek* 'pop music', *lied* 'song', *melodie* 'melody'

This paper will provide results for the categorization tasks that have been defined for the lexical semantics workshop 'Bridging the gap between

[1] e.g. dependency relations that qualify *apple* might be 'object of *eat*' and 'adjective *red*'. This gives us dependency triples like $< apple, obj, eat >$.

semantic theory and computational simulations' at ESSLLI 2008.[2] The workshop provides three different categorization (clustering) tasks:

- concrete noun categorization
- abstract/concrete noun discrimination:
- verb categorization

The three tasks will be carried out according to the two approaches described above. In the evaluation of the various tasks, we will try to determine whether the difference between the 'bag of words' approach and the syntactic approach is responsible for different clustering outputs.

## 2 Methodology

### 2.1 General remarks

The research has been carried out for Dutch, mainly because this enabled us to use the Alpino parser (van Noord, 2006), a dependency parser for Dutch. The test sets that were provided have been translated into Dutch by three translators, and – when multiple translations were found – the majority translation has been taken as the final one. The frequencies of the Dutch words are by and large comparable to the frequencies of their English counterparts. High frequent words (*dog*, *cat*) and low-frequent ones (*snail*, *turtle*) in one language generally have the same order of magnitude in the other, although exceptions occur (*eagle*). Table 1 compares the frequencies of words of the *animal* class in the British National Corpus and the Twente Nieuws Corpus. The results for the other words are similar.

All data has been extracted from the TWENTE NIEUWS CORPUS (Ordelman, 2002), a corpus of $\pm$ 500M words of Dutch newspaper text. The corpus is consistently divided into paragraphs, constituting our window for the bag of words approach. The whole corpus has been parsed with the Alpino parser, and dependency triples have been extracted.

The clustering solutions have been computed with the clustering program CLUTO (Karypis, 2003), using the 'rbr' option as clustering algorithm (this is an algorithm that repeatedly bisects the matrix until the desired numbers of clusters is reached; application of this algorithm was prescribed by the workshop task description).

### 2.2 Bag of words

For the bag of words approach, matrices have been constructed that contain co-occurrence frequencies of nouns (verbs) together with the most frequent words of the corpus in a context window. As a context window, we selected the paragraphs of the newspaper. The resulting matrix has been adapted with POINTWISE MUTUAL INFORMATION (PMI) (Church and Hanks, 1990).

The final test matrix has been constructed in two different ways:

1. a small matrix is extracted, containing only the frequencies of the words in the test set. The output is a matrix of e.g. 45 nouns by 2K co-occurring words;
2. a large matrix is extracted, containing the frequencies of a large number of words (including the test words). The output is a matrix of e.g. 10K nouns by 2K co-occurring words. After applying PMI, the test words are extracted from the large matrix.

The choice of method has a considerable impact on the final matrix, as the results of the PMI computation are rather different. In the first case, only the test words are taken into account to normalize the features; in the second case, the features are normalized with regard to a large set of words in the corpus. The difference will lead to different clustering results. The first method will be coined LOCAL PMI (LOCPMI), the second GLOBAL PMI (GLOPMI).

We have experimented with two kinds of dimensionality reduction: LATENT SEMANTIC ANALYSIS (LSA, Landauer et al. (1997; 1998)), in which a SINGULAR VALUE DECOMPOSITION (SVD) is computed of the original co-occurrence frequency matrix[3], and NON-NEGATIVE MATRIX FACTORIZATION (Lee and Seung, 2000), in which a factorization of the original frequency is calculated by minimizing Kullback-Leibler divergence between the

[2] http://wordspace.collocations.de/doku.php/esslli:start

[3] The original method of LSA uses the frequency of words by documents as input; we used frequencies of words by co-occurring words in a context window.

| NOUN.ENG | FREQ.BNC | LOGFREQ.BNC | NOUN.DU | FREQ.TWNC | LOGFREQ.TWNC |
|---|---|---|---|---|---|
| chicken | 2579 | 7.86 | kip | 7663 | 8.94 |
| eagle | 1793 | 7.49 | arend | 113 | 4.72 |
| duck | 2094 | 7.65 | eend | 3245 | 8.08 |
| swan | 1431 | 7.27 | zwaan | 1092 | 7.00 |
| owl | 1648 | 7.41 | uil | 559 | 6.33 |
| penguin | 600 | 6.40 | pinguïn | 146 | 4.98 |
| peacock | 578 | 6.36 | pauw | 221 | 5.40 |
| dog | 12536 | 9.44 | hond | 17651 | 9.77 |
| elephant | 1508 | 7.32 | olifant | 2708 | 7.90 |
| cow | 2611 | 7.87 | koe | 9976 | 9.21 |
| cat | 5540 | 8.62 | kat | 5822 | 8.67 |
| lion | 2155 | 7.68 | leeuw | 2055 | 7.63 |
| pig | 2508 | 7.83 | varken | 5817 | 8.67 |
| snail | 543 | 6.30 | slak | 712 | 6.56 |
| turtle | 447 | 6.10 | schildpad | 498 | 6.21 |

Table 1: The frequencies of English words in the BNC vs. the frequencies of Dutch words in the TWNC

original matrix and its factorization according to the constraint that all values have to be non-negative. But since the dimensionality reduction models did not bring about any improvement over the simple bag of word models, dimensionality reduction models have not been included in the evaluation.

### 2.3 Syntax-based

The syntax-based approach makes use of matrices that contain the co-occurrence frequencies of nouns (verbs) by their dependencies. Typically, the feature space with the syntax-based method is much larger than with simple co-occurrences, but also much sparser. The resulting matrix is again adapted with PMI.

Again, the matrix can be constructed in two different ways:

1. a small matrix, containing only the frequencies of the test words by the dependencies with which the word occurs. The output is a matrix of e.g. 45 nouns by 100K dependencies;

2. a large matrix, containing the frequencies of a large number of words (including the test words). The output is e.g. a matrix of 10K nouns by 100K dependencies. The final test words are extracted afterwards.

The choice of method again has a large impact on the final matrix with regard to PMI.

### 2.4 Evaluation measures

There are two external evaluation measures available in CLUTO – ENTROPY and PURITY – which have been chosen as evaluation measures for the workshop task. Entropy measures how the various semantic classes are distributed within each cluster, and purity measures the extent to which each cluster contains words from primarily one class (Zhao and Karypis, 2001). Both measures run from 0 to 1. Low entropy measures and high purity values indicate a successful clustering.

## 3 Results & Evaluation

### 3.1 Concrete noun categorization

#### 3.1.1 Introduction

In the concrete noun categorization task, the goal is to cluster 44 concrete nouns in a number of classes on various levels of generality:

- 2-way clustering: cluster nouns in two top classes *natural* and *artefact*;

- 3-way clustering: cluster nouns in three classes *animal*, *vegetable* and *artefact*;

- 6-way clustering: cluster nouns in six classes *bird*, *groundAnimal*, *fruitTree*, *green*, *tool* and *vehicle*.

In the next sections, we will evaluate how bag of words models and syntactic models are coping with this clustering task, and compare both methods.

#### 3.1.2 Bag of words

Table 2 gives the clustering results of the bag of words methods for different clustering sizes.

| method | n-way | entropy | purity |
|---|---|---|---|
| LOCPMI | 2 | .930 | .614 |
| | 3 | .489 | .750 |
| | 6 | .339 | .636 |
| GLOPMI | 2 | .983 | .545 |
| | 3 | .539 | .705 |
| | 6 | .334 | .682 |

Table 2: A comparison of different clustering results for concrete noun categorization — bag of words approach

None of the bag of words models is particularly good at noun categorization: the LOCPMI and GLOPMI have similar results. The results do show that bag of word models are better in categorizing on a more specific level: the more specific the clustering, the better the scores are.

Figure 1 shows the confusion matrix for the GLOPMI 6-way clustering.

| cluster | bird | grou | frui | gree | tool | vehi |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 2 |
| 2 | 1 | 0 | 4 | 5 | 2 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 5 |
| 4 | 0 | 0 | 0 | 0 | 3 | 0 |
| 5 | 6 | 8 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 7 | 0 |

Figure 1: Confusion matrix

The clusters found by the algorithm are still quite sensible; cluster 1 for example looks like this:

- *aardappel* ‘potatoe’, *ananas* ‘pineapple’, *banaan* ‘banana’, *champignon* ‘mushroom’, *kers* ‘cherry’, *kip* ‘chicken’, *kom* ‘bowl’, *lepel* ‘spoon’, *maïs* ‘corn’, *peer* ‘pear’, *sla* ‘lettuce’ *ui* ‘oignon’

Clearly, the algorithm has found a food-related cluster, with fruits, vegetables, a meat term (‘chicken’) and kitchen tools (‘bowl’, ‘spoon’).

The two- and three-way clusterings of the bag of words models are less sensible.

#### 3.1.3 Syntax-based

Table 3 gives the clustering results for the syntax-based algorithms for different clustering sizes.

| method | n-way | entropy | purity |
|---|---|---|---|
| LOCPMI | 2 | .939 | .636 |
| | 3 | .344 | .818 |
| | 6 | .118 | .886 |
| GLOPMI | 2 | .000 | 1.000 |
| | 3 | .000 | 1.000 |
| | 6 | .173 | .841 |

Table 3: A comparison of different clustering results for concrete noun categorization — syntactic approach

LOCPMI scores the best result with regard to the most specific (6-way) clustering, but only slightly better than GLOPMI. When the clustering task becomes more abstract, GLOPMI clearly outperforms the local model: the 2-way and 3-way clusterings are optimal in the global model, whereas the local models score worse results with increasing abstractness.

