A Comparison of Windowless and Window-Based Computational Association Measures as Predictors of Syntagmatic Human Associations

Justin Washtell
School of Computing
University of Leeds
washtell@comp.leeds.ac.uk

Katja Markert
School of Computing
University of Leeds
markert@comp.leeds.ac.uk

Abstract

Distance-based (windowless) word association measures have only very recently appeared in the NLP literature and their performance compared to existing windowed or frequency-based measures is largely unknown. We conduct a large-scale empirical comparison of a variety of distance-based and frequency-based measures for the reproduction of syntagmatic human assocation norms. Overall, our results show an improvement in the predictive power of windowless over windowed measures. This provides support to some of the previously published theoretical advantages and makes windowless approaches a promising avenue to explore further. This study also serves as a first comparison of windowed methods across numerous human association datasets. During this comparison we also introduce some novel variations of window-based measures which perform as well as or better in the human association norm task than established measures.

1 Introduction

Automatic discovery of semantically associated words has attracted a large amount of attention in the last decades and a host of computational association measures have been proposed to deal with this task (see Section 2). These measures traditionally rely on the co-ocurrence frequency of two words in a corpus to estimate a relatedness score. There has been a recent emergence of distance-based language modelling techiques in NLP (Savicki and Hlavacova, 2002; Terra and Clarke, 2004) in which the number of tokens separating words is the essential quantity. While some of this work has considered distance-based alternatives to conventional association measures (Hardcastle, 2005; Washtell, 2009), there has been no principled empirical evaluation of these measures as predictors of human association. We remedy this by conducting a thorough comparison of a wide variety of frequency-based and distance-based measures as predictors of human association scores as elicited in several different *free word association tasks*.

In this work we focus on first-order association measures as predictors of *syntagmatic* associations. This is in contrast to second and higher-order measures which are better predictors of paradigmatic associations, or word *similarity*. The distinction between syntagmatic and paradigmatic relationship types is neither exact nor mutually exclusive, and many paradigmatic relationships can be observed syntagmatically in the text. Roughly in keeping with (Rapp, 2002), we hereby regard *paradigmatic* assocations as those based largely on word similarity (i.e. including those typically classed as synonyms, antonyms, hypernyms, hyponyms etc), whereas *syntagmatic* associations are all those words which strongly invoke one another yet which cannot readily be said to be similar. Typically these will have an identifiable semantic or grammatical relationship (meronym/holonym: *stem – flower*, verb/object: *eat – food* etc), or may have harder-to-classify topical or idiomatic relationships (*family – Christmas*, *rock – roll*).

We will show in Section 3.2 that syntagmatic relations by themselves constitute a substantial 25-40% of the strongest human responses to cue words. Although the automatic detection of these assocations in text has received less attention than that of paradigmatic associations, they are nonetheless important in applications such as the resolution of bridging anaphora (Vieira and Poesio, 2000).[1] Furthermore, first-order associations

[1] where for example resolving *my house – the windows* to *the windows of my house* can be aided by the knowledge that windows are often (syntagmatically) associated with houses.

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 628–637,
Singapore, 6-7 August 2009.

are often the basis of higher-order vector word-space models used for predicting paradigmatic relationships: i.e. through the observation of words which share similar sets of syntagmatic associations. Therefore improvements made at the level we are concerned with may reasonably be expected to carry through to applications which hinge on the identification of paradigmatic relationships.

After a discussion of previous work in Section 2, we formulate the exact association measures and parameter settings which we compare in Section 3, where we also introduce the corpora and human association sets used. Then, by using evaluations similar to those described in (Baroni et al., 2008) and by Rapp (2002), we show that the best distance-based measures correlate better overall with human association scores than do the best window based configurations (see Section 4), and that they also serve as better predictors of the strongest human associations (see Section 5).

2 Related Work

Measures based on co-ocurrence frequency. The standard way of estimating the syntagmatic association of word pairs in a corpus is to examine the frequency of their co-occurence, and then usually to compare this to some expected frequency. There are a host of measures which exist for this purpose. After raw co-occurrence frequency, the simplest and most prevalent in the literature is Pointwise Mutual Information, famously used by Church (1989) (as the *association ratio*). This is defined as the log of the ratio of the observed co-occurrence frequency to the frequency expected under independence. More sophisticated and statistically-informed measures include t-Score, z-Score, Chi-Squared and Log-Likelihood (see Evert (2005) for a thorough review).

All of these measures have in common that they require co-occurrence frequency to be specified, and therefore require some definition of a region within which to count co-occurrences. This region might be the entirety of a document at one extreme, or a bigram at the other. A versatile and hugely popular generalised approach is therefore to consider a "'window"' of w words, where w can be varied to suit the application. Unsurprisingly, it has been found that this is a parameter which can have a significant impact upon performance (Yarowsky and Florian, 2002; Lamjiri et al., 2004; Wang, 2005). While choosing an optimum window size for an application is often subject to trial and error, there are some generally recognized trade-offs between small versus large windows, such as the impact of data-sparseness, and the nature of the associations retrieved (Church and Hanks, 1989; Church and Hanks, 1991; Rapp, 2002)

Measures based on distance between words in the text. The idea of using distance as an alternative to frequency for modelling language has been touched upon in recent literature (Savicki and Hlavacova, 2002; Terra and Clarke, 2004; Hardcastle, 2005). Washtell (2009) showed that it is possible to build distance-based analogues of existing syntagmatic association measures, by using the notions of mean and expected distance rather than of frequency. These measures have certain theoretical qualities - notably scale-independence and relative resilience to data-sparseness - which might be expected to provide gains in tasks such as the reproduction of human association norms from corpus data. The specific measure introduced by Washtell, called Co-Dispersion, is based upon an established biogeographic dispersion measure (Clark and Evans, 1954). We provide a thorough empirical investigation of Co-Dispersion and some of its derivatives herein.

