Appropriately Incorporating Statistical Significance in PMI

Om P. Damani and Shweta Ghonge
IIT Bombay
India
{damani,shwetaghonge}@cse.iitb.ac.in

Abstract

Two recent measures incorporate the notion of statistical significance in basic PMI formulation. In some tasks, we find that the new measures perform worse than the PMI. Our analysis shows that while the basic ideas in incorporating statistical significance in PMI are reasonable, they have been applied slightly inappropriately. By fixing this, we get new measures that improve performance over not just PMI but on other popular co-occurrence measures as well. In fact, the revised measures perform reasonably well compared with more resource intensive non co-occurrence based methods also.

1 Introduction

The notion of *word association* is used in many language processing and information retrieval applications and it is important to have low-cost, high-quality association measures. Lexical co-occurrence based word association measures are popular because they are computationally efficient and they can be applied to any language easily. One of the most popular co-occurrence measure is Pointwise Mutual Information (PMI) (Church and Hanks, 1989).

One of the limitations of PMI is that it only works with relative probabilities and ignores the absolute amount of evidence. To overcome this, recently two new measures have been proposed that incorporate the notion of statistical significance in basic PMI formulation. In (Washtell and Markert, 2009), statistical significance is introduced in PMI$_{sig}$ by multiplying PMI value with the square root of the evidence. In contrast, in (Damani, 2013), cPMId is introduced by bounding the probability of observing a given deviation between a given word pair's co-occurrence count and its expected value under a null model where with each word a global unigram generation probability is associated. In Table 1, we give the definitions of PMI, PMI$_{sig}$, and cPMId.

While these new measures perform better than PMI on some of the tasks, on many other tasks, we find that the new measures perform worse than the PMI. In Table 3, we show how these measures perform compared to PMI on four different tasks. We find that PMI$_{sig}$ degrades performance in three out of these four tasks while cPMId degrades performance in two out of these four tasks. The experimental details and discussion are given in Section 4.2.

Our analysis shows that while the basic ideas in incorporating statistical significance are reasonable, they have been applied slightly inappropriately. By fixing this, we get new measures that improve performance over not just PMI, but also on other popular co-occurrence measures on most of these tasks. In fact, the revised measures perform reasonably well compared with more resource intensive non co-occurrence based methods also.

2 Adapting PMI for Statistical Significance

In (Washtell and Markert, 2009), it is assumed that the statistical significance of a word pair association is proportional to the square root of the evidence. The question of what constitutes the evidence is answered by taking the lesser of the frequencies of the two words in the word pair, since at most that many pairings are possible. Hence the PMI value is multi-

Method	Formula	Revised Formula
PMI (Church and Hanks, 1989)	$log \frac{f(x,y)}{f(x)*f(y)/W}$	
PMI$_{sig}$ (Washtell and Markert, 2009)	$log \frac{f(x,y)}{f(x)*f(y)/W} * \sqrt{\mathbf{min(f(x),f(y))}}$	PMIs: $log \frac{f(x,y)}{f(x)*f(y)/W} * \sqrt{\mathbf{max(f(x),f(y))}}$
cPMId (Damani, 2013)	$log \frac{d(x,y)}{\mathbf{d(x)*d(y)}/D+\sqrt{\mathbf{d(x)}}*\sqrt{\frac{\ln \delta}{(-2.0)}}}$	sPMId: $log \frac{d(x,y)}{\mathbf{max(d(x),d(y))*min(d(x),d(y))}/D+\sqrt{\mathbf{max(d(x),d(y))}}*\sqrt{\frac{\ln \delta}{(-2.0)}}}$

Terminology:

W	Total number of words in the corpus	D	Total number of documents in the corpus
$f(x), f(y)$	unigram frequencies of x, y respectively in the corpus	$d(x), d(y)$	Total number of documents in the corpus containing at least one occurrence of x and y respectively
$f(x, y)$	Span-constrained (x, y) word pair frequency in the corpus	$d(x, y)$	Total number of documents in the corpus having at-least one span-constrained occurrence of the word pair (x, y)
δ	a parameter varying between 0 and 1		

Table 1: Definitions of PMI and its statistically significant adaptations. The sub-parts in bold represent the changes between the original formulas and the revised formulas. The product $max(d(x), d(y)) * min(d(x), d(y))$ in sPMId formula can be simplified to $f(x) * f(y)$, however, we left it this way to emphasize the transformation from cPMId.

plied by $\sqrt{min(f(x), f(y))}$ to get PMI$_{sig}$.