Figure 2 shows the confusion matrix for the best-performing 6-way clustering (LOCPMI). The results of the global model are quite similar.

| cluster | bird | grou | frui | gree | tool | vehi |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 4 | 5 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 7 | 0 |
| 3 | 6 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 7 |
| 5 | 0 | 0 | 0 | 0 | 6 | 0 |
| 6 | 1 | 8 | 0 | 0 | 0 | 0 |

Figure 2: Confusion matrix

Upon examining the results, the decisions made by the algorithm look quite reasonable:

- One bird ('chicken') is classified as *grounAnimal*;
- fruits and vegetables are assigned to one single cluster;
- the *tools* class is split up into two different clusters, so that a division is made between 'active' and 'passive' tools:
  - *beitel* 'chisel', *hamer* 'hammer', *mes* 'knife', *pen* 'pen', *potlood* 'pencil', *schaar* 'scissors', *schroevendraaier* 'screwdriver';
  - *beker* 'cup', *fles* 'bottle', *ketel* 'kettle', *kom*, 'bowl', *lepel* 'spoon', *telefoon* 'telephone'.

It is interesting to note that the difference between fruit and vegetables nonetheless is present in the data. When clustering the words from the subsets *fruitTree* and *green* into two classes, they are properly split up:

- *kers* 'cherry', *banaan* 'banana', *peer* 'pear', *ananas* 'pineapple';
- *champignon* 'mushroom', *maïs* 'corn', *sla* 'lettuce', *aardappel* 'potatoe', ui 'oignon'.

#### 3.1.4 Comparison of both approaches

Globally, the syntax-based approach seems more apt for concrete noun clustering. Both approaches have similar results for the most specific classification (6-way clustering), but the syntax-based approach performs a lot better on a more abstract level. The conclusion might be that the bag of words approach is able to cluster nouns into 'topics' (cfr. the cluster containing words that relate to the topic 'food'), but has difficulties generalizing beyond these topics. The syntax-based approach, on the other hand, is able to generalize beyond the topics, discovering features such as 'agentness' and 'naturalness', allowing the words to be clustered in more general, top-level categories.

### 3.2 Abstract/Concrete Noun Discrimination

#### 3.2.1 Introduction

The evaluation of algorithms discriminating between abstract and concrete nouns consists of three parts:

- In the first part, 30 nouns (15 with high concreteness value and 15 with low concreteness value) are clustered in two clusters, HI and LO;
- in the second part, 10 nouns with average concreteness value are added to the two-way clustering, to see whether they end up in the HI or the LO cluster;
- in the third part, a three-way clustering of the 40 nouns (15 HI, 10 ME, 15 LO) is performed.

In the next sections, both bag of word models and syntax-based models are again evaluated with regard to these parts.

#### 3.2.2 Bag of words

Table 4 gives the clustering results of the bag of words methods for different clustering sizes.

| method | part | entropy | purity |
|---|---|---|---|
| LOCPMI | part 1 | .470 | .867 |
| | part 3 | .505 | .750 |
| GLOPMI | part 1 | .000 | 1.000 |
| | part 3 | .605 | .700 |

Table 4: A comparison of different clustering results for abstract/concrete noun discrimination — bag of words approach

The GLOPMI model outperforms the LOCPMI in the discrimination of abstract and concrete nouns (part 1). The LOCPMI scores a bit better in discriminating the ME nouns (part 3).

Interestingly enough, the result of part 2 is for both LOCPMI and GLOPMI the same:

- *geur* 'smell', *vervuiling* 'pollution' and *weer* 'weather' are assigned to the HI cluster;
- *uitnodiging* 'invitation', *vorm* 'shape', *rijk* 'empire', *fundament* 'foundation', *ruzie* 'fight', *pijn* 'ache' and *ceremonie* 'ceremony' are assigned to the LO cluster.

#### 3.2.3 Syntax-based

Table 5 gives the clustering results of the syntax-based methods for different clustering sizes.

The local PMI method gets the best results: The 2-way clustering as well as the 3-way clustering are accurately carried out.

| method | part | entropy | purity |
|---|---|---|---|
| LOCPMI | part 1 | .000 | 1.000 |
| | part 3 | .000 | 1.000 |
| GLOPMI | part 1 | .000 | 1.000 |
| | part 3 | .367 | .750 |

Table 5: A comparison of different clustering results for abstract/concrete noun discrimination — syntactic approach

Part 2 gives different results for the LOCPMI and GLOPMI method:

- The local method classifies *rijk* 'empire' as HI and the other 9 words as LO;
- the global method classifies *weer* 'weather', *uitnodiging* 'invitation', *ceremonie* 'ceremony' as HI and the other 7 words as LO.

#### 3.2.4 Comparison of both approaches

The syntax-based approach outperforms the bag of words approach, although the bag of words approach is also able to make an accurate distinction between concrete and abstract nouns with the GLOPMI model. Again, the explanation might be that the syntax-based method is able to discover general features of the nouns more easily. Nevertheless, the results show that discrimination between concrete and abstract nouns is possible with the bag of words approach as well as the syntax-based approach.

### 3.3 Verb categorization

#### 3.3.1 Introduction

The goal of the verb categorization task is to cluster 45 verbs into a number of classes, both on a more general and a more specific level:

- 5-way clustering: cluster the verbs into 5 more general verb classes: *cognition*, *motion*, *body*, *exchange* and *changeState*;
- 9-way clustering: cluster the verbs into 9 fine-grained verb classes: *communication*, *mentalState*, *motionManner*, *motionDirection*, *changeLocation*, *bodySense*, *bodyAction*, *exchange* and *changeState*.

#### 3.3.2 Bag of words

Table 6 gives the clustering results of the bag of words methods for different clustering sizes.

| method | n-way | entropy | purity |
|---|---|---|---|
| LOCPMI | 5 | .478 | .622 |
| | 9 | .419 | .578 |
| GLOPMI | 5 | .463 | .600 |
| | 9 | .442 | .556 |

Table 6: A comparison of different clustering results for verb categorization — bag of words approach

There are no large differences between the local and global PMI method: both methods score about the same. The more specific classification (9-way clustering) scores slightly better, but the differences are small.

Figure 3 shows the confusion matrix for the best-performing 5-way clustering (GLOPMI). The results of the local model are again similar.

| cluster | cogn | moti | body | exch | chan |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 5 | 1 |
| 2 | 5 | 1 | 4 | 0 | 0 |
| 3 | 5 | 2 | 0 | 0 | 0 |
| 4 | 0 | 7 | 5 | 0 | 0 |
| 5 | 0 | 5 | 0 | 0 | 4 |

Figure 3: Confusion matrix

The first cluster mainly contains exchange verbs. The second cluster is a combination of cognition and body verbs. It is interesting to note that the body verbs with a particular emotional connotation ('cry', 'smile', also 'listen' and 'feel') end up in a cluster together with the cognition verbs. The body verbs without an emotional connotation ('breathe', 'drink', 'eat', also 'smell') end up in a cluster together with (body) movements (cluster 4). Cluster 3 seems a business-related cluster, given the fact that *bestuur* is an ambiguous verb in Dutch, meaning 'to drive' as well as 'to manage'.

The complete clustering is given below:

- *betaal* 'pay', *herstel* 'repair', *koop* 'buy', *leen* 'lend' , *merk* 'notice', *verkoop* 'sell', *verwerf* 'acquire'

- *ga_weg* ‘leave’, *herinner* ‘remember’, *huil* ‘cry’, *lach* ‘smile’, *lees* ‘read’, *luister* ‘listen’, *praat* ‘talk’, *vergeet* ‘forget’, *voel* ‘feel’, *weet* ‘know’

- *bestuur* ‘drive’, *controleer* ‘check’, *evalueer* ‘evaluate’, *spreek* ‘speak’, *suggereer* ‘suggest’,*verzoek* ‘request’, *zend* ‘send’

- *adem* ‘breathe’, *beweeg* ‘move’, *draag* ‘carry’, *drink* ‘drink’, *duw* ‘push’, *eet* ‘eat’, *kijk* ‘look’, *loop* ‘run’, *ruik* ‘smell’, *sta_op* ‘rise’, *trek* ‘pull’, *wandel* ‘walk’

- *breek* ‘break’ , *dood* ‘kill’ , *ga_binnen* ‘enter’, *kom_aan* ‘arrive’, *rij* ‘ride’, *sterf* ‘die’, *val* ‘fall’, *verniel* ‘destroy’, *vlieg* ‘fly’

#### 3.3.3 Syntax-based

Table 7 gives the clustering results of the syntax-based methods for different clustering sizes.

| method | n-way | entropy | purity |
|---|---|---|---|
| LOCPMI | 5 | .516 | .644 |
| | 9 | .432 | .489 |
| GLOPMI | 5 | .464 | .667 |
| | 9 | .408 | .556 |

Table 7: A comparison of different clustering results for verb categorization — syntactic approach

The global PMI approach yields slightly better results than the local one, but differences are again small. The more specific clustering is slightly better than the more general one.

Figure 4 shows the confusion matrix for the best-performing 5-way clustering (GLOPMI).

| cluster | cogn | moti | body | exch | chan |
|---|---|---|---|---|---|
| 1 | 0 | 8 | 1 | 0 | 0 |
| 2 | 1 | 2 | 0 | 5 | 1 |
| 3 | 7 | 0 | 3 | 0 | 0 |
| 4 | 2 | 2 | 0 | 0 | 4 |
| 5 | 0 | 3 | 6 | 0 | 0 |

Figure 4: Confusion matrix

The first cluster contains many motion verbs; the second one has many exchange verbs, and the third one contains many cognition verbs. The fourth cluster contains mainly change verbs, but also non-related cognition and motion verbs, and the fifth one contains mostly motion verbs.

