Lexical Co-occurrence, Statistical Significance, and Word Association

Dipak L. Chaudhari
Computer Science and Engg.
IIT Bombay
dipakc@cse.iitb.ac.in

Om P. Damani
Computer Science and Engg.
IIT Bombay
damani@cse.iitb.ac.in

Srivatsan Laxman
Microsoft Research India
Bangalore
slaxman@microsoft.com

Abstract

Lexical co-occurrence is an important cue for detecting word associations. We propose a new measure of word association based on a new notion of statistical significance for lexical co-occurrences. Existing measures typically rely on global unigram frequencies to determine expected co-occurrence counts. Instead, we focus only on documents that contain both terms (of a candidate word-pair) and ask if the distribution of the observed spans of the word-pair resembles that under a random null model. This would imply that the words in the pair are not related strongly enough for one word to influence placement of the other. However, if the words are found to occur closer together than explainable by the null model, then we hypothesize a more direct association between the words. Through extensive empirical evaluation on most of the publicly available benchmark data sets, we show the advantages of our measure over existing co-occurrence measures.

1 Introduction

Lexical co-occurrence is an important indicator of word association and this has motivated several co-occurrence[1] measures for word association like PMI (Church and Hanks, 1989), LLR (Dunning, 1993), Dice (Dice, 1945), and CWCD (Washtell and Markert, 2009). In this paper, we present a new measure of word association based on a new notion of statistical significance for lexical co-occurrences. In general, a lexical co-occurrence could refer to a pair of words that co-occur in a large number of documents; or it could refer to a pair of words that, although co-occur only in a small number of documents, occur close to each other within those documents. We formalize these ideas and construct a significance test that allows us to detect different kinds of co-occurrences within a single unified framework (a feature which is absent in current measures for co-occurrence). Another distinguishing feature of our measure is that it is based solely on the co-occurrence counts in the documents containing both words of the pair, unlike all existing measures which also take global unigram frequencies in account.

We need a null hypothesis that can account for an observed co-occurrence as a pure chance event and this in-turn requires a corpus generation model. Documents in a corpus can be assumed to be generated independent of each other. Existing co-occurrence measures further assume that each document is drawn from a multinomial distribution based on global unigram frequencies. The main concern with such a null model is the overbearing influence of the unigram frequencies on the detection of word associations. For example, the association between *anomochilidae* (dwarf pipe snakes) and *snake* could go undetected in our wikipedia corpus, since less than 0.1% of the pages containing *snake* also contained *anomochilidae*. Also, under current models, the expected *span*[2] of a word pair is very sensitive to the associated unigram frequencies: the expected span of a word pair composed of low frequency unigrams is much larger than that with high frequency unigrams. This is contrary to how word associa-

[1] We use the term co-occurrence to refer to a pair of words that co-occur in a document with an arbitrary number of intervening words.

[2] The *span* of an occurrence of a word-pair is the 'unsigned distance' between the positions of the corresponding word occurrences.

tions appear in language, where semantic relationships manifest with small inter-word distances irrespective of the underlying unigram distributions.

Based on these considerations we employ a null model that represents each document as a bag of words [3]. A random permutation of the associated bag of words gives a linear representation for the document. Under this null model, the locations of an unrelated pair of words will likely be randomly distributed in the documents in which they co-occur. If the observed span distribution of a word-pair resembles that under the (random permutation) null model, then the relation between the words is not strong enough for one word to influence the placement of the other. However, if the words are found to occur closer together than explainable by our null model, then we hypothesize a more direct association between the words. Therefore, this null model detects biases in span distributions of word-pairs while being agnostic to variations in global unigram frequencies.

In this paper, we propose a new measure of word association based on the statistical significance of the observed span distribution of a word-pair. We perform extensive experiments on all the publicly available benchmark data sets[4] and compare our measure against other popular co-occurrence measures. Our experiments demonstrate the advantages of our measure over all the competing measures. The ranked list of word associations output by our measure has the best correlation with the corresponding gold-standard in three (out of seven) data sets in our experiments, while remaining in the top three in other four datasets. While different measures perform best on different data sets, our measure outperforms other measures by being consistently either the best measure or very close to the best measure on all the data sets. The average deviation of our measure's correlation with the gold-standard from the best measure's correlation with the gold-standard (average taken across all the datasets) is 0.02, which is the least average deviation among all the measures, the next best deviations being 0.04 and 0.06.

The paper is organized as follows. We present our notion of statistical significance of span distribution in Section 2. Algorithm for computing the proposed word association measure is described in Section 3. We discuss related work in Section 4. Performance evaluation is presented in Section 5 followed with conclusions in Section 6.

2 Lexically significant co-occurrences

Evidence for significant lexical co-occurrences can be gathered at two levels in the data – document-level and corpus-level. First, at the document level, we may find that for a given word-pair, a surprisingly high proportion of its occurrences *within* a document have smaller spans than they would have by random chance. Second, at the corpus-level, we may find a pair of words appearing closer-than-random in multiple documents in the corpus. We now describe how to combine both kinds of evidence to decide whether the nearby occurrences of a word-pair are statistically significant or not.

Let the *frequency* f of a word-pair α in a document D, be the maximum number of *non-overlapped occurrences* of α in D. A set of occurrences of a word-pair is said to be non-overlapped if the words corresponding to one occurrence from the set do not appear in-between the words corresponding to any other occurrence from the set.

Let $\widehat{f}^x$ denote the maximum number of non-overlapped occurrences of α in D with span less than a given threshold x. We refer to $\widehat{f}^x$ as the *span-constrained frequency* of α in D. Note that $\widehat{f}^x$ cannot exceed f.