In (Damani, 2013), statistical significance is introduced by bounding the probability of observing a given number of word-pair occurrences in the corpus, just by chance, under a null model of independent unigram occurrences. For this computation, one needs to decide what constitutes a random trial when looking for a word-pair occurrence. Is it the occurrence of the first word (say x) in the pair, or the second (say y). In (Damani, 2013), occurrences of x are arbitrarily chosen to represent the sites of the random trial. Using Hoeffdings Inequality:

$$P[f(x, y) \geq f(x) * f(y)/W + f(x) * t] \\ \leq \exp(-2 * f(x) * t^2)$$

By setting $t = \sqrt{\ln \delta/(-2 * f(x))}$, we get δ as an upper bound on probability of observing more than $f(x) * f(y)/W + f(x) * t$ bigram occurrences in the corpus, just by chance. Based on this *Corpus Level Significant PMI*(cPMI) is defined as:

$$cPMI(x, y) = log \frac{f(x, y)}{f(x) * f(y)/W + f(x) * t}$$

$$= log \frac{f(x, y)}{f(x) * f(y)/W + \sqrt{f(x)} * \sqrt{\ln \delta/(-2)}}$$

In (Damani, 2013), several variants of cPMI are introduced that incorporate different notions of statistical significance. Of these *Corpus Level Significant PMI based on Document count*(cPMId - defined in Table 1) is found to be the best performing, and hence we consider this variant only in this work.

2.1 Choice of Random Trial

While considering statistical significance, one has to decide what constitutes a random trial. When looking for a word-pair (x, y)'s occurrences, y can potentially occur near each occurrence of x, or x can potentially occur near each occurrence of y. Which of these two set of occurrences should be considered the sites of random trial. We believe that the occurrences of the more frequent of x and y should be considered, since near each of these occurrences the other word could have occurred. Hence $f(x)$ and $f(y)$ in cPMI definition should be replaced with $max(f(x), f(y))$ and $min(f(x), f(y))$ respectively. Similarly, $d(x)$ and $d(y)$ in cPMId formula should be replaced with $max(d(x), d(y))$ and $min(d(x), d(y))$ respectively to give a new measure *Significant PMI based on Document count*(sPMId).

Using the same logic, $\sqrt{min(f(x), f(y))}$ in PMI$_{sig}$ formula should be replaced with $\sqrt{max(f(x), f(y))}$ to give the formula for a new measure *PMI-significant*(PMIs). The definitions of sPMId and PMIs are also given in Table 1.

3 Related Work

There are three main types of word association measures: Knowledge based, Distributional Similarity based, and Lexical Co-occurrence based.

Based on Firth's *You shall know a word by the company it keeps* (Firth, 1957), distributional similarity based measures characterize a word by the distribution of other words around it and compare

Method	Formula
ChiSquare (χ^2)	$\sum_{i,j} \frac{(f(i,j)-Ef(i,j))^2}{Ef(i,j)}$
Dice (Dice, 1945)	$\frac{f(x,y)}{f(x)+f(y)}$
GoogleDistance (L.Cilibrasi and Vitany, 2007)	$\frac{max(\log d(x),\log d(y))-\log d(x,y)}{\log D-min(\log d(x),\log d(y))}$
Jaccard (Jaccard, 1912)	$\frac{f(x,y)}{f(x)+f(y)-f(x,y)}$
LLR (Dunning, 1993)	$\sum_{\substack{x' \in \{x,\neg x\} \\ y' \in \{y,\neg y\}}} f(x',y') log \frac{f(x',y')}{f(x')f(y')}$
nPMI (Bouma, 2009)	$\frac{log \frac{f(x,y)}{f(x)*f(y)/W}}{log \frac{1}{f(x,y)/W}}$
Ochiai (Janson and Vegelius, 1981)	$\frac{f(x,y)}{\sqrt{f(x)f(y)}}$
PMI^2 (Daille, 1994)	$log \frac{\frac{f(x,y)}{f(x)*f(y)/W}}{\frac{1}{f(x,y)/W}} = log \frac{f(x,y)^2}{f(x)*f(y)}$
Simpson (Simpson, 1943)	$\frac{f(x,y)}{min(f(x),f(y))}$
SCI (Washtell and Markert, 2009)	$\frac{f(x,y)}{f(x)\sqrt{f(y)}}$
T-test	$\frac{f(x,y)-Ef(x,y)}{\sqrt{f(x,y)\left(1-\frac{f(x,y)}{W}\right)}}$