The complete clustering is given below:

- *beweeg* ‘move’, *duw* ‘push’, *kijk* ‘look’, *loop* ‘run’, *rij* ‘ride’, *trek* ‘pull’, *vlieg* ‘fly’, *wandel* ‘walk’, *zend* ‘send’

- *bestuur* ‘drive’, *betaal* ‘pay’, *controleer* ‘check’, *draag* ‘carry’, *koop* ‘buy’, *leen* ‘lend’, *verkoop* ‘sell’, *verniel* ‘destroy’, *verwerf* ‘acquire’

- *adem* ‘breathe’, *herinner* ‘remember’, *lees* ‘read’, *merk* ‘notice’, *praat* ‘talk’, *spreek* ‘speak’, *sugerreer* ‘suggest’, *vergeet* ‘forget’, *voel* ‘feel’, *weet* ‘know’

- *breek* ‘break’, *dood* ‘kill’, *evalueer* ‘evaluate’, *herstel* ‘repair’, *kom_aan* ‘arrive’, *sterf* ‘die’, *val* ‘fall’, *verzoek* ‘request’

- *drink* ‘drink’, *eet* ‘eat’, *ga_binnen* ‘enter’, *ga_weg* ‘leave’, *huil* ‘cry’, *lach* ‘smile’, *luister* ‘listen’, *ruik* ‘smell’, *sta_op* ‘rise’

#### 3.3.4 Comparison of both approaches

The performance of the bag of words model and the syntax-based model is similar; neither of both really outperforms the other. The more specific clustering solutions are slightly better than the general ones.

There is considerable difference between the clustering solutions found by the bag of words approach and the syntax-based approach. Again, this might be due to the kind of similarity found by the models. The bag of words approach seems to be influenced by topics again (the business related cluster), whereas the syntax-based model might be influenced by more general features (‘motion’ in the first cluster). But given the evaluation results of the verb clustering, these are very tentative conclusions.

## 4 Conclusions & Future Work

The evaluation results presented in the former section indicate that semantic space models are fruitful models for the induction of actual semantic classes.

Especially the noun categorizations – the concrete noun categorization and the concrete/abstract noun discrimination task – perform very well. Verb categorization is a more difficult task for the semantic models presented in this paper: the results are worse than those for the noun categorizations.

In general, the syntax-based approach yields better results than the bag of words approach. This might be due to the fact that bag of words models get at a kind of topical semantic similarity, whereas the syntax-based model might be able to extract more abstract properties of the words. These are, however, tentative conclusions for a tendency present in the data. More research is needed to found this statement.

In most cases, a global method of calculation PMI yields better results. If the algorithm is able to normalize the word vectors according to the distributions of a large number of words in the data, the clustering solution is generally better. There are, however, some exceptions.

In the future, other semantic models need to be investigated for the categorization of verbs. Including subcategorization information in the models might be beneficial for the clustering of verbs, as well as generalizing among the feature space of the verb's dependencies (e.g. by using semantic noun clusters instead of nouns).

One last issue for future work is the comparison between small clustering tasks like the ones presented above, and the clustering solutions of a large clustering framework, in which a large number of words are captured.

Nevertheless, the results of the present evaluation tasks indicate that word space models are a suitable tool for the induction of semantic classes.

## References


Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22–29.

Z. Harris. 1985. Distributional structure. In Jerrold J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press.

George Karypis. 2003. CLUTO - a clustering toolkit. Technical Report #02-017, nov.

Thomas Landauer and Se Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychology Review*, 104:211–240.

Thomas Landauer, Peter Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:295–284.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562.

R.J.F. Ordelman. 2002. Twente Nieuws Corpus (TwNC), August. Parlevink Language Techonology Group. University of Twente.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, Leuven.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical Report #01-40.

# Decorrelation and Shallow Semantic Patterns for Distributional Clustering of Nouns and Verbs

**Yannick Versley**
SFB 441
University of Tübingen
versley@sfs.uni-tuebingen.de

## Abstract

Distributional approximations to lexical semantics are very useful not only in helping the creation of lexical semantic resources (Kilgariff et al., 2004; Snow et al., 2006), but also when directly applied in tasks that can benefit from large-coverage semantic knowledge such as coreference resolution (Poesio et al., 1998; Gasperin and Vieira, 2004; Versley, 2007), word sense disambiguation (McCarthy et al., 2004) or semantical role labeling (Gordon and Swanson, 2007).

We present a model that is built from Web-based corpora using both shallow patterns for grammatical and semantic relations and a window-based approach, using singular value decomposition to decorrelate the feature space which is otherwise too heavily influenced by the skewed topic distribution of Web corpora.

## 1 Introduction

It is well-established that human learning of lexical items beyond a certain point is driven by considering the *contexts* in which a word occurs, and it has been confirmed by McDonald and Ramscar (2001) that few occurrences of a word in informative contexts suffice to influence similarity judgements for marginally known words.

Computational models of word semantics based on this assumption are not only attractive for psychological modelling of language, but also for the purposes of automatic text processing, especially for applications where manual ontology construction would be infeasible or overly expensive, or to aid manual construction of lexical resources (cf. Kilgariff et al. 2004).

A common approach (Philips, 1985; Hindle, 1990) is to represent the context a word appears in by the words occurring in that context, weighting more heavily the context elements that co-occur more often than expected for random co-occurrences.

It is possible to group the approaches to use collocate features into two main areas:

- relation-free methods aim to directly use vectors of collocate words as a representation without distinguishing the relation between the target word and its collocates. Thus, related terms such as *doctor*, *hospital* and *treatment* which share many collocates, would be assigned a high similarity value.

- relation-based methods use collocate words together with grammatical relations, so that one noun being a frequent subject and another being a frequent object of a given word would not increase their similarity score – in the hospital example, a context like *the doctor treats the patient* would not contribute to the similarity value of *doctor* and *patient*.

Different methods of extracting word features will pick up different aspects of the denoted concept, from general topic, to sentiment, to ontologically relevant features such as exterior appearance.

In the remainder of this paper, I will start from the hypothesis that basing distributional similarity measures on context elements that are informative (in the sense that they implicitly or explicitly reflect the ontological principles of the targeted taxonomy) is preferable, and, by extension, that explicitly using syntactico-semantic relations yields better results.

## 2 Experimental Setting

To be useful in real-world tasks, both the size of the vocabulary and the size of the corpus should be large enough, as smaller samples would not contain enough contexts for many of the rarer words. This precludes approaches that rely on large numbers of search engine queries, such as the ones by Markert and Nissim (2005), Almuhareb and Poesio (2005), or Geleijnse and Korst (2006), as achieving significant coverage would necessitate an order of magnitude more effort than the (already very significant) weeks or months of running search engine queries that are necessary for a smaller sample.

On the other hand, the time consumption of full parsing means that approximate methods can be a better fit for processing very large corpora: Curran and Moens (2002) find that the rather large time that full parsing takes (even with a fast parser such as Lin's (1998b) MiniPar) can be reduced by using a reimplementation of Grefenstette's (1992) Sextant system for approximate parsing, which uses a chunker and considers simple neighbourhood relationships between chunks to extract compound, subject and object relations. Since the Sextant reimplementation only uses chunks, it is much faster (by a factor of 27), while the accuracy for the extracted relations is rather close to that of full parsing; Curran also remarks that a simple window-based approach is even faster and can still achieve good quality on large corpora, even though it is inferior to the syntax-based approaches.

In the following, we will explore the use of two large, Web-based datasets, namely UK-WaC (Ferraresi, 2007), as well as Google's n-gram database[1] for unsupervised noun and verb clustering, evaluated on the corresponding datasets proposed by the workshop organisers.

Besides a purely window-based approach, which we will present in section 4, we will present an approach that uses shallow patterns to approximate syntactic and semantic relationships, in section 3; even though some of the relations need more processing in different languages (most notably verb arguments, which are nontrivial to identify in languages with free word order such as German or Czech, or between compound parts in languages with synthetic compounds), we can show that this approach is not only computationally relatively inexpensive but also yields high-quality clustering results for verb clustering, where current approaches do not consider semantic relations at all.

[1] Thorsten Brants, Alex Franz (2006): Web 1T 5-gram Version 1, LDC2006T13

### 2.1 Relational Features for Nouns

Most older approaches to distributional similarity focus on syntactic relations, such as the compound noun, adjective-noun, subject and object relations that Grefenstette (1992) extract from his SEXTANT shallow parser, or the larger set of relations that Lin (1998a) extracts by full parsing.

Clustering words using such ontologically motivated patterns has been used by Evans (2003), who uses hypernymy patterns such as those popularised by Hearst (1992) to cluster named entities, and by Almuhareb and Poesio (2005), who use a pattern inspired by Berland and Charniak's (1999) to cluster nouns by their attributes. Using pattern search on the World Wide Web, Almuhareb and Poesio are able to achieve very good results. Some researchers such as Pantel et al. (2004) use supervised training to learn patterns corresponding to a single relation; going past single ontological relations, Baroni and Lenci (2008) use supervised learning of surface patterns corresponding to relations out of an inventory of 20 relations.

For our experiments, we used a combination of syntactic patterns targeting the same relations as Grefenstette (1992), variants of the hypernymy and meronymy-related patterns popularised by Hearst (1992) and Berland and Charniak (1999), respectively, as well as coordinate structures ($X$ and/or $Y$); in contrast to Cederberg and Widdows (2003), we use second-order associations (regarding as similar terms which are coordinated with the same feature words) and do not see coordination as an indication for similarity of the conjuncts.

### 2.2 Relational Features for Verbs

Clustering and classification of verbs in the literature McCarthy (2000); Schulte im Walde and Brew (2002) often makes heavy use of information about argument structure, which is hard to come by without parsing; Stevenson and collaborators (Stevenson and Merlo, 1999; Joanis et al., 2007) use shallower

| UK-Wac relation | entropy | purity |
|---|---|---|
| nv | 0.209 | 0.818 |
| vn$^{-1}$ | 0.244 | 0.750 |
| jjn$^{-1}$ | 0.205 | 0.773 |
| **nn** | **0.172** | **0.841** |
| nn$^{-1}$ | 0.218 | 0.795 |
| cc:and | 0.241 | 0.750 |
| cc:and$^{-1}$ | 0.210 | 0.750 |
| cc:or | 0.203 | 0.767 |
| cc:or$^{-1}$ | 0.200 | 0.795 |
| Y's X | 0.566 | 0.475 |
| Y's X $^{-1}$ | 0.336 | 0.725 |
| X of Y | 0.437 | 0.655 |
| X of Y $^{-1}$ | 0.291 | 0.750 |
| **Google n-grams relation** | **entropy** | **purity** |
| of the | 0.516 | 0.579 |
| of the$^{-1}$ | 0.211 | 0.818 |
| and other | 0.237 | 0.744 |
| and other$^{-1}$ | 0.458 | 0.632 |
| such as | 0.335 | 0.692 |
| such as$^{-1}$ | 0.345 | 0.675 |

Table 1: Shallow patterns for nouns

features of which some do not necessitate parsed input, but they concentrate on verbs from three classes and it is not certain whether their features are informative enough for larger clustering tasks.