2.1 Document-level significant co-occurrence

To assess the statistical significance of the word-pair α we ask if the span-constrained frequency $\widehat{f}^x$ (of α) is more than what we would expect in a document of size ℓ containing f 'random' occurrences of α. Our intuition is that if two words are associated in some way, they will often appear close to each other in the document and so the distribution of the spans will typically exhibit a bias toward values less than a suitably chosen threshold x.

[3] There can be many ways to associate a bag of words with a document. Details of this association are not important for us, except that the bag of words provides some kind of quantitative summary of the words within the document.

[4] We exclude very small data sets of 80 word pairs or less. Sizes of the seven datasets we used range from 351 word-pairs to 83,713 word-pairs.

Definition 1 *Consider the null hypothesis that the linear representation of a document is generated by choosing a random permutation of the bag of words associated with the document. Let ℓ be the length of the document and f denote the frequency of a word-pair in the document. For a given a span threshold x, we define $\pi_x(\widehat{f^x}, f, \ell)$ as the probability under the null that the word-pair will appear in the document with a span-constrained frequency of* at least $\widehat{f^x}$.

Observe that $\pi_x(0, f, \ell) = 1$ for any $x > 0$; also, for $x \geq \ell$ we have $\pi_x(f, f, \ell) = 1$ (i.e. all f occurrences will always have span less than x for $x \geq \ell$). However, for typical values of x (i.e. for $x \ll \ell$) the probability $\pi_x(\widehat{f^x}, f, \ell)$ decreases with increasing $\widehat{f^x}$. For example, consider a document of length 400 with 4 non-overlapped occurrences of α. The probabilities of observing at least 4, 3, 2, 1 and 0 occurrences of α within a span of 20 words are 0.007, 0.09, 0.41, 0.83, and 1.0 respectively. Since $\pi_{20}(3, 4, 400) = 0.09$, even if 3 of the 4 occurrences of α have span less than 20 words, there is 9% chance that the occurrences were a consequence of a random event. As a result, if we desired a confidence-level of at least 95%, we would have to declare observed co-occurrences of α as *insignificant*.

Given an ϵ ($0 < \epsilon < 1$) and a span threshold x (≥ 0) the document D is said to *support* the hypothesis "α is an ϵ-significant word-pair within the document" if we have $[\pi_x(\widehat{f^x}, f, \ell) < \epsilon]$. We refer to ϵ as the *document-level* evidence of the lexical co-occurrence of α.

2.2 Corpus-level significant co-occurrence

We now describe how to aggregate evidence for lexical significance by considering the occurrence of α across multiple documents in the corpus. Let $\{D_1, \ldots, D_K\}$ denote the set of K documents (from out of the entire corpus) that contain at least one occurrence of α. Let ℓ_i be the length of D_i, f_i be the frequency of α in D_i, and, $\widehat{f_i^x}$ be the *span-constrained frequency* of α in D_i. Define indicator variables z_i, $i = 1, \ldots, K$ as:

$$z_i = \begin{cases} 1 & \text{if } \pi_x(\widehat{f_i^x}, f_i, \ell_i) < \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

As discussed previously, z_i indicates whether "α is an ϵ-significant word-pair within the document D_i." Note that we view $\widehat{f_i^x}$ as the only random quantity here, with x fixed by the user, and ℓ_i and f_i fixed given the document D_i and word-pair α. Let $Z = \sum_{i=1}^{K} z_i$; Z models the number of documents (out of K) that support the hypothesis "α is an ϵ-significant word-pair." The expected value of Z is given by

$$\begin{aligned} E(Z) &= \sum_{i=1}^{K} E(z_i) \\ &= \sum_{i=1}^{K} \pi_x(g_{\epsilon,x}(f_i, \ell_i), f_i, \ell_i) \quad (2) \end{aligned}$$

where $g_{\epsilon,x}(f_i, \ell_i)$ is given by *Definition 2* below.

Definition 2 *Given a document of length ℓ in which a word-pair has a frequency of f, and given a span threshold x, we define $g_{\epsilon,x}(f, \ell)$ as the smallest r for which the inequality $[\pi_x(r, f, \ell) < \epsilon]$ holds.*

Note that $g_{\epsilon,x}(f, \ell)$ is well-defined since $\pi_x(r, f, \ell)$ is non-increasing with respect to r. For the example given earlier, $g_{0.2,20}(4, 400) = 3$ and $g_{0.05,20}(4, 400) = 4$. Since each document in the corpus is assumed to be generated independently, z_i's are independent random variables and we can bound the deviation of the observed value of Z from its expectation using Hoeffding's Inequality – for any $t > 0$, we have

$$\begin{aligned} P[Z \geq E(Z) + Kt] &\leq \exp(-2Kt^2) \\ &= \delta \quad (3) \end{aligned}$$

Recall that Z models the number of documents supporting the hypothesis "α is an ϵ-significant word-pair."). Thus, the upper-bound δ ($= \exp(-2Kt^2)$), $0 < \delta < 1$ denotes the upper-bound on the probability that just due to random chance, more than $(E(Z) + Kt)$ documents out of K will support the hypothesis "α is an ϵ-significant word-pair." We call δ the *corpus-level* evidence of the lexical co-occurrence α. For example, in our corpus, the word-pair (*canyon, landscape*) occurs in $K = 416$ documents. For $\epsilon = 0.1$, we have ϵ-significant occurrences in $Z = 33$ documents (out of 416) , while $E(Z) = 14.34$. Suppose we want to be 99% sure that the occurrences of (*canyon, landscape*) in the 33 documents were a consequence of non-random phenomena. Let $\delta = 1 - 0.99 = 0.01$. By setting

word-1	*word-2*		
	$(0.1, 0.1)$	$(0.1, 0.4)$	$(0.4, 0.1)$
algae	green	mold	pool
amuse	entertain	clown	amaze
damn	hell	mad	bad
rat	dirty	ugly	disease
sedative	drug	narcotic	calm
topping	chocolate	flavour	caramel
umbrella	rain	dry	shade
unknown	known	dark	secret
worm	insect	dirt	fishing
wrap	cover	seal	bandage

Table 1: Examples of word-pairs from *Florida* dataset having statistically significant co-occurrences in the wikipedia corpus for different (ϵ, δ) combinations under a span constraint of 20 words.