Measures based on syntactic relations. Several researchers (Lin, 1998; Curran, 2003; Pado and Lapata, 2007) have used word space models based on grammatical relationships for detecting and quantifying (mostly paradigmatic) word associations. In this paper, we will not use syntactic relation measures for two main reasons. Firstly these depend on the availability of parsers, which is not a given for many languages. Secondly, this may not be the most pertinent approach for predicting human free associations, in which certain observed relationsips can be hard to express in terms of syntactic relationships.

3 Methodology

Similar to (Rapp, 2002; Baroni et al., 2008, among others), we use comparison to human assocation datasets as a test bed for the scores produced by computational association measures. An alternative might be to validate scores against those derived from a structured resource such as WordNet.

Table 1: Human association datasets

Name	Origin	Cues	Respondents
Kent	Kent & Rosanoff (1910)	100	$\sim$ 1000
Minnesota	Russell & Jenkins (1954)	100	$\sim$ 1000
EAT	Kiss et al (1973)	8400	100
Florida	Nelson et al (1980)	5019	$\sim$ 140

However, relatedness measures for WordNet are many and varied and are themselves the subject of evaluation (Pedersen et al., 2004). Although human association datasets have their own peculiarities, they do at least provide some kind of definite Gold Standard. Yet another alternative might be to incorporate our computational association scores into an application (such as anaphora resolution), and measure the performance of that, but noise from other submodules would complicate evaluation. We leave such extensions to possible future work.

We use evaluations similar to those used before (Rapp, 2002; Pado and Lapata, 2007; Baroni et al., 2008, among others). However, whereas most existing studies use only one dataset, or hand-selected parts thereof, we aim to evaluate measures across four different human datasets. In this way we hope to get as unbiased a picture as possible.

3.1 Association data

The datasets used are listed in Table 1. While the exact experimental conditions may differ, the datasets used were all elicited using the same basic methodology: by presenting individual words (*cues*) to a number of healthy human subjects and asking in each case for the word that is most immediately or strongly evoked. An association score can then be derived for each cue/response pair in a dataset by dividing the number of participants providing a given response by the number who were presented with the cue word. In Table 1, *respondents* refers to the number of people from whom a response was solicited for each cue word in a study (this is not to be confused with the number of unique responses).

Of these four datasets, one (Kent & Rosanoff) appears not to have been previously used in any peer-reviewed study of corpus-derived lexical association. It is worth noting that some of these datasets are quite dated, which might affect correlations with corpus-derived scores, as culture and contemporary language have a fundamental impact upon the associations humans form (White and Abrams, 2004).

3.2 Frequency of Syntagmatic Associations

To verify that strong human associations do include a large number of syntagmatic associations, we manually annotated all pairs consisting of a cue and its strongest human response in the `Minnesota` and `Kent` datasets as expressing either a syntagmatic or a paradigmatic relationship. The overall set to be annotated consisted of 200 pairs.

Annotators were given short (half-page) guidelines on syntagmatic and paradigmatic assocations, stating that very similar items (including hyponyms/hypernyms) as well as antonyms were to be judged as paradigmatic whereas words that do not fulfil this criterion are to be judged as syntagmatic. The two annotators were the authors of this paper (one native and one near-native speaker). After independent annotation, agreement was measured at a percentage agreement of 91/93% and a kappa of 0.80/0.82 for `Minnesota` and `Kent`, respectively. Therefore, the distinction can be made with high reliability.

Overall, 27/39% of the human responses were syntagmatic in the Kent/Minnesota datasets, showing that syntagmatic relations make up a large proportion of even the strongest human associations.

3.3 Corpora

We use two randomized subsets of the British National Corpus (BNC), a representative 100 million word corpus of British English (Burnard, 1995): one 10 million word sample, and a 1 million word sample. A vocabulary of approximately 33,000 word types was used. The selected words included approximately 24,000 word types comprising all cue and target words from the multiple sets of human association norms to be used in this study. To these were added a top-cut of the most frequent words in the BNC, until the total of 33,000 word types was reached. The resultant set included ap-

proximately the 24,000 most common word types in the BNC, with the remaining 9000 words types therefore comprising relatively uncommon words taken from the human associative responses.

The words included in the vocabulary accounted for over 94.5% of tokens in the corpus. Although statistics for the remaining word types in the BNC were not gathered, their corresponding tokens were left in the corpus so that these could be properly accounted for when calculating distances and window spans.

In order to maximize matching between word types in the corpus and association norms, all words in both were normalized by converting to lower-case and removing hyphens and periods. Words consisting entirely of numerals, or numerals and punctuation, and all "phrasal" associative responses (those containing spaces) were discarded. The 33,000 word count was satisfied after making these normalizations.