Schulte im Walde (2008) uses both grammatical relations output by a full parser and and part-of-speech classes co-occurring in a 20 word window to cluster German verbs. Comparing her clustering to gold standard classifications extracted from GermaNet (a German wordnet) and German FrameNet and another gold-standard using classes derived from human associations. She found that the different gold standards preferred different classes of grammatical relations: while GermaNet clustering results were best using subjects of nontransitive verb occurrences, FrameNet results were best when using adverbs, and the human association were best matched using NP and PP dependents on verbs.

In addition to syntactic correlates such as those investigated by Schulte im Walde (2008), we use several patterns targeted at more semantic relations.

Chklovski and Pantel (2004) extract 29,165 pairs of transitive verbs that co-occur with the same subject and object role, using Lin and Pantel's (2001) DIRT (Discovery of Inference Rules from text) approach, and then classify the relation between these verbs into several relations using Web patterns indicating particular relations (*similarity*, *strength*, *antonymy*, *enablement*, and *succession*.

Besides detecting conjunctions of verbs (allowing other words in between, but requiring the part-of-speech tags to match to exclude matches like "see how scared I *was* and *started* to calm me"), and capturing general within-sentence co-occurrence of verbs, we also tried to capture discourse relations more explicitly by limiting to certain discourse markers, such as *that*, *because*, *if*, or *while*.

## 3 Clustering Results

To determine the weight for an association in the vector calculated for a word, we use the pointwise mutual information value:

$$mi^{+}(w_1, w_2) = \max\left(0, \log \frac{p(X = w_1 | Y = w_2)}{p(X = w_1)}\right)$$

We then use the vectors of $mi^{+}$ values for clustering in CLUTO using repeated bisecting $k$-means with cosine similarity.[2]

For the nouns, we use a the last noun before a verb as an approximation of subjecthood (`vn`), the next head noun as an approximation for direct objects (`nv`), as well as adjective modifiers (`jjn`), and noun compounds (`nn`) on UK-WaC using the provided lemmas. Using Berland and Charniak's patterns A and B ($Y$'s $X$, $X$ of $Y$) on UK-WaC, we found that a surface string search (using Minnen et al.'s (2001) morphological analyser to map word forms to their lemmas) on the Google n-gram dataset gave superior results. We used the same surface string search for Hearst's *X and other Ys* and *Ys such as X* patterns (restricting the "Ys" part to plural nouns to improve the precision). As the Hearst-style patterns are relatively rare, the greater quantity of data from the Google n-grams outweighs the drawback of having no part of speech tagging and only approximate lemmatisation.

Both on UK-WaC and on Google's n-gram dataset, we find a stark asymmetry in the clusterings

[2]Note that the resulting clusters can vary depending on the random initialisation, which means that re-running CLUTO later can result in slightly better or worse clustering.

| UK-Wac relation | entropy | purity |
|---|---|---|
| nv$^{-1}$ | 0.398 | 0.556 |
| vn | 0.441 | 0.511 |
| rv$^{-1}$ | **0.342** | **0.622** |
| vi | 0.397 | 0.556 |
| vv | 0.423 | 0.533 |
| vv$^{-1}$ | 0.378 | 0.556 |
| that | 0.504 | 0.467 |
| that$^{-1}$ | 0.479 | 0.489 |
| because | 0.584 | 0.378 |
| because$^{-1}$ | 0.577 | 0.400 |
| if | 0.508 | 0.444 |
| if$^{-1}$ | 0.526 | 0.444 |
| while | **0.477** | **0.511** |
| while$^{-1}$ | 0.502 | 0.444 |
| by $X$ing | 0.488 | 0.489 |
| by $X$ing$^{-1}$ | 0.380 | 0.600 |
| then | 0.424 | 0.533 |
| then$^{-1}$ | 0.348 | 0.600 |
| cc:and | 0.278 | 0.711 |
| cc:and$^{-1}$ | 0.329 | 0.622 |
| cc:or | **0.253** | **0.733** |
| cc:or$^{-1}$ | 0.323 | 0.667 |

Table 2: Shallow patterns for verbs

of meronymy patterns, probably due to the fact that parts or attributes provide useful information, but the nouns in the evaluation set are not meaningful parts of other objects.

Considering the verbs, we found that a preceding adverb (rv) provided the most useful information, but other patterns, such as subject-verb (nv), and verb-object (vn), as well as using the following preposition (vi) to approximate the distribution of prepositional modifiers of the verb, give useful results, as much as the following verb (vv), which we used for a very rough approximation of discourse relations. Using verbs linked by subordinate conjunctions such as *if*, *that*, or *because*, performs comparatively poorly, however.

A third group of patterns is inspired by the patterns used by Chklovski and Pantel (2004) to approximate semantic relations between verbs, namely *enablement* relations expressed with gerunds (linking the previous verb with the gerund in sentences such as "Peter *altered* the design by *adding* a green button"), *temporal succession* by relating any verb that is modified by the adverb *then* with its preceding verb, and *broad similarity* by finding pairs of coordinated verbs (i.e., having a coordination between them and marked with the same part-of-speech tag).

Noun compounds for nouns and preceding adverbs for verbs already give slightly better clusterings than an approach simply considering words co-occurring in a one-word window (see table 3), with coordination and some of the semantic patterns also yielding results on par with (for nouns) syntactic relations.

## 4 Window-based approach with decorrelation

As reported by Curran and Moens (2002), a simple cooccurrence-window-based approach, while inferior to approaches based on full or shallow parsing, is amenable to the treatment of much larger data quantities than parsing-based approaches, and indeed, some successful work such as Rapp (2003) or Ravichandran et al. (2005) does not use syntactic information at all.

In this section, we report the results of our approach using window-based cooccurrence on Google's n-gram dataset, using different weighting functions, window sizes, and number of feature words. As a way to minimize the way of uninformative collocates, we simply excluded the 500 most frequent tokens for use as features, using the next most frequent $N$ words (for $N$ in 8k, 24k, 64k, 512k).

Besides the positive mutual information measure introduced earlier, we tried out a simple logarithmic weighting function:

$$Log(w_1, w_2) = \log(1 + C(w_1, w_2))$$

(where $C(w_1, w_2)$ is the raw count for $w_1$ and $w_2$ co-occurring in a window), and the entropy-weighted variant used by Rapp (2003):

$$LogEnt(w_1, w_2) = \log(1+C(w_1, w_2))\cdot H(X|\,Y{=}w_2)$$

This weighting function emphasizes features (i.e., values for $w_2$) which co-occur with many different target words.

Generally, we found that the window-based approach gave the best results with mutual information weighting (with clustering entropy values for verbs between 0.363, for using 8k features with a

window size of 1 word around the target word, and 0.504, for using 512k features with a window size of 4) than for the other methods (which yielded entropy values between 0.532, for 64k features with a window of 2 words and logarithmic weighting and 0.682, for 8k features with a window size of 4 words and log-entropy weighting). This difference is statistically very significant ($p < 0.0001$ for a paired t-test between $mi^+$ and $Log$ over combinations of three different window sizes and four different vocabulary sizes).

To see if singular value decomposition would improve the clustering results, we collected co-occurrence vectors for the clustering target verbs in addition to a collection of frequent verbs that we obtained by taking the 2000 most frequent verbs or nouns and eliminating verbs that correspond to very frequent noun forms (e.g., to machine), as well as all non-nouns (e.g. gonna), arriving at a set of 1965 target verbs, and 1413 target nouns, including the items to be clustered. Even though using this larger data set makes it more difficult to experiment with larger feature spaces, we saw the possibility that just using the words from the data set would create an artificial difference from the transformation one would get using SVD in a more realistic setting and the transformation obtained in the experiment.

Using singular value decomposition for dimensionality reduction only seems to have a very small positive effect on results by itself: using mutual information weighting, we get from 0.436 to between 0.408 (for 100 dimensions), with other weighting functions, dimensionality values, or vocabulary sizes perform even worse.

This is in contrast to Rapp (2003), who achieved vastly better results with SVD and log-entropy weighting than without in his experiments using the British National Corpus, and in parallel to the findings of Baroni and Lenci (2008), who found that Rapp's results do not carry over to a web-based corpus such as UK-WaC. Looking at table 4, we find it plausible that the window-based approach tends to pick up topic distinctions instead of semantic regularities, which gives good results on a carefully balanced corpus such as the BNC, but drowns other information when using a Web corpus with a (typically) rather biased topic distribution.[3]

Examining the singular vectors and values we get out of the SVD results, we find that the first few singular values are very large, and the corresponding vectors seem to represent more a topic distinction than a semantic one. Parallel to this, the results for the SVD of log-weighted data is plateauing after the first few singular vectors are added, quite possibly due to the aforementioned drowning of information by the topical distinctions. To relieve this, we altered the size of singular values before clustering, either by taking the square root of the singular values, which has the effect of attenuating the effect of the singular vectors with large values, or by setting all singular values to 1, creating a feature space that has a spherically symmetric data distribution (usually referred to as decorrelation or whitening). As can be seen in figure 1, decorrelation yields clearly superior results, even though they are clearly much noisier, yielding wildly varying results with the addition of just a few more dimensions. For the decorrelated vectors, we find that depending on the other parameters, positive mutual information is either significantly better ($p \approx 0.0001$ for paired t-test over results for different dimension numbers with a window size of 1 and 8k features), or insignificantly worse ($p \approx 0.34$ for a window size of 2 and 24k features). We attribute the fact that the best clustering result for the window-based approach was achieved with log-entropy weighting to the fact that the log and log-entropy based vectors are noisier and have more variance (with respect to number of dimensions), thus possibly yielding artifacts of overfitting the small test data set; however, further research will be necessary to confirm or deny this.