$t = \sqrt{\ln \delta/(-2K)} = 0.07$, we get $E(Z) + Kt = 43.46$. Only if Z was 44 or more, there would be less than 1% chance of that being a random phenomena. Thus, we cannot be 99% sure that the observed co-occurrences in the 33 documents are non-random. Hence, our test declares (*canyon, landscape*) as *insignificant* at $\epsilon = 0.1, \delta = 0.01$. We now summarize our significance test in the definition below.

Definition 3 (Significant lexical co-occurrence) *Consider a word-pair α and a set of K documents containing at least one occurrence each of α. Fix a span threshold of x (> 0), a document-level evidence of ϵ ($0 < \epsilon < 1$) and a corpus-level evidence of δ ($0 < \delta < 1$). Let Z denote the number of documents (out of K) that support the hypothesis "α is ϵ-significant within the document." The word-pair α is said to be (ϵ, δ)-significant if we have $[Z \geq E(Z) + Kt]$, where $t = \sqrt{\log \delta/(-2K)}$ and $E(Z)$ is given by Eq. (2). The ratio $[Z/(E(Z) + Kt)]$ is called the* Co-occurrence Significance Ratio (CSR) *for α.*

2.3 Discussion

The significance test of *Definition 3* gathers both document-level and corpus-level evidence from data in calibrated amounts. Prescribing ϵ fixes the strength of the document-level hypothesis in our test, while, δ, controls the extent of corpus-level evidence we need to declare a word-pair as significant. A small δ demands that there must be multiple documents in the corpus, each of which, individually have some evidence of relatedness for the pair of words.

By running the significance test with different values of ϵ and δ, the CSR test can be used to detect different types of lexically significant co-occurrences. For example, the strongest lexical co-occurrences would have both strong document-level evidence (low ϵ) as well as high corpus-level evidence (low δ). Informally, these would represent pairs of words that appear multiple times with small spans within a document, in many documents, and in-practice, we find that multi-word expressions or pairs of words separated by stop words tend to dominate this type. On the other hand, a higher ϵ would represent word-pairs that appear relatively farther apart within a document, or a higher δ would represent word-pairs that appear together in relatively fewer documents. Note that to detect co-occurrences that exclusively correspond to (say) low ϵ and high δ, we would have to run the test with low ϵ and high δ, and then remove word-pairs that were also found significant at low ϵ and low δ.

In Table 1, we present some examples of different types of co-occurrences. The table lists word-pairs that were found to be statistically significant for different choices of (ϵ, δ). Note that a word-pair is reported under $(\epsilon = 0.1, \delta = 0.4)$ or $(\epsilon = 0.4, \delta = 0.1)$ only if it was not also found significant under other two parameter settings. The strongest correlations are the word-pairs corresponding to $(\epsilon = 0.1, \delta = 0.1)$ e.g., *algae-green*, *rat-dirty* and *worm-insect*. Different sets of weaker co-occurrences are detected depending on whether we relaxed δ or ϵ. For example, *algae-mold* is significant at a higher δ, while *algae-pool* is significant for higher ϵ.

The semantic notion of word association is an abstract concept and different kinds of associations (with potentially different statistical characterizations) may be preferred by human judges in different situations. While in Section 5, we discuss in detail various datasets used, the evaluation methodology, and the performance of CSR across datasets, we wish to point out here that in 3 out of 5 cross-validation runs for *wordsim* dataset, the best performing CSR parameters were $x = 50w$, $\epsilon = 0.1$ and $\delta = 0.9$, while in 3 out of 5 runs for *Minnesota* dataset, the best performing CSR parameters were $x = 20w$, $\epsilon = 0.3$ and $\delta = 0.5$. This gives us some indication that different kinds of word associations were preferred in different data sets.

3 Computing Co-occurrence Significance Ratio(CSR)

There are three main steps for computing CSR and the pseudocodes for these are listed in Procedures 1, 2 & 3. Of these, the first two can be run offline since they do not depend on the text corpus. They need to be run only once, after which CSR can be computed for any word-pair on any given corpus of documents. We describe these steps in the subsections below.

3.1 Computing histogram $hist_{f,\ell,x}(\cdot)$

The first step is to compute a histogram for the span-constrained frequency, $\widehat{f}^x$, of a word-pair whose frequency is f in a document of length ℓ, given a chosen span threshold of x (under our null model).