In order to maximize the variety of the language in the samples, the subsets were built from approximately the first 2000 words only of each randomly selected document from the BNC (a similar strategy to that used in constructing the 1 million word Brown Corpus). Both a 10 million word and a 1 million word sample were constructed in this fashion, allowing us to also examine the effects of varying corpus size and content.

3.4 Association measures used

3.4.1 Frequency-based measures

In the following, x is the cue word and y a (possible) response word. Therefore $p(x)$ is the probability of observing x, and $p(\bar{x})$ refers to the probability of not observing x.

Pointwise Mutual Information (hereonin PMI) was introduced in Section 2. For ranking word pairs, we can neglect the usual logarithm.

$$PMI = \frac{p(x,y)}{p(x)p(y)}$$

PMI is infamous for its tendency to attribute very high association scores to pairs involving low frequency words, as the denominator is small in such cases, even though the *evidence* for association in such cases is also small. This can result in some unlikely associations. There exist a number of alternative measures which factor in the amount of evidence to give an estimate of the *significance of association*. One popular and statistically appealing such measure is Log-Likelihood (LL) (Dunning, 1993). LL works on a similar principle to PMI but considers the ratio of the observed to expected co-occurrence frequencies for *all contingencies* (i.e. including those where the words do not co-occur). LL, as it most frequently appears in the literature, is not actually a measure of positive association: it also responds to significant *negative* association. Therefore LL is arguably not suited to the task in hand. Krenn & Evert (2001) experiment with one-tailed variants of LL and Chi-Squared measures, although they do not define these variants. Here, we construct a one-tailed variant of LL by simply reversing the signs of the terms which respond to negative association.

$$\begin{aligned} LL_{1tail} &= p(x,y)\log\frac{p(x,y)}{p(x)p(y)} - p(x,\bar{y})\log\frac{p(x,\bar{y})}{p(x)p(\bar{y})} \\ &- p(\bar{x},y)\log\frac{p(\bar{x},y)}{p(\bar{x})p(y)} + p(\bar{x},\bar{y})\log\frac{p(\bar{x},\bar{y})}{p(\bar{x})p(\bar{y})} \end{aligned}$$

LL does not have a clear analogue amongst the distance-based measures (introduced in Section 3.4.2), whereas PMI for instance does. We therefore construct variants of PMI and other measures which take the *amount of evidence* into account in a way which can be directly reproduced in the distance domain. For this we borrow from Sackett (2001) who asserts that, all other things being equal, statistical significance is proportional to the square root of the sample size. There are a number of ways one might quantify sample size. We take a consistent approach across the various distance-based and frequency-based measures: we assume sample size to be equivalent to the lesser of the frequencies of the two words as this represents the total number of words available for pairing, with fewer observed pairs therefore being considered to constitute negative evidence.

$$PMI_{sig} = \sqrt{\min(p(x),p(y))}\frac{p(x,y)}{p(x)p(y)}$$

All of the above measures are symmetric. Human associative responses however are not (Michelbacher et al., 2007): a person's tendency to give the response *because* to the cue *why* does not necessarily reflect their tendency to give the response *why* to the cue *because*.[2] A simple asymmetric association measure is conditional probability (CP)

[2]This notion of assymmetry is not to be confused with

- the probability of observing the response, given that the cue has already occurred.

$$CP = p(y|x) = \frac{p(x,y)}{p(x)}$$

CP suffers from the fact that it does not account at all for the general frequency of the response word. It therefore tends to favour very frequent words, such as function words. An obvious solution would be to divide CP by the frequency of the response word, however this merely results in PMI which is symmetric. By multiplying CP with PMI (and taking the root, to simplify) we obtain a measure which is asymmetric yet does not overtly favour frequent response words.[3] We refer to this herein as Semi-Conditional Information (SCI).

$$SCI = \frac{p(x,y)}{p(x)\sqrt{p(y)}}$$

We also explore variants of both CP and SCI with the additional significance correction presented for PMI_{sig}. These can be easily inferred from the formulae above.

3.4.2 Distance-based Measures

Co-Dispersion (herein CD), introduced by Washtell (2009), is defined as the ratio of the mean observed distance to the expected distance, where the expected distance is derived from the frequency of the more frequent word type. Distance refers to the number of tokens separating an occurrence of one word and the *nearest* occurrence of another word. Pairs spanning an intervening occurrence of *either* word type or a document boundary are not considered. Note that here we specify only the generalised mean M, as we wish to keep the specific choice of mean as a parameter to be explored,

$$CD = \frac{1/\max(p(x),p(y))}{M(dist_{xy1} \ldots dist_{xyn})}$$

where $dist_{xyi}$ is i^{th} observed distance between some occurrence of word type x and its nearest preceding or following occurrence of word type y, and n is the total number of such distances observed (being at most equal to the frequency of the rarer word).

In cases where many occurrences of the less frequent word were not able to be paired, raw CD gives midleading results. This is because unpairable words themselves provide useful negative evidence which CD ignores. A more appropriate measure can be formed in which the mean distance is calculated using the frequency of the less frequent word, regardless of whether this many distances were actually observed. This gives us Neutrally-Weighted Co-Dispersion (NWCD). Note that for convenience, we keep the standard definition of the mean and introduce a correction factor instead.

$$NWCD = \frac{n}{\min(p(x),p(y))} \frac{1/\max(p(x),p(y))}{M(dist_{xy1} \ldots dist_{xyn})}$$

An asymmetric association measure can be formed in a similar manner. Instead of calculating the mean using the frequency of the less frequent word as described above, we explicitly use the frequency of the cue word (which in some cases may actually *exceed* the number of distances observed). This gives us Cue-Weighted Co-Dispersion (CWCD).

$$CWCD = \frac{n}{p(x)} \frac{1/\max(p(x),p(y))}{M(dist_{xy1} \ldots dist_{xyn})} \quad (1)$$

In addition to these measures, we also explore significance-corrected forms NWCD_{sig} and CWCD_{sig}, by introducing the same sample size term employed by PMI_{sig}, CP_{sig} and SCI_{sig}. Again, these can readily be inferred from the existing formulae in the above two sections.