## 5 Results and Discussion

To get a better clustering than would be possible using single features, we tried combinations of the most promising single features by first normalizing the individual feature vectors by their $L_p$ norm,

[3] Cf. table 4: besides the first two vectors, which seem to identify frequency or content/navigation distinction, the second and third singular vector are clearly influenced by dominant web genres, with a pornography vs. regulatory documents axis for v2 and a Unix/programming vs. newswire documents axis for vector v3.

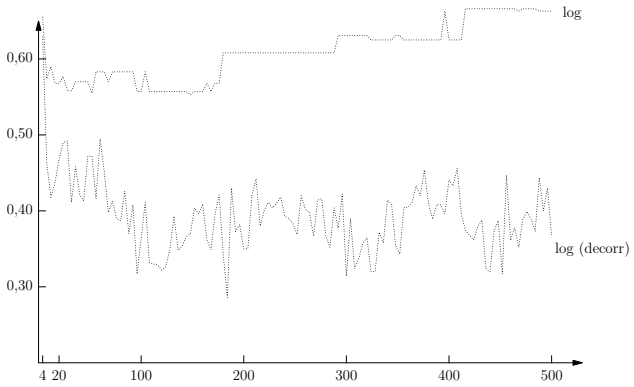



Figure 1: Influence of decorrelation on clustering quality (4-word window, 8k features)

Clustering Entropy in verb clustering vs. number of dimensions; lower is better

| noun clustering relation | entropy | purity |
|---|---|---|
| win(1), 64k features, $mi^+$ | 0.221 | 0.818 |
| best SVD+decorrelation | 0.196 | 0.795 |
| nn | 0.172 | 0.841 |
| cc:or$^{-1}$ | 0.200 | 0.795 |
| nv$^{-1}$+jjn$^{-1}$+and other, 7cl. | **0.034** | **0.977** |
| **verb clustering relation** | **entropy** | **purity** |
| win(1), 1M features, $mi^+$ | 0.376 | 0.600 |
| best SVD+decorrelation | 0.280 | 0.711 |
| rv$^{-1}$ | 0.342 | 0.622 |
| cc:or | 0.253 | 0.733 |
| cc:and+then$^{-1}$ | **0.218** | **0.778** |

Table 3: Results overview

for $p = 1.5$.[4] We then concatenate the normalized vectors for the different relations to get the vector used in clustering. As can be seen in table 3, the window-based approach comes near the best results for a single syntax-based pattern, whereas the semantically motivated patterns work better than either syntactic patterns or the window-based approach. The best combinations we found involve several of the semantically motivated patterns and, in the case of nouns, also informative syntactic relations the key seems to be that the different relation focus on complementary aspects of the classes. While the decorrelation-based approach is an improvement over a simpler window-based approach, it does not seem possible to get much larger improvements; however, it should be said that both window size and feature space were constrained due to limitations of the Google n-gram data on one hand and memory limitations on the other.

The resulting clusters generally seem rather sensible, although they sometimes incorporate distinctions that are slightly different from those in the gold standard: in many clusterings, the class of birds and ground animals are split according to a different pattern, e.g. domestic and wild animals. Some other divisions are very consistently found in all clusterings: Even in the best clustering, artifacts are split into a container-like group including *bottle, bowl, cup* and others, and a handle-like artifact group including *chisel, hammer, screwdriver*, and fruits and vegetables are merged into one group unless the number of clusters is increased to seven. *chicken* also seems to be consistently misclustered as a cooking ingredient rather than an animal.

For the verbs, the communication verbs are split into the non-directive verbs *read*, *speak* and *talk*, which are clustered with two mental state verbs which are less action-focused, *know* and *remember*, as well as *listen*, which the gold standard categorizes as a body sense verb, whereas the more directive communication verbs *request* and *suggest* are clustered together with the more action-focused mental state verbs *check* and *evaluate*, and *repair*, which the gold standard groups with the state change verbs (*break*, *die*, *kill*).

## 6 Outlook

We presented two approaches for using distributional statistics extracted from large Web-based corpora to cluster nouns and verbs: one using shallow patterns to extract syntactically and semantically motivated relations, and the other using a small window size together with Google's n-gram dataset, showing how manipulating the SVD-transformed representation helps overcome problems that are due to the skewed topic distribution of Web corpora. We also showed how multiple relations can be combined to arrive at high-quality clusterings that are better

[4] The Lebesgue norm $L_p = \left(\sum |x_i|^p\right)^{1/p}$ has the euclidean norm $L_2$ as a special case. For $1 \leq p < 2$, the $L_p$-norm is larger than the euclidean norm if there are multiple non-zero values in a vector; we think that normalizing by the $L_{1.5}$ norm rather than $L_2$ norm has the beneficial effect of slightly emphasizing relations with a smaller feature space.

| v0: $\lambda = 56595$ | | v1: $\lambda = 2043.5$ | | v2: $\lambda = 2028.7$ | | v3: $\lambda = 1760.5$ | |
|---|---|---|---|---|---|---|---|
| fundraise | *0.0000 | ensure | *-9999.99 | f–ck | a–s | configure | src |
| exhilarate | *Reserved | determine | *Verzeichnis | suck | p–ssy | filter | header |
| socialize | *Advertise | process | *-99 | *amend | *pursuant | *accuse | *father |
| pend | *Cart | identify | *-999 | *comply | *Agreement | *murder | *whom |

Table 4: Singular vectors for the largest singular values (8k features, 4-word window)
Most important target verbs (left) and features (right), starred words have a negative weight in the vector. Some explicit words in vector 2 have been redacted by replacing middle letters with a dash.

| noun clusters | | |
|---|---|---|
| banana | cat | bottle[1] |
| cherry | cow | bowl[1] |
| pear | dog | kettle[1] |
| pineapple | elephant | pencil[1] |
| chisel[1] | lion | pen[1] |
| hammer[1] | pig | spoon[1] |
| knife[1] | snail | telephone[1] |
| scissors[1] | turtle | |
| screwdriver[1] | | |
| duck | *chicken* | boat |
| eagle | corn | car |
| owl | lettuce | helicopter |
| peacock | mushroom | motorcycle |
| penguin | onion | ship |
| swan | potato | truck |
| **verb clusters** | | |
| breathe | drive | carry |
| cry | fly | pull |
| drink | ride | push |
| eat | run | send |
| | walk | |
| acquire | break | feel |
| buy | destroy | look |
| lend | die | notice |
| pay | kill | smell |
| sell | *fall* | *smile* |
| check[2] | know[2] | arrive |
| evaluate[2] | remember[2] | enter |
| *repair* | *listen* | leave |
| request[3] | read[3] | rise |
| suggest[3] | speak[3] | *move* |
| | talk[3] | *forget* |

Table 5: Resulting verb and noun clusters
(Each cluster is one column. Italicized items are the only members of their class in the cluster)

than would be possible using either single relations or the best results achieved using the window-based approach.

Several open questions remain for future research: One would be the use of supervised learning approaches to perform automatic weighting and/or acquisition of patterns. The other one would be a question of how these approaches can be scaled up to the size needed for real-world applications. While the most important issue for the window-based approach is the use of Singular Value Decomposition, which scales poorly with both the size of the dataset due to nonlinear growth of computation time as well as memory consumption, the relation-based approach may suffer from data sparsity when considering rare words, especially using the rarer semantic relations; however, an approach like the ones by Snow et al. (2006) or Baroni and Lenci (2008) that is able to learn patterns from supervised training data may solve this problem at least partially.

**Acknowledgements** I am grateful to the two anonymous reviewers for helpful comments on an earlier version of this paper. The research reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of Collaborative Research Centre (Sonderforschungsbereich) 441 "Linguistic Data Structures".

## References

Almuhareb, A. and Poesio, M. (2005). Finding concept attributes in the web. In *Proc. of the Corpus Linguistics Conference*.

Baroni, M. and Lenci, A. (2008). Concepts and word spaces. *Italian Journal of Linguistics*, to appear.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL-1999*.

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the preci-

sion and recall of automatic hyponymy extraction. In *Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*.

Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proc. EMNLP 2004*.

Curran, J. and Moens, M. (2002). Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Evans, R. (2003). A framework for named entity recognition in the open domain. In *RANLP 2003*.

Ferraresi, A. (2007). Building a very large corpus of English obtained by Web crawling: ukWaC. Master's thesis, Università di Bologna.

Gasperin, C. and Vieira, R. (2004). Using word similarity lists for resolving indirect anaphora. In *ACL'04 workshop on reference resolution and its applications*.

Geleijnse, G. and Korst, J. (2006). Learning effective surface text patterns for information extraction. In *Proceedings of the EACL 2006 workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Gordon, A. S. and Swanson, R. (2007). Generalizing semantic role annotations across syntactically similar verbs. In *ACL 2007*.

Grefenstette, G. (1992). Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *ACL Student Session 1992*.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING 92)*.

Hindle, D. (1990). Noun classification from predicate argument structures. In *Proceedings of the 28th annual meeting of the Association for Computational Linguistics*.

Joanis, E., Stevenson, S., and James, D. (2007). A general feature space for automatic verb classification. *Natural Language Engineering*, forthcoming:1–31.

Kilgariff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. In *EURALEX 2004*.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proc. CoLing/ACL 1998*.

Lin, D. (1998b). Dependency-based evaluation of Minipar. In *Workshop on the Evaluation of Parsing Systems*.

Lin, D. and Pantel, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Markert, K. and Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402.

McCarthy, D. (2000). Using semantic prefernces to identify verbal participation in role switching alternations. In *Proc. NAACL 2000*.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279, Morristown, NJ, USA. Association for Computational Linguistics.

McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proc. 23rd Annual Conference of the Cognitive Society*.

Minnen, G., Caroll, J., and Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale knowledge acquisition. In *Proc. Coling 2004*.

Philips, M. (1985). *Aspects of Text Structure: An investigation of the lexical organization of text*. Elsevier, Amsterdam.

Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *AAAI Spring Symposium on Learning for Discourse*.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proc. Ninth Machine Translation Summit*.