Definition 4 *Given a document of length ℓ and a span threshold of x, we define $hist_{f,\ell,x}(\widehat{f}^x)$ as the number of ways to embed f non-overlapped occurrences of a word-pair in the document such that exactly $\widehat{f}^x$ occurrences have span less than x.*

Procedure 1 $ComputeHist(f, \ell, x)$ – Offline

Input f - number of non-overlapped occurrences; ℓ - document length; x - span threshold
Computes $hist_{f,\ell,x}[\cdot]$ as per *Definition 4*

```
1:  Initialize hist_{f,ℓ,x}[f̂^x] ← 0 for f̂^x = 0, ..., f
2:  if f > ℓ then
3:      return
4:  if f = 0 then
5:      hist_{f,ℓ,x}[0] ← 1
6:      return
7:  for i ← 1 to (ℓ − 1) do
8:      for j ← (i + 1) to ℓ do
9:          hist_{f−1,ℓ−j,x} ← ComputeHist(f − 1, ℓ − j, x)
10:         for k ← 0 to f − 1 do
11:             if (j − i) < x then
12:                 hist_{f,ℓ,x}[k + 1] ← hist_{f,ℓ,x}[k + 1]
                                          + hist_{f−1,ℓ−j,x}[k]
13:             else
14:                 hist_{f,ℓ,x}[k] ← hist_{f,ℓ,x}[k] + hist_{f−1,ℓ−j,x}[k]
```

Procedure 1 lists the pseudocode for computing the histogram $hist_{f,\ell,x}$. The main steps involve selecting a start and end position for embedding the very first occurrence (lines 7-8) and then recursively calling $ComputeHist(\cdot,\cdot,\cdot)$ (line 9). The i-loop selects a start position for the first occurrence of the word-pair, and the j-loop selects the end position. The recursion step now computes the number of ways to embed the remaining $(f-1)$ non-overlapped occurrences in the remaining $(\ell - j)$ positions. Once we have $hist_{f-1,\ell-j}$, we check whether the occurrence introduced at positions (i, j) will contribute to the $\widehat{f}^x$ count. If $(j - i) < x$, whenever there are k span-constrained occurrences in positions $(j + 1)$ to ℓ, there will be $(k + 1)$ span-constrained occurrences in positions 1 to ℓ. Thus, we increment $hist_{f,\ell}[k + 1]$ by the quantity $hist_{f-1,\ell-j}[k]$ (lines 10-12). However, if $(j - i) > x$, there is no contribution to the span-constrained frequency from the (i, j) occurrence, and so we increment $hist_{f,\ell}[k]$ by the quantity $hist_{f-1,\ell-j}[k]$ (lines 10-11, 13-14). Finally, we note that in our implementation we use memorization to avoid redundant recursive calls.

3.2 Computing $\pi_x(\cdot, f, \ell)$ distribution

Procedure 2 $ComputePiDist(f, \ell, x)$ – Offline

Input f - number of non-overlapped occurrences; ℓ - document length; x - span threshold
Computes Distribution $\pi_x[f, \ell, \cdot]$ as per *Definition 1* and $g_{\epsilon,x}[f, \ell]$ as per *Definition 2*

```
1: N[f, ℓ, x] = Σ_{k=0}^{f} hist_{f,ℓ,x}[k]
2: for f̂^x ← 0 to f do
3:     N_x[f̂^x, f, ℓ] ← Σ_{k=f̂^x}^{f} hist_{f,ℓ,x}[k]
4:     π_x[f̂^x, f, ℓ] ← N_x[f̂^x, f, ℓ] / N[f, ℓ, x]
5: g_{ϵ,x}[f, ℓ] ← min{r | π_x[r, f, ℓ] < ϵ}
```

The second offline step is computation of the $\pi_x(\cdot, f, \ell)$ distribution. We store the number of ways of embedding f non-overlapped occurrences of a word-pair in a document of length ℓ in the array $N[f, \ell]$. Similarly, the array $N_x[\widehat{f}^x, f, \ell]$ stores the number of ways of embedding f non-overlapped occurrences of the word-pair in a document of length ℓ, such that at least $\widehat{f}^x$ of the f occurrences have span less than x. To compute $N[f, \ell, x]$ and $N_x[\widehat{f}^x, f, \ell]$, we need the histogram $hist_{f,\ell,x}[\cdot]$ which is the output of *Procedure 1*. *Procedure 2* lists the pseudocode for computing $\pi_x(\widehat{f}^x, f, \ell)$ from $N(f, \ell)$ and $N_x(\widehat{f}^x, f, \ell)$ given $hist_{f,\ell}$ from *Procedure 1* (For the sake of readability the pseudocode does not describe some optimizations that we used in our implementation).

The *Procedure 1* is exponential in f and ℓ but it does not depend on the data corpus. Hence, we can run the Procedures 1 and 2 off-line, and publish the $\pi_x[]$ and $g_{\epsilon,x}[]$ tables for various x, $\widehat{f}^x$, f and

ℓ. Using these tables[5], anyone wishing to compute CSR needs to only run Procedure 3.

3.3 Computing CSR for a given word-pair

Procedure 3 $ComputeCSR(\alpha, \epsilon, \delta, x)$

Input α - word-pair; ϵ - document-level evidence; δ - corpus-level evidence; x - span threshold; Corpus of documents
Computes $CSR(\alpha)$ - Co-occurrence Significance Ratio (CSR) for α as per *Definition 3*

1: $\mathcal{D} \leftarrow \{D_1, \ldots, D_K\}$ // Set of documents from the corpus that each contain at least one occurrence of α.
2: $t \leftarrow \sqrt{\log \delta/(-2K)}$
3: $Z \leftarrow 0$ and $Z_E \leftarrow 0$
4: **for** $i \leftarrow 1$ to K **do**
5: $\quad \ell_i$ = Length of D_i
6: $\quad f_i$ = Frequency of α in D_i
7: $\quad \widehat{f}_i^x$ = Span-constrained frequency of α in D_i
8: $\quad$ **if** $\pi_x[\widehat{f}_i^x, f_i, \ell_i] < \epsilon$ **then**
9: $\quad\quad z_i \leftarrow 1$
10: $\quad$ **else**
11: $\quad\quad z_i \leftarrow 0$
12: $\quad Z \leftarrow Z + z_i$
13: $\quad Z_E \leftarrow Z_E + \pi_x[g_\epsilon[f_i, \ell_i, x], f_i, \ell_i]$
14: $CSR(\alpha) = Z/(Z_E + Kt)$

Procedure 3 implements the significance test given in *Definition 3* and requires that the $\pi_x[]$ and $g_{\epsilon,x}[]$ tables have already been computed offline.