3.5 Co-occurrence Parameters

For frequency-based co-occurrence statistics, the principle parameter is the window size. We will use five window sizes separated by a constant scaling factor, chosen so as to span those most commonly encountered in the literature, with some extension towards the upper end. We use w to represent this parameter, with $w = 2$ implying a window size of +/-2. The parameter values explored

that of *direction* in the text. While the two may correlate, one can find ample counter-examples: *jerky* triggers *beef* more strongly than *beef* triggers *jerky*.

[3] Note that Wettler & Rapp (1993) introduced a more general asymmetric measure for predicting human associations, by employing an exponent parameter to $p(y)$. Our formulisation is equivalent to their measure with an exponent of 0.5, whereas they found an exponent of 0.66 to be most effective in their empirical study. Exponents of 0 and 1 result in CP and PMI respectively.

are $w = 2$, $w = 10$, $w = 50$, $w = 250$ and $w = 1250$. We examine such large window sizes so as to give a fairer comparison with the distance approach which is not bounded by a window, and in acknowledgement of the fact that the entire document as context has been used with some success in other application areas (most notably information retrieval).

For distance-based statistics, the principle parameter is the function via which the various observed distances between tokens are reduced to a single mean value. In this investigation we will explore five means. These are the power means with exponents (which herein we refer to as m) ranging from -2 to +2. These give us the quadratic mean or RMS ($m = 2$), the arithmetic mean ($m = 1$), the geometric mean ($m = 0$), the harmonic mean ($m = -1$), and the inverse quadratic mean ($m = -2$).

4 Task I: Correlations on word pairs

One of the ESSLLI Workshop shared tasks (Baroni et al., 2008) required the evaluation of correlation between a small, manually selected subset of human cue-response scores from the `EAT` dataset and automatic scores for the same word pairs. Here, tather than focusing on word pairs which meet certain grammatical and frequency criteria we test on all pairs. For the `EAT` and `Florida` datasets, this amounts to many tens of thousands of cue-response pairs. Although this makes the task of correlation harder, it means we can attribute a great deal of statistical significance to the results and make our observations as general as possible.

4.1 Evaluation Measures, Upper Bounds and Baselines

For evaluating agreement between corpus-derived associations and human associations, we use Spearman's Rank correlation. This is appropriate because we are primarily interested in the relative ranking of word pair associations (in order to predict particularly strong responses, for example). Although some studies have used Pearson's correlation, the various association measures explored here are not linear within each another and it would be inappropriate to evaluate them under the assumption of a linear relationship with the human norms.

Two of the human datasets, `Kent` and `Minnesota`, though collected independently, are based on the same set of 100 cue words established by Kent (1910). Therefore by performing a rank correlation of these two datasets with one another, (each of which was produced by pooling the responses of some 1000 people) we can get a useful upper-bound for correlations: if a computer-based system were to exceed this upper-bound in correlations with either dataset, then we would need to suspect it of over-fitting.

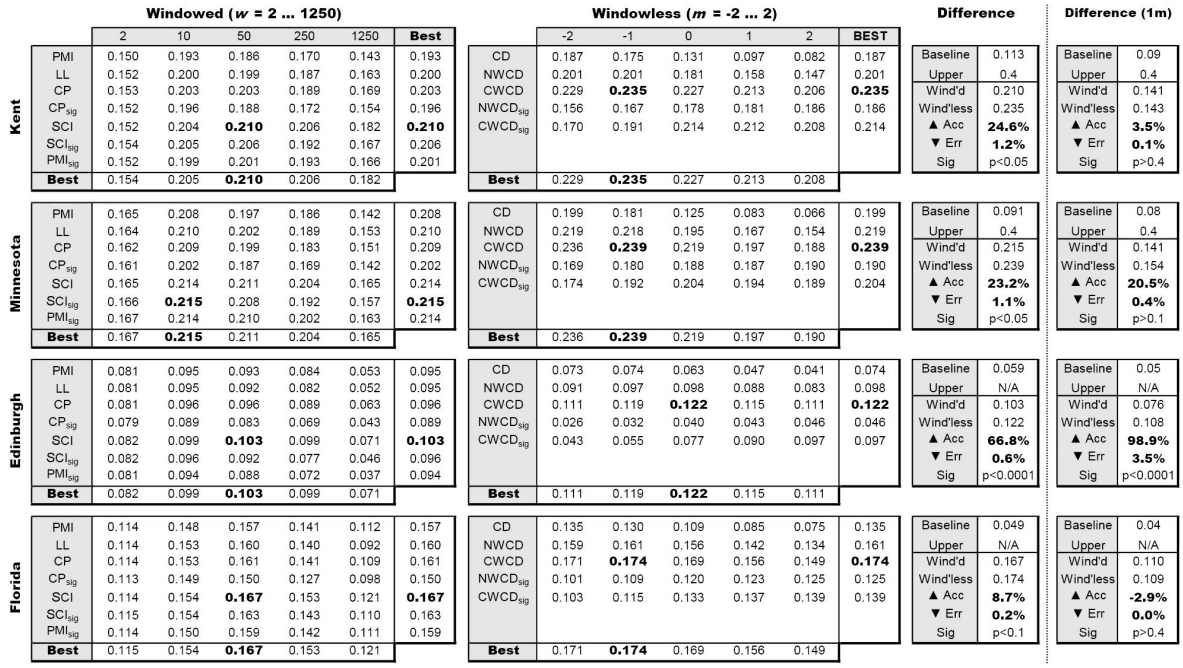
As a baseline, we use the corpus frequency of the response word. The simple assumption is that the more frequent a word is, the more likely it is to appear as a human response independent of the cue given. This is also the simplest formulation which does not assign equal scores to the various possible responses, and which is therefore capable of producing a rank-list of predictions.