Ravichandran, D., Pantel, P., and Hovy, E. (2005). Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proc. ACL 2005*.

Schulte im Walde, S. (2008). Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, to appear.

Schulte im Walde, S. and Brew, C. (2002). Inducing german semantic verb classes from purely syntactic subcategorization information. In *Proc. ACL 2002*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*.

Stevenson, S. and Merlo, P. (1999). Automatic verb classification using grammatical features. In *Proc. EACL 1999*.

Versley, Y. (2007). Antecedent selection techniques for high-recall coreference resolution. In *Proc. EMNLP 2007*.

# Does Latent Semantic Analysis Reflect Human Associations?

**Tonio Wandmacher**
Institute of Cognitive Science
University of Osnabrück
Albrechtstr. 28,
49069, Germany
twandmac@uos.de

**Ekaterina Ovchinnikova**
Institute of Cognitive Science
University of Osnabrück
Albrechtstr. 28,
49069, Germany
eovchinn@uos.de

**Theodore Alexandrov**
Center for Industrial Math.
University of Bremen
Bibliothekstr. 2,
28359, Germany
theodore@
math.uni-bremen.de

## Abstract

In the past decade, Latent Semantic Analysis (LSA) was used in many NLP approaches with sometimes remarkable success. However, its abilities to express semantic relatedness have been not yet systematically investigated. In this work, the semantic similarity measures as provided by LSA (based on a term-by-term matrix) are compared with human free associations. Three tasks have been performed: (i) *correlation* with human association norms, (ii) *discrimination* of associated and unassociated pairs and (iii) *prediction* of the first human response. After a presentation of the results a closer look is taken to the statistical behavior of the data, and a qualitative (example-based) analysis of the LSA similarity values is given as well.

## 1 Introduction

In its beginnings, Latent Semantic Analysis aimed at improving the vector space model in information retrieval. Its abilities to enhance retrieval performance were remarkable; results could be improved by up to 30%, compared to a standard vector space technique (Dumais, 1995). It was further found that LSA was able to retrieve documents that did not even share a single word with the query but were rather semantically related.

This finding was the headstone for many subsequent researches. It was tried to apply the LSA approach to other areas, such as automated evaluation of student essays (Landauer et al., 1997) or automated summarization (Wade-Stein and Kintsch, 2003). In (Landauer and Dumais, 1997), even an LSA-based theory of knowledge acquisition was presented.

Many researches have made claims on the analytic power of LSA. It is asserted that LSA does not return superficial events such as simple contiguities, but is able to describe semantic similarity between two words (cf. Wade-Stein and Kintsch, 2003). The extracted word relations are referred to as latent, hidden or deep (cf. Landauer et al. 1998), however, only few articles address the nature of this deepness.

Some steps in this direction were taken by Landauer and Dumais (1997) and later by Rapp (2003). In these works, LSA-based similarities were used to solve a synonym test, taken from the TOEFL[1]. However, the results achieved can only be seen as a first indication for the capacity of LSA.

We try to make a little step further. The main objective of this work is therefore not improvement, but evaluation and a better understanding of the method. The present investigation is carried out in the framework of the *Lexical Semantics Workshop: Bridging the gap between semantic theory and computational simulations* at ESSLLI'08[2], which is devoted to discovering of the relationships between word spaces computed by corpus-based distributional models and human semantic spaces. In this paper, we concentrate on exploration of the correlation between the LSA semantic similarity measures and human free associations[3].

[1] Test Of English as a Foreign Language

[2] *http://wordspace.collocations.de/doku.php/esslli:start*.

[3] See *http://wordspace.collocations.de/doku.php/data:correlation_with_free_association_norms*.

The paper is structured as follows. In section 2 we briefly introduce the LSA method. We then (section 3) give an overview on related work exploring the semantic and associative capacities of LSA. In section 4 we describe the workshop tasks on free associations and provide the results that we have obtained. In section 5 we present a detailed quantitative and qualitative analysis of the achieved results. In the final section we draw conclusions and discuss open issues.

## 2 Latent Semantic Analysis: Method

LSA is based on the vector space model from information retrieval (Salton and McGill, 1983). Here, a given corpus of text is first transformed into a term×context matrix $A$, displaying the occurrences of each word in each context. The decisive step in the LSA process is then a *singular value decomposition* (SVD) of the matrix which enables the mapping of this matrix to a subspace. The resulting lower-dimensional matrix is the best reduced-rank least-squares approximation of the original matrix. According to the proponents of LSA this reduction plays an important role for the uncovering of important relations which are hidden (or '*latent*') in the original matrix.

In its original form (cf. Deerwester et al. 1990), LSA is based on a co-occurrence matrix of terms in documents; such a matrix is normally extremely sparse[4], and it is obvious that this matrix grows with the number of documents in the training corpus. Moreover, the notion of document varies strongly over different corpora: a document can be only a paragraph, an article, a chapter or a whole book, no hard criteria can be defined. Therefore, another type of matrix can be used, as described by (Schütze, 1998) and (Cederberg and Widdows, 2003), which is not based on occurrences of terms in documents but on other co-occurring terms (term×term-matrix). The two sets of terms need not be identical, one can also define a (usually smaller) set of *index terms* $I = (i_1, ..., i_m)$. The size of the matrix is then independent of the size of the training data, so that much larger corpora can be used for training.

After applying SVD, each word is represented as a vector of $k$ dimensions, and for every word pair $w_i$, $w_j$ of the vocabulary we can calculate a similarity value $\cos(w_i, w_j)$, based on the $cosine$ between their respective vectors.

Figure 1 summarizes the processing steps involved in training an LSA-based semantic space.

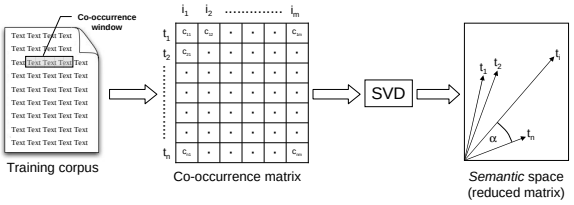



Figure 1: Schematic overview on the generation of an LSA-based semantic space

In the following we will apply this kind of model, based on an SVD-reduced term-by-term co-occurrence matrix, to the different tasks, and we will compute term similarity by measuring the cosine of term vectors in the reduced space.

## 3 Related Work

Considering the large number of works applying LSA for various purposes, it is a surprising matter of fact that only little research was done in order to better understand the kind of relatedness that distributional approaches like LSA are able to reflect.

In (Landauer and Dumais, 1997) a theory of knowledge acquisition and representation is presented, assuming that the meaning induction mechanisms performed by LSA are very similar to those of humans. As an example task, LSA is applied to solve the TOEFL synonym test, and it could be shown that the results of LSA are the same as those of the average foreign student passing the *TOEFL* (LSA: 64.4%; human participants: 64.5%). In (Rapp, 2003), an LSA model, based on a term×term matrix and trained on much more data, was able to solve even 92.5% of the synonym questions.

In (Wandmacher, 2005) term relations (nearest neighbors) generated by LSA as well as a first-order co-occurrence approach are systematically analyzed and compared. It could be shown that only a small part of the relations are systematically related (e. g. by hyponymy or synonymy), the largest part of the nearest neighbors of a term were loose associations.

[4] In (Wandmacher, 2005) a matrix was used that had less than 0.08% non-zero elements.

While the error rate for LSA was lower than for the first-order approach, no substantial differences between the results of the two methods could be determined. It could however be observed that a crucial factor for the quality of the nearest neighbors is the specificity of a term.

The correspondence of human association and first-order co-occurrence was investigated in (Wettler et al., 2005). Here, 100 stimulus words from the Kent-Rosanoff word association test with associations selected from the *Edinburgh Associative Thesaurus* (cf. the following section) were predicted with the help of associationist learning theory. This theory states that the associative strength between two events $i$ and $j$ increases by a constant fraction of the maximally possible increment whenever these two events cooccur. This idea was applied to the coocurence of terms in the British National Corpus. The achieved results appear to be very promising. For 29 of the 100 stimulus words the model produced the primary associative response.

## 4 Tasks and Results

The main goal of our analysis was to find out to what extent free associations can be explained and predicted by statistical similarity measures computed by LSA. In order to address this issue, the workshop organizers have proposed the three tasks described below. Different training and test data sets containing association pairs were provided for each of the three tasks[5].

*Free associations* are the first words that come to the mind of a native speaker when he or she is presented a stimulus word. The degree of free association between a stimulus (*cue*) and a response (*target*) is quantified by the percentage of test subjects who produced *target* when presented with *cue*.

### 4.1 Method

For training we used 108M words from two British newspapers (*The Times*, *The Guardian*) of the years 1996 to 1998. Using the *Infomap* NLP toolkit[6], developed at Stanford University's CSLI, we generated a term×term co-occurrence matrix of size 80.000×3.000, closed-class words not occurring in the test data were disregarded. The vocabulary ($|V| = 80.000$) as well as the index terms ($|I| = 3.000$) were determined by corpus frequency, and terms occurring less than 24 times in the corpus were excluded from the vocabulary. We calculated several spaces for co-occurrence windows of $\pm 5, \pm 25, \pm 50, \pm 75$ words, respectively; the window did not cross article boundaries. The results presented in the following are obtained using the $\pm 75$-window space, if not mentioned otherwise. The matrix was reduced by SVD to 300 dimensions; term similarity was determined by measuring the truncated cosine of the angle between the corresponding term vectors. Since negative cosine values can occur but are meaningless for similarity measurements (i. e. terms having a negative similarity value are not more dissimilar than those having a value of 0), negative values are set to 0.

### 4.2 Discrimination

This task consists in discrimination between three classes of association strengths:

- the FIRST set – strongly associated cue-target pairs given by more than 50% of test subjects as first responses,
- the HAPAX set – cue-target pairs that were produced by a single test subject,
- the RANDOM set – random combinations of headwords from EAT that were never produced as a cue-target pair.