The first step is to determine the subset $\mathcal{D}$ of documents containing the given word-pair (line 1). Then we compute t based on δ and K (the size of $\mathcal{D}$) (line 2). Next we determine how many of the K documents support the hypothesis "α is ϵ-significant within the document" (lines 3-12). The expected number of documents supporting the hypothesis is accumulated in Z_E (line 13). CSR is then computed as the ratio of Z to $(Z_E + Kt)$ (line 14).

3.4 Run-time overhead

The computation of Co-occurrence Significance Ratio (CSR) as given in Definition 3 might appear more complex than the simple formulae for other co-occurrence measures given in Table 2. However, bulk of the complexity in calculating CSR lies in the one-time (data independent) off-line computation of the $\pi_x[]$ and $g_{\epsilon,x}[]$ tables. Once these tables are published, the cost of comparing CSR for a given word pair is comparable to the cost of computing any other (spanned) measure in Table 2. The main data-dependent computations for a spanned measure are in determining span-constrained frequencies; all other steps are simple arithmetic operations or memory lookups. To illustrate this, Procedure 4 gives details of computing PMI. The comparison of Procedures 3 and 4 shows their almost parallel structures. The main overhead in these procedures is incurred in line 7, where span-constrained frequencies in a given document are computed.

Procedure 4 $ComputePMI(a, b)$

Input (x, y) - word pair;
Computes PMI (Table 2) for (x, y).

1: let $\mathcal{D} = \{D_1, \ldots, D_K\}$ // set of documents containing at least one occurrence of α.
2: N = total number of words in corpus
3: (f_x,f_y) = unigram frequencies of x, y in corpus
4: (p_x,p_y) = (f_x/N,f_y/N)
5: $\widehat{f} = 0$
6: **for** $i \leftarrow 1$ to K **do**
7: $\quad \widehat{f}_i$ = span-constrained frequency of α in D_i
8: $\quad \widehat{f} = \widehat{f} + \widehat{f}_i$
9: $\hat{p}_{x,y} = \widehat{f}/N$
10: $PMI = log(\frac{\hat{p}_{x,y}}{p_x p_y})$

4 Related Work

Existing word association measures can be divided into three broad categories: (i) *Co-occurrence measures* that rely on co-occurrence frequencies of both words in a corpus in addition to the individual unigram frequencies (Table 2), (ii) *Distributional similarity-based measures* that characterize a word by the distribution of other words around it (Agirre et al., 2009; Bollegala et al., 2007; Chen et al., 2006; Wandmacher et al., 2008), and (iii) *Knowledge-based measures* that use knowledge-sources like thesauri, semantic networks, or taxonomies (Milne and Witten, 2008; Hughes and Ramage, 2007; Gabrilovich and Markovitch, 2007; Yeh et al., 2009; Strube and Ponzetto, 2006; Finkelstein et al., 2002; Liberman and Markovitch, 2009).

In this paper, we focus on comparison with other co-occurrence measures. These measures are used in several domains like ecology, psychology, medicine, and language processing. Table 2 lists several measures chosen from all these domains. Except Ochiai (Ochiai, 1957; Janson and Vegelius, 1981) and the recently introduced

[5] http://www.cse.iitb.ac.in/~damani/papers/EMNLP11/resources.html

Method	Formula
CSR (this work)	$Z/(E(Z)+Kt)$
CWCD (Washtell and Markert, 2009)	$\frac{\hat{f}(x,y)}{p(x)} \frac{1/max(p(x),p(y))}{M}$
Dice (Dice, 1945)	$\frac{2\hat{f}(x,y)}{f(x)+f(y)}$
LLR (Dunning, 1993)	$\sum_{\substack{x' \in \{x,\neg x\} \\ y' \in \{y,\neg y\}}} p(x',y') log \frac{p(x',y')}{p(x')p(y')}$
Jaccard (Jaccard, 1912)	$\frac{\hat{f}(x,y)}{f(x)+f(y)-\hat{f}(x,y)}$
Ochiai (Janson and Vegelius, 1981)	$\frac{\hat{f}(x,y)}{\sqrt{f(x)f(y)}}$
Pearson's χ^2 test	$\sum_{\substack{x' \in \{x,\neg x\} \\ y' \in \{y,\neg y\}}} \frac{\left(\hat{f}(x',y')-E\hat{f}(x',y')\right)^2}{E\hat{f}(x',y')}$
PMI (Church and Hanks, 1989)	$log \frac{p(x,y)}{p(x)p(y)}$
SCI (Washtell and Markert, 2009)	$\frac{p(x,y)}{p(x)\sqrt{p(y)}}$
T-test	$\frac{\hat{f}(x,y)-E\hat{f}(x,y)}{\sqrt{\hat{f}(x,y)\left(1-\frac{\hat{f}(x,y)}{N}\right)}}$

N	Total number of tokens in the corpus
$f(x), f(y)$	unigram frequencies of x, y in the corpus
$p(x), p(y)$	$f(x)/N, f(y)/N$
$\hat{f}(x,y)$	Span-constrained (x,y) word pair frequency in corpus
$\hat{p}(x,y)$	$\hat{f}(x,y)/N$
M	Harmonic mean of the spans of $\hat{f}(x,y)$ occurrences
$E\hat{f}(x,y)$	Expected value of $\hat{f}(x,y)$

Table 2: Co-occurrence measures.