4.2 Task I Results

Figure 1 shows the Spearman's rank correlation co-efficients across all paramaterisations of all association measures (frequency-based on the left, and distance-based on the right), with each human dataset, for the 10 million word corpus. Emboldened are the best performing windowed and windowless configurations for each dataset. The difference of these figures over the baseline is highly significant ($p < 0.0001$ in most cases). The panels to the right show summary statistics for these figures, and for the 1 million word corpus (for which full figures are not included owing to space limitations). These statistics include the performance of the baseline, where relevant the estimated upper-bound (see Section 4.1), and the difference in performance of the distance-based method over the window-based. The accuracy and error figures are based on the co-efficients of determination (r^2) and are expressed both as a relative improvement in accuracy (how much closer (r^2) is to 1 under the distance-based approach) and reduction in error (how much further r^2 is from zero). Also the significance of the difference in the r values is given.

4.3 Discussion

The two-way Spearman's rank correlations between the `Kent` and `Minesota` datasets suggested an upper bound of $r = 0.4$. In theory, a large proportion of this agreement is accounted for by paradigmatic associations which we are not likely to fully reproduce with these first-order measures. By this standard, the general levels of



Kent

Windowed (w = 2 ... 1250)	2	10	50	250	1250	Best
PMI	0.150	0.193	0.186	0.170	0.143	0.193
LL	0.152	0.200	0.199	0.187	0.163	0.200
CP	0.153	0.203	0.203	0.189	0.169	0.203
CP_{sig}	0.152	0.196	0.188	0.172	0.154	0.196
SCI	0.152	0.204	**0.210**	0.206	0.182	**0.210**
SCI_{sig}	0.154	0.205	0.206	0.192	0.167	0.206
PMI_{sig}	0.152	0.199	0.201	0.193	0.166	0.201
Best	0.154	0.205	**0.210**	0.206	0.182	

Windowless (m = -2 ... 2)	-2	-1	0	1	2	BEST
CD	0.187	0.175	0.131	0.097	0.082	0.187
NWCD	0.201	0.201	0.181	0.158	0.147	0.201
CWCD	0.229	**0.235**	0.227	0.213	0.206	**0.235**
$NWCD_{sig}$	0.156	0.167	0.178	0.181	0.186	0.186
$CWCD_{sig}$	0.170	0.191	0.214	0.212	0.208	0.214
Best	0.229	**0.235**	0.227	0.213	0.208	

Difference	
Baseline	0.113
Upper	0.4
Wind'd	0.210
Wind'less	0.235
▲ Acc	**24.6%**
▼ Err	**1.2%**
Sig	p<0.05

Difference (1m)	
Baseline	0.09
Upper	0.4
Wind'd	0.141
Wind'less	0.143
▲ Acc	**3.5%**
▼ Err	**0.1%**
Sig	p>0.4

Minnesota

Windowed (w = 2 ... 1250)	2	10	50	250	1250	Best
PMI	0.165	0.208	0.197	0.186	0.142	0.208
LL	0.164	0.210	0.202	0.189	0.153	0.210
CP	0.162	0.209	0.199	0.183	0.151	0.209
CP_{sig}	0.161	0.202	0.187	0.169	0.142	0.202
SCI	0.165	0.214	0.211	0.204	0.165	0.214
SCI_{sig}	0.166	**0.215**	0.208	0.192	0.157	**0.215**
PMI_{sig}	0.167	0.214	0.210	0.202	0.163	0.214
Best	0.167	**0.215**	0.211	0.204	0.165	

Windowless (m = -2 ... 2)	-2	-1	0	1	2	BEST
CD	0.199	0.181	0.125	0.083	0.066	0.199
NWCD	0.219	0.218	0.195	0.167	0.154	0.219
CWCD	0.236	**0.239**	0.219	0.197	0.188	**0.239**
$NWCD_{sig}$	0.169	0.180	0.188	0.187	0.190	0.190
$CWCD_{sig}$	0.174	0.192	0.204	0.194	0.189	0.204
Best	0.236	**0.239**	0.219	0.197	0.190	

Difference	
Baseline	0.091
Upper	0.4
Wind'd	0.215
Wind'less	0.239
▲ Acc	**23.2%**
▼ Err	**1.1%**
Sig	p<0.05