For each of the cue–target pairs, excluding those which contained terms not being present in our vocabulary, we have computed LSA similarity values. We obtained results for 300 of the 301 suggested pairs of the test data set, using a discrimination threshold of 0.23 between FIRST and HAPAX, and a threshold of 0.02 for discrimination between HAPAX and RANDOM, which showed to be optimal for the training data set. The following table shows the discrimination results for all classes considered[7]:

[5] The data sets are based on a database of English association norms, the Edinburgh Associative Thesaurus (EAT). Cf. also: *http://www.eat.rl.ac.uk/*.

[6] *http://infomap-nlp.sourceforge.net/*

[7] HOR stands for HAPAX or RANDOM; Accuracy = Right * 100 / (Right+Wrong).

| | Right | Wrong | Accuracy |
|---|---|---|---|
| FIRST (th=0.23) | 50 | 50 | 50% |
| HAPAX (th=0.02) | 63 | 32 | 68% |
| RANDOM | 68 | 17 | 78.2% |
| Total (F/H/R) | 181 | 119 | 60.33% |
| HoR | 189 | 11 | 94.5% |
| FIRST/HoR | 239 | 61 | 79.66% |

### 4.3 Correlation

The task is to predict free association strength (ranging from 0 to 1) for a given list of cue-target pairs, quantified by the proportion of test subjects that gave this target as a response to the stimulus cue. Pairs in the training and test set have been selected by stratified sampling so that association strength is uniformly distributed across the full range.

We have computed LSA similarity values for 239 of the 241 suggested pairs, achieving the *Pearson* correlation of 0.353 between the human scores and the LSA values; the *Kendall* correlation coefficient is 0.263. Both are significant with a $p$-value$<$0.01.

### 4.4 Response Prediction

In this task, models have to predict the most frequent responses for a given list of stimulus words. The data sets contain cue-target pairs with the association strength of the target response and the association strength of the second (unknown) response. The cues were selected from the EAT in such a way that the association strength of the dominant response must be $\geqslant$0.4, and at least three times as high as that of the second response. For the first response prediction we have computed the LSA similarity between cues and all terms in our vocabulary for 199 pairs from 201. The resulting average rank of the correct response is 51.89 (if the correct response is not among the suggested candidates, it is assigned rank 100 regardless of the number of suggestions). The distribution of the target ranks is as follows:

| **Target rank** | 1 | 2 | 3 | 4 | 5 | 6 | 7-99 | 100 |
|---|---|---|---|---|---|---|---|---|
| **Frequency** | 31 | 10 | 7 | 5 | 6 | 7 | 43 | 89 |

### 4.5 Co-occurrence Window

The size of the co-occurrence window on which the input matrix is based is a crucial factor establishing relatedness. Previous works using term$\times$term matrices employed rather small windows: Lund and Burgess (1996) used windows of $\pm 8$ words, Cederberg and Widdows (2003) used $\pm 15$ words and Rapp (2003) used a window of $\pm 2$ words only.

To get a better understanding of this parameter, we calculated models for different window sizes ($\pm 5, \pm 25, \pm 50, \pm 75$ words) and tested them on the above described tasks[8].

| | $\pm 5$ | $\pm 25$ | $\pm 50$ | $\pm 75$ |
|---|---|---|---|---|
| Correlation ($r$) | 0.254 | 0.326 | 0.347 | 0.354 |
| Disc. (Acc.) | 54.67 | 55.67 | 58.67 | 60.33 |
| Pred. (Av. Rank) | 62.61 | 54.11 | 52.69 | 51.89 |

The results for all three tasks are quite univocal: The performance improves with the size of the co-occurrence window. This is of course only a first and rather coarse-grained observation, but it indicates that this parameter deserves more attention in the application of distributional models.

## 5 Analysis of the Results

In this section, we will analyse our results by comparing the similarity values produced by LSA with the human scores.

### 5.1 Quantitative Analysis

#### 5.1.1 Correlation Analysis

As the reliability of a statistical analysis depends on the size of considered sample, in this section we examine not only the test set (of size 239) but the test and training sets altogether (of size 278). Since the distributions of human values both of the training and test sets are the same, the training values can be regarded as sampled from the same general population.

The calculated Pearson and Kendall correlation coefficients are close to those reported for the training set (see section 4), and are 0.353 and 0.263, correspondingly. Both are significant with $p < 0.01$. The Spearman correlation is 0.384 and is also significant. This confirms a significant monotone and, moreover, linear dependence between the human and LSA values.

As an initial step, let us visually examine figure 2 which depicts for each pair of terms (i) its human and LSA values against its ordinal number (rank),

[8] Due to computational restrictions we were not able to calculate co-occurrence matrices for windows larger than $\pm 75$.

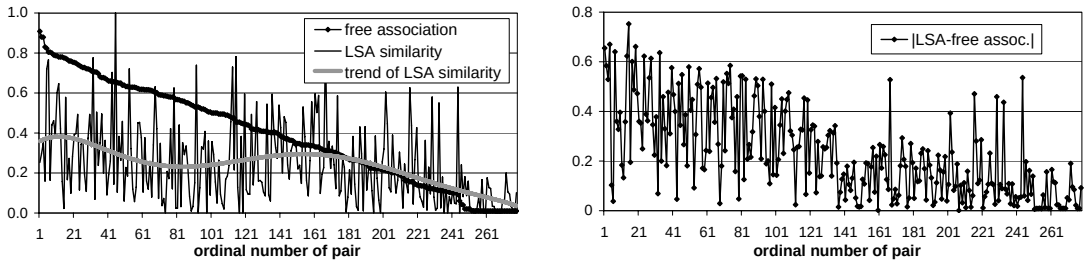



Figure 2: The human and LSA values (left) and their absolute differences (right).

(ii) the absolute difference between these two values, where the pairs are sorted by their human values. The behavior of the observed characteristics seems to differ at around the 136'th pair (human value $\approx$ 0.4). For the pairs with higher ranks (i.e. $\geq$ 0.4) the LSA values approximate human values on average and the difference between the LSA and human values looks like noise with constant variance. For the pairs with lower ranks the averaged LSA values show no clear dependence on the human values.

Based on these observations, we state the hypothesis of separation of the whole data set into two groups: *high human association group* $G_1$ with the human values $>$0.4 and *low human association group* $G_2$ with the values $<$0.4, where there is no correlation between LSA and human values in the first group $G_1$ in contrast to $G_2$.

For testing the stated hypothesis, we calculated the following characteristics between the human and the LSA values in each group and for the whole data set: (i) mean absolute difference, (ii) Pearson and Kendall correlation coefficients, and their significance and, furthermore, (iii) in each group we tested the hypothesis of randomness of the LSA values (using the *Abbe criterion*). The results are given in table 1; they show that in the high human association group $G_1$ there is no dependence between the human and the LSA values; moreover the mean absolute difference between these values is large (0.35), and it considerably exceeds the mean difference over the whole data set (0.23). At the same time, the results for the low human association group $G_2$ indicate a significant linear correlation producing small mean absolute difference (0.12).

Thus, we confirmed our hypothesis of difference between the groups $G_1$ and $G_2$. The existence of these groups demonstrates the fact that low association can be easily established, whereas correct estimation of high association strength seems to be complicated (cf. section 5.2). This observation conforms with the good discrimination results reported for the `RANDOM` group and bad results for the `FIRST` group. We would like to note that the Pearson and Kendall correlations between the LSA and human values calculated for the prediction data set (where all human values $\geq$0.4) are insignificant, which additionally confirms our hypothesis of independence between the LSA similarity and the human association values for pairs with a high latter value.

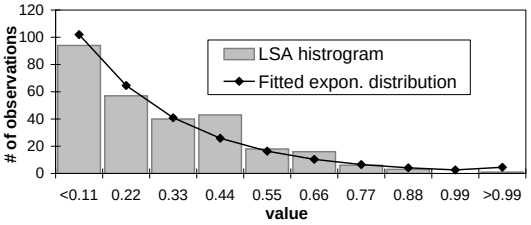



Figure 3: Histogram of the LSA values and the fitted exponential density.

The next interesting conclusion can be derived considering the histogram of LSA values (see figure 3; recall that the human values are uniformly distributed). Though the hypothesis of an exponential distribution of the LSA values is rejected with $p$-value$<$0.01, it becomes obvious that LSA underestimates association strength as compared with human scores. Moreover, all but one of the 12 pairs with the highest LSA values ($\geqslant$0.63) have high human values ($\geq$0.45), see table 2. Thus, it is possible that the pairs with high LSA values also have high human values but not vice versa.

### 5.1.2 Prediction Analysis

In the following a closer look is taken on the results of the prediction task. First, though for each

| Group | Mean abs. diff. | Pearson corr. | Kendall corr. | Randomness of LSA values |
|---|---|---|---|---|
| $G_1$ | 0.35 | 0.211 ($-$) | 0.172 ($+$) | Not rejected ($p$-value=0.43) |
| $G_2$ | 0.12 | 0.514 ($+$) | 0.393 ($+$) | Rejected ($p$-value=0.00) |
| $G_1 \cup G_2$ (whole data set) | 0.23 | 0.353 ($+$) | 0.263 ($+$) | Not rejected ($p$-value=0.07) |

Table 1: Intragroup properties, the signs $-$ or $+$ indicate significance of the correlation coefficients with $p$-value$<$0.01.

| cue | target | human value | LSA value |
|---|---|---|---|
| ha | ha | 0.66 | 1.00 |
| inland | revenue | 0.31 | 0.84 |
| four | five | 0.45 | 0.78 |
| question | answer | 0.71 | 0.78 |
| good | bad | 0.80 | 0.77 |
| grammar | school | 0.53 | 0.74 |
| below | above | 0.47 | 0.73 |
| daughter | son | 0.63 | 0.72 |
| vehicle | car | 0.82 | 0.72 |
| parish | church | 0.66 | 0.70 |
| boy | girl | 0.78 | 0.65 |
| sing | song | 0.60 | 0.63 |

Table 2: The 12 pairs with the highest LSA values.

cue the first association value is at least three time larger than the second association value (see section 4), we do not detect the same effect for LSA. The first and second LSA nearest neighbor values differ in only 1.1 times on average (vs. 8.6 times for the human values). It means that for every cue, the LSA similarity values of the most strongly related terms are very close. Second, it is interesting to note that in the human data when the large first association values ($\geq$0.65) increase, the second association values decrease, see figure 4. For LSA values no such effect is observed. A possible interpretation of this fact is that for humans a first strong association suppresses the others.