CWCD[6] (Washtell and Markert, 2009) all other measures are well-known in the NLP community (Pecina and Schlesinger, 2006). Our results show that Ochiai and Chi-Square have almost identical performance, differing only in 3rd decimal digits. Rankings produced by Chi-square is almost monotonic with respect to the rankings produced by Ochiai. This is because, for most word pairs (x,y), $[f(x) \ll N]$, $[f(y) \ll N]$, $[f(x,y) \ll f(x)]$, and $[f(x,y) \ll f(y)]$. Therefore three of the four terms in the Chi-square summation become zero[7] and the fourth term approximates to the square of Ochiai. Similarly Jaccard and Dice coincide. While presenting our experimental results, we report these pairs of measures together.

[6]CWCD was reported in (Washtell and Markert, 2009) as the best performing variant among the so-called windowless (or spanless) measures. In our experiments, we implemented windowed (spanned) version of the CWCD measure.

[7]For example, $\hat{f}(x,\neg y) - Ef(x,\neg y) = f(x) - N \times pf(x) \times pf(\neg y) = f(x) - \frac{1}{N} \times f(x) \times f(\neg y) = f(x) - \frac{1}{N} \times f(x) \times N = 0$.

Aspect	Data Set	No. of Respondents	No. of Word Pairs	No. of Filtered Word Pairs
Semantic relatedness	wordsim (Finkelstein et al., 2002)	16	353	351
Free-Association	Edinburg (Kiss et al., 1973)	100	325,588	83,713
	Florida (Nelson et al., 1980)	5,019	65,523	59,852
	Goldfarb-Halpern (Goldfarb and Halpern, 1984)	316	410	384
	Kent (Kent and Rosanoff, 1910)	1,000	14,576	14,086
	Minnesota (Russell and Jenkins, 1954)	1,007	10,447	9,649
	White-Abrams (White and Abrams, 2004)	440	745	652

Table 3: Characteristics of data sets used.

5 Performance Evaluation

Two main aspects of word association studied in literature are: a) *semantic relatedness*, and b) *free association*. *Semantic relatedness* encompasses many different relationships between words, like synonymy, meronymy, antonymy, and functional association (Budanitsky and Hirst, 2006). *Free association* refers to the first response-words that come to mind when presented with a stimulus. (ESSLLI, 2008). We experiment with all the publicly available datasets that come with gold standard judgement of these aspects, except the very small ones with less than 80 word-pairs[8].

5.1 Datasets

Details[9] of the datasets used in our experiments are listed in Table 3. Each data set comes with a gold-standard of human judgments - a ranked list of association scores for the word-pairs in the data set. The *wordsim* dataset was prepared by asking the subjects to estimate the relatedness of the word pairs on a

[8](MillerCharles (Miller and Charles, 1991), Rubenstein-Goodenough (Rubenstein and Goodenough, 1965) and TOEFL (Landauer and Dumais, 1997))

[9]We removed word-pairs containing multiword expressions. For data sets with more than 10,000 word-pairs, we filtered out pairs that contain stop words listed in (StopWordList, 2010). For Edinburg (size 275393 after previous filtering), we further filtered word-pairs where the response was supported by only one respondant. Original and filtered data sets are available at http://www.cse.iitb.ac.in/~damani/papers/EMNLP11/resources.html

	Edinburg (83,713)	Florida (59,852)	Kent (14,086)	Minnesota (9,649)	White-Abrams (652)	Goldfarb-Halpern (384)	wordsim (351)
CSR	**0.25**	**0.30**	0.42	0.31	**0.34**	0.10	0.63
CWCD	0.23	0.23	0.40	0.30	0.21	**0.19**	0.54
Dice (Jaccard)	0.20	0.27	**0.43**	**0.32**	0.21	0.09	0.59
LLR	0.20	0.26	0.40	0.29	0.18	0.03	0.51
Ochiai (χ^2)	0.24	**0.30**	**0.43**	0.31	0.29	0.08	0.62
PMI	0.22	0.25	0.36	0.26	0.22	0.11	**0.69**
SCI	0.24	0.27	0.38	0.27	0.23	0.06	0.37
TTest	0.17	0.23	0.37	0.26	0.17	-0.02	0.45

Table 4: Comparison of the average Spearman coefficients obtained across five cross-validation runs by different measures. The best performing measure for each data-set is shown in bold. All standard deviations for Edinburg and Florida were less than 0.01, for Kent and Minnesota were between 0.01 and 0.02, for White-Abrams were between 0.05 and 0.08, for Goldfarb-Halpern between 0.05 and 0.15 and for wordsim were between 0.02 and 0.15. Number of word-pairs in each dataset is shown in brackets against its name.