Difference (1m)	
Baseline	0.08
Upper	0.4
Wind'd	0.141
Wind'less	0.154
▲ Acc	**20.5%**
▼ Err	**0.4%**
Sig	p>0.1

Edinburgh

Windowed (w = 2 ... 1250)	2	10	50	250	1250	Best
PMI	0.081	0.095	0.093	0.084	0.053	0.095
LL	0.081	0.095	0.092	0.082	0.052	0.095
CP	0.081	0.096	0.096	0.089	0.063	0.096
CP_{sig}	0.079	0.089	0.083	0.069	0.043	0.089
SCI	0.082	0.099	**0.103**	0.099	0.071	**0.103**
SCI_{sig}	0.082	0.096	0.092	0.077	0.046	0.096
PMI_{sig}	0.081	0.094	0.088	0.072	0.037	0.094
Best	0.082	0.099	**0.103**	0.099	0.071	

Windowless (m = -2 ... 2)	-2	-1	0	1	2	BEST
CD	0.073	0.074	0.063	0.047	0.041	0.074
NWCD	0.091	0.097	0.098	0.088	0.083	0.098
CWCD	0.111	0.119	**0.122**	0.115	0.111	**0.122**
$NWCD_{sig}$	0.026	0.032	0.040	0.043	0.046	0.046
$CWCD_{sig}$	0.043	0.055	0.077	0.090	0.097	0.097
Best	0.111	0.119	**0.122**	0.115	0.111	

Difference	
Baseline	0.059
Upper	N/A
Wind'd	0.103
Wind'less	0.122
▲ Acc	**66.8%**
▼ Err	**0.6%**
Sig	p<0.0001

Difference (1m)	
Baseline	0.05
Upper	N/A
Wind'd	0.076
Wind'less	0.108
▲ Acc	**98.9%**
▼ Err	**3.5%**
Sig	p<0.0001

Florida

Windowed (w = 2 ... 1250)	2	10	50	250	1250	Best
PMI	0.114	0.148	0.157	0.141	0.112	0.157
LL	0.114	0.153	0.160	0.140	0.092	0.160
CP	0.114	0.153	0.161	0.141	0.109	0.161
CP_{sig}	0.113	0.149	0.150	0.127	0.098	0.150
SCI	0.114	0.154	**0.167**	0.153	0.121	**0.167**
SCI_{sig}	0.115	0.154	0.163	0.143	0.110	0.163
PMI_{sig}	0.114	0.150	0.159	0.142	0.111	0.159
Best	0.115	0.154	**0.167**	0.153	0.121	

Windowless (m = -2 ... 2)	-2	-1	0	1	2	BEST
CD	0.135	0.130	0.109	0.085	0.075	0.135
NWCD	0.159	0.161	0.156	0.142	0.134	0.161
CWCD	0.171	**0.174**	0.169	0.156	0.149	**0.174**
$NWCD_{sig}$	0.101	0.109	0.120	0.123	0.125	0.125
$CWCD_{sig}$	0.103	0.115	0.133	0.137	0.139	0.139
Best	0.171	**0.174**	0.169	0.156	0.149	

Difference	
Baseline	0.049
Upper	N/A
Wind'd	0.167
Wind'less	0.174
▲ Acc	**8.7%**
▼ Err	**0.2%**
Sig	p<0.1

Difference (1m)	
Baseline	0.04
Upper	N/A
Wind'd	0.110
Wind'less	0.109
▲ Acc	**-2.9%**
▼ Err	**0.0%**
Sig	p>0.4

Figure 1: Correlations for window-based and windowless measures on a 10 million word corpus

correlation seen here (for these datasets $r = 0.235$ and $r = 0.239$ respectively) seem very reasonable.

What is immediately clear from Figure 1 is that, for the range of parameters tested here, we see a relatively small but statistically significant improvement across four of the five datasets when adopting the distance-based approach.

The correlations are unsurprisingly lower across the board for the much smaller 1 million word corpus. Here, the best distance-based measure statistically significantly outperforms the best window-based one (with a significance level of $p < 0.0001$) on one out of four datasets, while the differences are not great enough to be considered statistically significant on the other three datasets. There is therefore some evidence that the benefits observed with the larger corpus hold in the presence of limited data, which is in support of the general theory that distance-based methods capture more information from the corpus at the co-occurrence level (Washtell, 2009). It remains clear, however, that no method is presently a substitute for using a larger corpus.

In terms of optimum configurations, we find that for the frequency-based approach with the larger corpus, a window size of around +/-10 to +/-50 words more or less consistently produces the best results, irrespective of association the measure. Interestingly on the small corpus the tendency appears to be towards a somewhat larger window size than with the larger corpus. This may be related to the larger windows' increased resilience to data-sparseness. Somewhat surprisingly, we also see that our assymmetric association measures SCI and SCI_{sig} perform the best overall amongst the windowed measures, largely irrespective of the window or corpus, size.

In the large corpus, the best distance-based measure is the asymmetric $CWCD$, with the significance corrected measure $CWCD_{sig}$ showing greater strength in the small corpus: perhaps, again, for its improved reliability in the presence of very low-frequency data. The optimum mean for the distance-based parameterisations is somewhere around $m = -1$ (the harmonic) to $m = 0$ (the geometric). We find this unsurprising as the typical distribution of inter-word distances in a corpus is heavily skewed towards the smaller distances - indeed even a random corpus exhibits this characteristic with the distances following a geometric distribution.