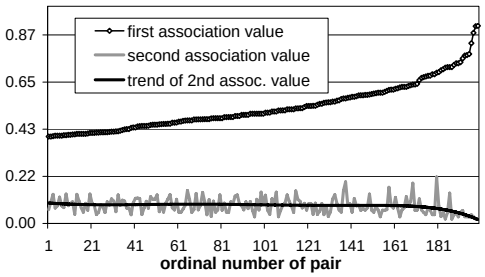



Figure 4: The association values for the prediction task.

### 5.1.3 Parts of Speech and Lexical-Semantic Relations

Wandmacher (2005) indicates that the quality of the term relations for a given cue may depend on its part of speech, e.g. LSA has found far more meaningful relations for nouns then for adjectives and verbs.[9] We have made different observations: For the correlation task the best result was achieved with adjectives (for adjectives the Pearson correlation is 0.68, for nouns 0.33, and for verbs 0.14) and for the prediction task there is no significant difference.[10]

We have also tagged relations for the prediction task test data set.[11] For syntagmatic relations the standard classification (cf. Cruse, 1986) was used: near-synonymy (association highlights a common part of the meaning, e.g. (*incorrect, wrong*)), opposition (association highlights an opposite part of the meaning, e.g. (*female, male*)), hypo-/hyperonymy (e.g. (*finch, bird*)), co-hyponymy (e.g. (*july, august*)), mero-/holonymy (e.g. (*deck, ship*)).[12] In order to estimate relations between terms belonging to different parts of speech we have distinguished following relations: collocation (e.g. (*wizard, oz*)), attribute-class relation (e.g. (*sugar, sweet*)), predicate-argument (e.g. (*eating, food*)), unsystematic association (which mostly express connection of terms via an implicit predicate, e.g. (*prefect, school*)). The information about the corresponding classes is given in table 3. We acknowledge that any tagging of this kind is highly subjective. Moreover, the number of pairs in some of our classes is definitely not enough to perform an analysis. Nevertheless, we decided to present these results, since the LSA values for the class of oppositions show dis-

[9] These results refer to German.

[10] Morphologically ambiguous words (e.g. *sting* or *shotgun*) were excluded from this analysis.

[11] The prediction data set was chosen because it contains meaningful associations only (cf. section 4).

[12] We do not mention relations that occurred less than 5 times in the data set, e.g. causation, presupposition etc.

tinctively better performance than others.

| relation | average rank | number of pairs |
|---|---|---|
| n.-syn. | 46.98 | 47 |
| oppos. | 24.42 | 31 |
| hypo. | 53.32 | 22 |
| mero. | 58.43 | 21 |
| co-hyp. | 40.50 | 6 |
| colloc. | 77.59 | 17 |
| attr.-cl. | 85.86 | 7 |
| pred.-arg. | 49 | 13 |
| assoc. | 62.65 | 31 |

Table 3: Average rank of targets and number of pairs in every class of relations for the prediction task data set.

### 5.2 Qualitative Analysis

In order to get a better understanding of what kind of information is reflected by LSA, we will take a look at some specific examples. First, we consider the term pairs that have got the highest LSA values ($\geq$0.63, see table 2). Obviously, LSA assigns a similarity of 1 to the pairs where cue and target are identical (e.g. (*ha, ha*)), whereas for human subjects such an association is not necessarily preferential. Then, LSA strongly associates oppositions, e.g. (*question, answer*), (*good, bad*), (*daughter, son*).[13] High LSA estimates for other semantic relations, such as collocations (e.g. (*inland, revenue*)), hyponyms (e.g. (*vehicle, car*)), co-hyponyms (e.g. (*four, five*)) etc., are found to be less regular and more corpus dependent.

The widest range of disagreements between LSA and human evaluations seems to be corpus-related. Since we have used a newspaper corpus, LSA extracted rather specific semantic neighbors for some of the terms. For example, terms from the food domain seem to stand out, possibly because of numerous commercial statements: e.g. for *fresh* the nearest neighbors are (*flavour*, 0.393), (*soup*, 0.368), (*vegetables*, 0.365), (*potato*, 0.362), (*chicken*, 0.36). Thus, the association (*fresh, lobster*) receiving a very low human value (0.01) is estimated by LSA at 0.2.

An interesting effect occurs for associations between some concepts and their salient properties, e.g. (*snow, white*) which is estimated at 0.408 by humans and at 0.09 by LSA. The nearest neighbors found by LSA for *snow* belong to the "weather forecast" domain: (*snowfalls*, 0.65), (*winds*, 0.624), (*weather*, 0.612), (*slopes*, 0.61), (*temperature*, 0.608). It is straightforward to suppose that since the feature of "being white" for snow is so natural for our language community, people do not talk much about it in newspapers.

Concerning word senses LSA is known to generate neighbors of the prominent meaning only and to suppress other domains (cf. Rapp, 2003; Wandmacher, 2005). This effect can lead both to over- and to underestimation in comparison with human values. For example the pair (*nurse, hospital*) gets a relatively high LSA value of 0.627 (while the human value is 0.156), because LSA has selected the nearest neighbors for *nurse* from only one (and very specific) domain: (*nurses*, 0.64), (*hospital*, 0.627), (*patient*, 0.597), (*doctors*, 0.554), (*patients*, 0.525). On the other hand, (*eve, adam*) receives only 0.024 by LSA (while the human value is 0.567), because LSA has selected another meaning for the homonym *eve*: (*christmas*, 0.657), (*festive*, 0.535), (*yuletide* 0.456), (*festivities*, 0.453), (*presents*, 0.408).

Besides the already mentioned effects we have noticed some more regularities. It is often the case (for 9 out of 22 collocations in the correlation task data sets) that LSA assigns a low value ($< 0.1$) to term pairs forming a collocation, e.g. (*peg, clothes*, hum.: 0.225, LSA: 0.001), (*shotgun, wedding*, hum.: 0.402, LSA: 0.06), (*core, apple*, hum.: 0.776, LSA: 0.023). The problem here is that the terms in such collocations have no other overlap in their meanings (e.g. the nearest neighbors for *shotgun* are (*gun*, 0.536), (*pistol*, 0.506), (*shooting*, 0.463), (*shotguns*, 0.447), (*firearms*, 0.445), which most of the time have nothing to do with weddings) and the given collocations are rare in the corpus.

As for the auxiliary words (like prepositions, pronouns and conjunctions), LSA produces rather unstable results. A general observation is that the association strength for such pairs is mostly underestimated because of their low specificity (cf. Wandmacher, 2005). However, there is not enough data in the considered data sets to investigate this effect.

It is worth reminding that the semantic similarity estimated by LSA is symmetric, whereas it is obviously not the case for human scores. For example

[13] 15 from 19 oppositions found in the correlation task data sets have got LSA values >0.22.

the association of terms *wrong* and *right* which is assigned an LSA value of 0.493, is estimated by humans at 0.717 in the direction from *wrong* to *right* and at 0.42 in the opposite direction.

## 6 Discussion and Conclusion

In this paper, we have described the results of three tasks[14] in order to get an understanding of the relationships between human free associations and similarity measures produced by LSA. In reply to the title's question, we have to report that no strong correlation between human associations and LSA similarity could be discovered. Likewise, our prediction results are relatively bad (as compared to those by Wettler et al. 2005). However, Wettler et al. (2005) have used a lemmatized corpus, which is not the case for our study. The effect of lemmatization on the training data should be investigated in more detail.

We did however investigate the effect of the size of the co-occurrence window, and we have found larger windows (of around $\pm 75$ words) to provide significantly better results in all tasks than windows of smaller sizes.

Another effect that we have observed is that LSA estimates for weakly associated terms are much closer to those of humans than for strongly associated terms. Then, we have reported a regular underestimation by LSA. We have also pointed out the fact that the clear preference for one association in human responses is not established by LSA; the average distance between the first and the second LSA neighbor is much lower (section 5.1.2).

Furthermore, we have added some comments on the LSA similarity estimates for different parts-of-speech and kinds of lexical relations. Finally, we have tried to establish some qualitative regularities in the disagreements between LSA and human estimations (section 5.2).

For further investigation it will be interesting to look not only at the first words coming into the mind of a subject after being presented a cue but also at further associations. This will probably help to understand to which domains do these associations belong and to compare these domains with the domains found for the cue by LSA.

[14]The files containing our results can be found at *http://www.ikw.uos.de/∼twandmac/FA-Results-WOA.zip*.

## References


S. Cederberg and D. Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy. In *Proc. of CoNNL*.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, G.B.

S. C. Deerwester, S. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by Latent Semantic Analysis. *American Society of Information Science*, 41(6):391–407.

S. Dumais. 1995. Latent Semantic Indexing (LSI): TREC-3 Report. In D. Harman, editor, *Proc. of TREC-3*, volume 500-226, pages 219–230. NIST Special Puclication.

T. K. Landauer and S. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner. 1997. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto and P. Langley, editors, *Proc. of the 19th annual meeting of the Cognitive Science Society*, pages 412–417, Mawhwah, NJ. Erlbaum.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments and Computers*, 28(2):203–208.

R. Rapp. 2003. Word Sense Discovery Based on Sense Descriptor Dissimilarity. In *Proc. of the Machine Translation Summit IX*, New Orleans.

G. Salton and M. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, NY.

H. Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

D. Wade-Stein and E. Kintsch. 2003. Summary street: Interactive computer support for writing. Technical report, University of Colorado.

T. Wandmacher. 2005. How semantic is Latent Semantic Analysis? In *Proc. of TALN/RECITAL'05*, Dourdan, France, 6-10 june.

M. Wettler, R. Rapp, and P. Sedlmeier. 2005. Free word associations correspond to contiguities between words in texts. *Quantitative Linguistics*, 12(2-3):111–122.