	Edinburg (83,713)	Florida (59,852)	Kent (14,086)	Minnesota (9,649)	White-Abrams (652)	Goldfarb-Halpern (384)	wordsim (351)	Worst Rank	Avg. Deviation	Worst Deviation
CSR	**0.00 (1)**	**0.00 (1)**	0.01 (3)	0.01 (2)	**0.00 (1)**	0.09 (3)	0.06 (2)	**3**	**0.02**	**0.09**
CWCD	0.02 (4)	0.07 (7)	0.03 (4)	0.02 (4)	0.13 (5)	**0.00 (1)**	0.15 (5)	7	0.06	0.15
Dice (Jaccard)	0.05 (6)	0.03 (3)	**0.00 (1)**	**0.00 (1)**	0.13 (5)	0.10 (4)	0.10 (4)	6	0.06	0.13
LLR	0.05 (6)	0.04 (5)	0.03 (4)	0.03 (5)	0.16 (7)	0.16 (7)	0.18 (6)	7	0.09	0.18
Ochiai (χ^2)	0.01 (2)	**0.00 (1)**	**0.00 (1)**	0.01 (2)	0.05 (2)	0.11 (5)	0.07 (3)	5	0.04	0.11
PMI	0.03 (5)	0.05 (6)	0.07 (8)	0.06 (7)	0.12 (4)	0.08 (2)	**0.00 (1)**	8	0.06	0.12
SCI	0.01 (2)	0.03 (3)	0.05 (6)	0.05 (6)	0.11 (3)	0.13 (6)	0.32 (8)	8	0.10	0.32
TTest	0.08 (8)	0.07 (7)	0.06 (7)	0.06 (7)	0.17 (8)	0.21 (8)	0.24 (7)	8	0.13	0.24

Table 5: Comparison of deviations from the best performing measure on each data set. Number of word-pairs in each dataset is shown in brackets against its name. Figures in brackets against the deviation values denote the ranks of the measures in the corresponding data sets.

scale from 0 to 10 (Finkelstein et al., 2002). The methodology for collecting *free association* data is explained at (ESSLLI, 2008): The degree of free association between a stimulus (S) and response (R) is the percentage of respondents who respond R as the first response when presented with stimulus S.

These datasets are of varying size, and they were constructed at different point in time, in different geographies. This allows us to compare different measures comprehensively under varying range of circumstances. To the best of our knowledge, no previous work has reported such a detailed comparison of co-occurrence measures.

5.2 Resources Used

We use the Wikipedia (Wikipedia, April 2008) corpus with 2.7 million articles (total of 1.24 Gigawords). We did no pre-processing - no lemmatization or function-word removal. When counting document size (in words), punctuations were ignored. Documents larger than 1500 words were partitioned such that each part was at most 1500 words[10]. We indexed the corpus using Lucene search engine library and used Lucene APIs to obtain various statistics and documents containing given word-pairs.

5.3 Methodology

Each measure listed in Table 2 produces a ranked list of association scores for the word-pairs in a data set. We evaluate each measure by the Spearman's rank correlation between the ranking produced by the measure and the gold-standard ranking.

The span threshold (or window-width) x is a user-defined parameter in all measures. In addition, CSR has the parameters ϵ and δ. For any measure, the ranking of word-pairs will likely change with chang-

[10]While this limit can be raised using heavier computing resources, we believe that partitioning documents of sizes greater than 1500 words was reasonable (especially since typical span values we used were less than 50, much less than 1500).

Method	Resource	wordsim			Esslli (272)
		wordsim (353)	sim (203)	rel (252)	
PMI	Wikipedia	0.69	0.72	0.68	0.32
Ochiai (χ^2)	Wikipedia	0.62	0.68	0.62	**0.44**
Significance Ratio (CSR)	Wikipedia	0.63	0.70	0.64	0.43
Latent Semantic Analysis (Wandmacher et al., 2008)	Newspaper corpus	-	-	-	0.38
Graph Traversal (WN30g) (Agirre et al., 2009))	Wordnet	0.66	0.72	0.56	-
Bag of Words based Distributional Similarity (BoW) (Agirre et al., 2009))	Web corpus	0.65	0.70	0.62	-
Context Window based Distributional Similarity (CW) (Agirre et al., 2009))	Web corpus	0.60	**0.77**	0.46	-
Hyperlink Graph (Milne and Witten, 2008)	Wikipedia hyperlinks graph	0.69	-	-	-
Random Graph Walk (Hughes and Ramage, 2007)	WordNet	0.55	-	-	-
Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007) (reimplemented in (Yeh et al., 2009))	Wikipedia concepts	**0.75** **(0.71)**	-	-	-
Normalized Path-length (lch) (Strube and Ponzetto, 2006)	Wikipedia category tree	0.55	-	-	-
Thesarus based (Jarmasz, 2003)	Roget's thesaurus	0.55	-	-	-
Latent Semantic Analysis (Finkelstein et al., 2002)	Web corpus	0.56	-	-	-

Table 6: Comparison of co-occurrence based measures with knowledge-based and distributional similarity based measures. These other measures have not been applied to the free association datasets shown in Table 3. Data for missing entries is not available. Note that *sim* and *rel* are subsets of *wordsim* dataset. Number of word-pairs in each dataset is shown in brackets against its name.

ing parameter values. Hence we follow the standard methodology of fixing parameters through cross validation. Specifically, we partition the data into five folds, four of which are used for training and one hold-out fold is used for testing. For each measure, the parameter values that achieve best correlation with human judgments on 4 training folds are used to predict on the 1 hold-out testing fold. This experiment is repeated 5 times for different training and test folds. The average rank correlation obtained by each measure over 5 cross-validation runs is reported for each dataset. We varied ϵ and δ between 0.01 and 0.90 and x between 5 and 50 words.

5.4 Results

For each measure and for each data set, the average correlation over the 5 cross-validation runs is reported in Table 4. The corresponding standard deviations are mentioned in the table's caption. The best performing measure in each case is highlighted in bold. While different measures performed best on different data sets, the results in Table 4 shows that CSR performs consistently well across all data sets. In all data sets the correlation for CSR was always either the best or close to the best.