5 Task II: Agreement with strongest human associations

The correlation evalation presented considers all word pairs present in the human datasets. However, human association norms tend to contain a very long tail of *hapax legomena* - responses which were given by only one individual. Such responses are extremely difficult for corpus-based

association measures to predict, and given that there is so little consensus amongst human respondents over these items, it is probably not particularly useful to do so. Rather, it might be most useful to predict common or majority human responses.

5.1 Evaluation measure and Upper Bound

For the strongest human response to each cue in the human datasets, its rank was calculated amongst all $33,000$ possible responses to that cue, according to each association measure and parameterisation. Where there were tied scores for various responses, a median rank was assigned. As a rough upper bound, we would be impressed by a computer system which was able to predict the most popular human response as often as a randomly selected individual in the human experiments happened to chose the most popular response.

5.2 Task II Results

Figure 2 illustrates the range of computational association scores attributed to only the strongest human responses. The position of the strongest human response to each cue word, within the computationally-ranked lists of all possible responses, is plotted on the y-axis. For each association measure the points are ordered from best to worst along the x-axis. In the ideal case therefore, the most popular human response for every cue word would appear at rank 1 amongst the computer-generated responses, resulting in a horizonal line at y=1. Generally speaking therefore, the smaller the area above a line the better the performance of a measure.

Three summary statistics can be derived from Figure 2:

1) The number of most popular human responses that are correctly predicted by a measure is indicated by the x-position at which its line departs from y=1. This can be seen to be around 11% for CWCD$_{sig}$ and is zero for the two best PMI parameterizations, with other illustrated measures performing intermediately.

2) The width of the flat horizontal tails at the opposite corner of the figure indicate the proportion of the cue words for which a measure was unable to differentiate the strongest human response from the large contingent of zero association scores resulting from unobservable co-occurrences. This tail is non-existent for CWCD$_{sig}$, but afflicts some 25% and 62% of cue words under the two best PMI parameterizations, again with other illustrated measures performing intermediately.

3) The median rank of the most popular human response for each measure can be read of on the y-axis at the horizontal mid-point (indicated by a feint vertical line).

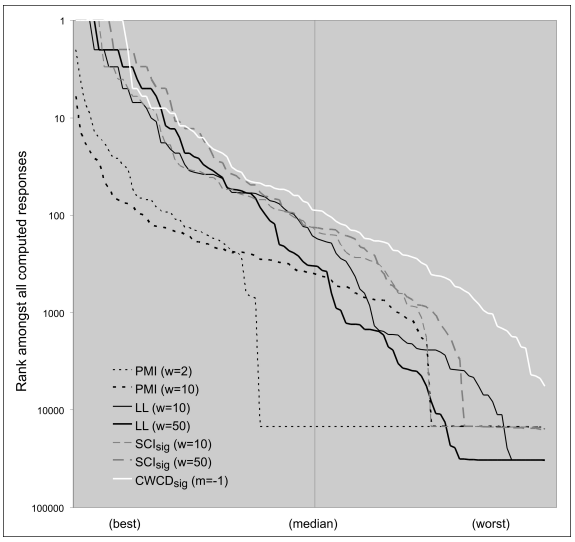


Figure 2: Agreement of computational measures with strongest human responses

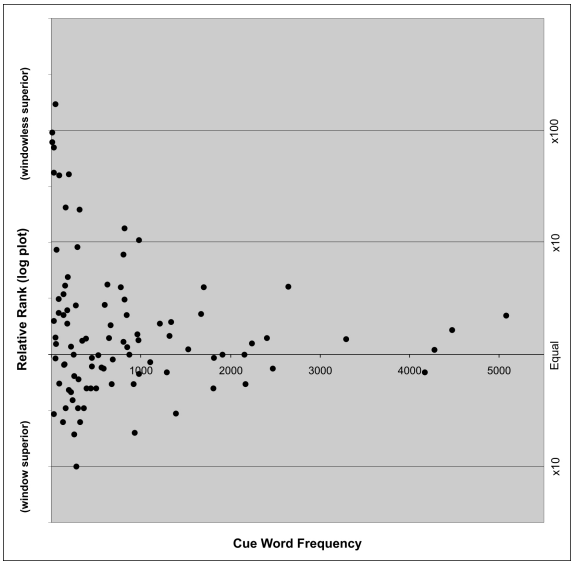


Figure 3: Relative agreement of computational measures with strongest human responses

The results shown are for the Kent dataset, and are highly typical. Included in the figure are the three frequency-based configurations with the highest median rank: SCI$_{sig}$ at window sizes $w = 10$ and $w = 50$, and standard LL at $w = 10$. Three

other frequency-based configurations are included for contrast. Also included is the single windowless configuration with the highest median rank - in this case CWCD$_{sig}$ using the harmonic mean. Several other windowless configurations (notably CWCD and the nearby means) and had very similar profiles.

Figure 3 shows the magnitude of the difference in the ranking of each of the same 100 strong human cue/response pairs, between the best windowless versus best windowed method. Points above the axis represent those cue/response pairs which the windowless method ranked more highly, and vice-versa. The points have been ordered on the x-axis according the the cue word frequency.

5.3 Discussion

Noteworthy, studying Figure 2, is the great sensitivity of the frequency-based measures to the window size parameter. There exists a cut-off point, linked to window size, beyond which the frequency-based measures are unable to make any differentiation between the desired human response and a large portion of the $33,000$ candidate responses. This is almost certainly due to a lack of evidence in the presence of very low frequency words. Log-Likelihood performs somewhat better in this respect, as it takes negative information into account.

Although the distance-based approach follows the same general trend as the other measures, it is nonetheless able to generate a distinct non-zero association score for *every* strong human response and overall it aptly ranks them more highly. A larger number these responses are actually ranked first (i.e. successfully predicted) by the distance-based approach. In fact this number is comparable to, and sometimes exceeds, the upper-bound of 10% implied by taking the average proportion of human respondents who give the most popular response to a given cue.

Whilst Figure 2 showed that overall the windowless method fairs better, on a per-cue basis (Figure 3) things are a little more interesting: For a little over a third of cue-words the windowed method actually appears to perform somewhat better. For the majority however, the windowless approach performs *considerably better* (note that the y-axis scale is logarithmic). It can also be seen that the difference between the methods is most pronounced for low frequency cue words, with responses to some cues exhibiting a relative ranking of around one-hundred times lower for the windowed method. This further supports the theory that the windowless methods are better able to exploit sparse data.

6 Conclusions and Future work

This paper presented the first empirical comparison of window-based and the relatively recently introduced windowless association measures, using their ability to reproduce human association scores as a testbed. We show that the best windowless measures are always at least as good as the best window-based measures, both when it comes to overall correlation with human association scores and predicting the strongest human response. In addition, for several human association sets, they perform significantly better. Although not all parameter settings and corpus sizes could be explored, we conclude that it is worthwhile investigating windowless association measures further. As a side-benefit, we have also introduced new variants of existing frequency-based association measures and shown them to perform as well as or better than their existing counterparts. Although these measures were semi-principled in their construction, a deeper understanding of why they work so well is needed. This may in turn lead to the construction of superior windowless measures.

In our own future work, we are especially interested in using higher-order windowless association measures for retrieving paradigmatic relations as well as exploring their use in various NLP applications.

7 Acknowledgements

We would like to extend sincere thanks to Reinhard Rapp for providing us with the Minnesota dataset in digital form, and additional thanks to Eric Atwell for his support.

References

M. Baroni, S. Evert, and A. Lenci, editors. 2008. *Esslli Workshop on Distributional Lexical Semantics*.

L. Burnard, 1995. *Users' Reference Guide, British National Corpus*. British National Corpus Consortium, Oxford, England.

K. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proc. of ACL-89*, pages 76–83.

K. Church and P. Hanks. 1991. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

P. Clark and F.C. Evans. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35:445–453.

J. Curran. 2003. *From distributional to semantic similarity*. Ph.D. thesis, University of Edinburgh.

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

S. Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut fr maschinelle Sprachverarbeitung, University of Stuttgart.

D. Hardcastle. 2005. Using the distributional hypothesis to derive coocurrence scores from the British National Corpus. In *Proc. of Corpus Linguistics*.

J. Jenkins. 1970. The 1952 Minnesota word association norms. In L. Postman and G. Keppel, editors, *Norms of word associations*, pages 1–38. Academic press.

G. Kent and A. Rosanoff. 1910. A study of association in insanity. *Amer. J. of Insanity*, pages 317–390.

G. Kiss, C. Armstrong, R. Milroy, and J. Piper. 1973. An associative thesaurus of English and its computer analysis. In A. Aitken, R. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*. Edinburgh University Press.

B. Krenn and S. Evert. 2001. Cam we do better than frequency? a case study on extracting pp-verb collocations. In *Proc. of the ACL Workshop on Collocations*.

A. Lamjiri, O. El Demerdash, and L. Kosseim. 2004. Simple features for statistical word sense disambiguation. In *Proc. of SENSEVAL-2004*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL-98*.

Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. 2007. Asymmetric association measures. In *Proc. of RANLP-2007*.

D. Nelson, C. McEvoy, J. Walling, and J. Wheeler. 1980. The University of South Florida homograph norms. *Behaviour Research Methods and Instrumentation*, 12:16–37.

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

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proc. of the 21st National Conference on Artificial Intelligence; 2004*.

R. Rapp. 2002. The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Proc of COLING 2002*.

D.L. Sackett. 2001. Why randomized controlled trials fail but needn't: 2. failure to employ physiological statistics, or the only formula a clinician-trialist is every likely to need (or understand). *Canadian Medical Association Journal*, 165(9):1226–1237.

P. Savicki and J. Hlavacova. 2002. Measures of word commonness. *Journal of Quantitative Linguistcs*, 9(3):215–231.

E. Terra and C. Clarke. 2004. Fast computation of lexical affinity models. In *Proc of COLING 2004*.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4), December.

X. Wang, 2005. *Robust Utilization of Context in Word Sense Disambiguation*, chapter Modeling and Using Context, pages 529–541. Springer Lecture Notes in Computer Science.

J. Washtell. 2009. Co-dispersion: A windowless approach to lexical association. In *Proc. of EACL-2009*.

M. Wettler and R. Rapp. 1993. Computation of word associations based on the co-ocurrences of words in large corpora. In *Proc. of the First Workshop on Very Large Corpora*.

K. White and L. Abrams. 2004. Free associations and dominance ratings of homophones for young and older adults. *Behaviour Research Methods, Instruments and Computers*, 36:408–420.

D. Yarowsky and R Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310.