As expected, our results are statistically more significant for the larger data sets, compared to the smaller ones. The standard deviations of the results are small for two largest data sets (less than 0.01 for Edinburg and Florida), gradually increasing (less than 0.02 for Kent and Minnesota), and becoming high (upto .15) for the three smallest datasets.

Although, among all measures, CSR has the best average correlation over all datasets, taking average of correlations across widely different dataset is not a meaningful way to decide on which measure to use. Ideally one would like to access an oracle to learn which measure will perform best on a particular unseen application dataset. Short of such an oracle, if one were to pick a fixed measure a-priori, then one would like to know how much worse off one is compared to the best measure for that dataset.

To compare different measures from this perspective, we compute the deviation of the correlation for each measure from the correlation of the best measure for each data set. These deviations are reported in Table 5, along with the corresponding ranks. The average deviation of CSR over all the data sets is 0.02, which is the least among all the measures, the next two being 0.04 and 0.06. CSR also has the least worst-deviation among all measures. Also, CSR is never ranked worse than 3 in any of the data sets. This is also the smallest worst-rank among all measures. Based on these results, we infer that CSR is overall the best performing co-occurrence based word association measure.

While the focus of our work is on the co-occurrence measures, for completeness, we present all the known results for knowledge and distributional similarity-based measures on the datasets un-

der consideration in Table 6. Note that in (Agirre et al., 2009), the *wordsim* data set was partitioned into two sets, namely *sim* and *rel*, and in *Esslli* shared task (ESSLLI, 2008), a 272 word pair subset of the *Edinburgh* dataset was chosen. To facilitate comparison, in addition to CSR, we also present results for PMI and Ochiai (Chi-Square) which are the best performing co-occurrence measures on *wordsim*, and *Esslli* datasets. For co-occurrence-based measures, we used 5-fold cross validation, which is inapplicable for parameterless measures. Results show that co-occurrence-based measures compare well with other resource-heavy measures.

6 Conclusions

In this paper, we introduced a new measure called CSR for word-association based on statistical significance of lexical co-occurrences. Our measure, while being agnostic to global unigram frequencies, detects skews in span distributions of word-pairs in documents containing both words. We carried out extensive evaluation on several benchmark datasets. Our experiments demonstrate the advantages of our measure over all the competing measures.

Acknowledgments

This work was supported in part by the Ministry of Human Resources Development, Government of India and by the Tata Research Development and Design Center (TRDDC). We thank Mr. Justin Washtell (University of Leeds) for providing us with various datasets.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL-HLT*.

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Measuring semantic similarity between words using web search engines. In *WWW*, pages 757–766.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguists*, 32(1):13–47.

Hsin-Hsi Chen, Ming-Shun Lin, and Yu-Chuan Wei. 2006. Novel association measures using web search with double checking. In *ACL*.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL*, pages 76–83.

L. R. Dice. 1945. Measures of the amount of ecological association between species. *Ecology*, 26:297–302.

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

ESSLLI. 2008. *Free association* task at lexical semantics workshop esslli 2008. `http://wordspace.collocations.de/doku.php/workshop:esslli:task`.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*.

Robert Goldfarb and Harvey Halpern. 1984. Word association responses in normal adult subjects. *Journal of Psycholinguistic Research*, 13(1):37–55.

T Hughes and D Ramage. 2007. Lexical semantic relatedness with random graph walks. In *EMNLP*.

P. Jaccard. 1912. The distribution of the flora of the alpine zone. *New Phytologist*, 11:37–50.

Svante Janson and Jan Vegelius. 1981. Measures of ecological association. *Oecologia*, 49:371–376.

Mario Jarmasz. 2003. Rogets thesaurus as a lexical resource for natural language processing. Technical report, University of Ottowa.

G. Kent and A. Rosanoff. 1910. A study of association in insanity. *American Journal 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 *The Computer and Literary Studies*, pages 379–382. Edinburgh University Press.

T. Landauer and S. Dumais. 1997. The latent semantic analysis theory of acquisition, induction, and representation of knowledge. In *Psychological Review*, volume 104/2, pages 211–240.

Sonya Liberman and Shaul Markovitch. 2009. Compact hierarchical explicit semantic representation. In *Proceedings of the IJCAI 2009 Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy (WikiAI09)*, Pasadena, CA, July.

G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *ACL*.

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.

A Ochiai. 1957. Zoogeografical studies on the soleoid fishes found in japan and its neighbouring regions-ii. *Bulletin of the Japanese Society of Scientific Fisheries*, 22.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *ACL*.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, October.

W.A. Russell and J.J. Jenkins. 1954. The complete minnesota norms for responses to 100 words from the kent-rosanoff word association test. Technical report, Office of Naval Research and University of Minnesota.

StopWordList. 2010. `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`. *The Information Retrieval Group, University of Glasgow*. Accessed: November 15, 2010.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, pages 1419–1424.

T. Wandmacher, E. Ovchinnikova, and T. Alexandrov. 2008. Does latent semantic analysis reflect human associations? In *European Summer School in Logic, Language and Information (ESSLLI'08)*.

Justin Washtell and Katja Markert. 2009. A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In *EMNLP*, pages 628–637.

Katherine K. White and Lise Abrams. 2004. Free associations and dominance ratings of homophones for young and older adults. *Behavior Research Methods, Instruments, & Computers*, 36(3):408–420.

Wikipedia. April 2008. `http://www.wikipedia.org`.

Eric Yeh, Daniel Ramage, Chris Manning, Eneko Agirre, and Aitor Soroa. 2009. Wikiwalk: Random walks on wikipedia for semantic relatedness. In *ACL workshop "TextGraphs-4: Graph-based Methods for Natural Language Processing"*.