Table 2: Definition of other co-occurrence measures being compared in this work. The terminology used here is same as that in Table 1, except that E in front of a variable name means the expected value of that variable.

Task	Semantic Relatedness	Sentence Similarity	Synonym Selection	
Dataset	WordSim	Li	ESL	TOEFL
Metric	Spearman Rank Correlation	Pearson Correlation	Fraction Correct	Fraction Correct
PMI	0.68	0.69	0.62	0.59
PMI$_{sig}$	0.67	**0.85**	0.58	0.56
cPMId	**0.72**	0.67	0.56	0.59
PMIs	0.66	**0.85**	0.66	**0.61**
sPMId	**0.72**	0.75	**0.70**	**0.61**
ChiSquare (χ^2)	0.62	0.80	0.62	0.58
Dice	0.58	0.76	0.56	0.57
GoogleDistance	0.53	0.75	0.09	0.19
Jaccard	0.58	0.76	0.56	0.57
LLR	0.50	0.18	0.18	0.27
nPMI	**0.72**	0.35	0.54	0.54
Ochiai/ PMI^2	0.62	0.77	0.62	0.60
SCI	0.65	**0.85**	0.62	0.60
Simpson	0.59	0.78	0.58	0.57
TTest	0.44	0.63	0.44	0.52
Semantic Net (Li et al., 2006)		0.82		
ESA (Gabrilovich and Markovitch, 2007)	0.74			
(reimplemented in (Yeh et al., 2009))	0.71			
Distributional Similarity (on web corpus) (Agirre et al., 2009))	0.65			
Context Window based Distributional Similarity (Agirre et al., 2009))	0.60			
Latent Semantic Analysis (on web corpus) (Finkelstein et al., 2002)	0.56			
WordNet::Similarity (Recchia and Jones, 2009)			0.70	0.87
PMI-IR3 (using context) (Turney, 2001)				0.73

Table 3: 5-fold cross-validation results for different co-occurrence measures. The results for the best, and second best co-occurrence measures for each data-set is shown in bold and underline respectively. Except GoogleDistance and LLR, all results for all co-occurrence measures are statistically significant at $p = .05$. For each task, the best known result for different non co-occurrence based methods is also shown.

two words for distributional similarity (Agirre et al., 2009; Wandmacher et al., 2008; Bollegala et al., 2007; Chen et al., 2006). They are also used for modeling the meaning of a phrase or a sentence (Grefenstette and Sadrzadeh, 2011; Wartena, 2013; Mitchell, 2011; G. Dinu and Baroni, 2013; Kartsaklis et al., 2013).

Knowledge-based measures 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).

Co-occurrence based measures (Pecina and Schlesinger, 2006) simply rely on unigram and bigram frequencies of the words in a pair. In this work, our focus is on the co-occurrence based measures, since they are resource-light and can easily be used for resource-scarce languages.

3.1 Co-occurrence Measures being Compared

Co-occurrence based measures of association between two entities are used in several domains like ecology, psychology, medicine, language processing, etc. To compare the performance of our newly introduced measures with other co-occurrence measures, we have selected a number of popular co-occurrence measures like ChiSquare (χ^2), Dice (Dice, 1945), GoogleDistance (L.Cilibrasi and Vitany, 2007), Jaccard (Jaccard, 1912), LLR (Dunning, 1993), Simpson (Simpson, 1943), and T-test from these domains.

In addition to these popular measures, we also experiment with other known variations of PMI like nPMI (Bouma, 2009), PMI^2 (Daille, 1994), Ochiai (Janson and Vegelius, 1981), and SCI (Washtell and Markert, 2009). Since PMI^2 is a monotonic transformation of Ochiai, we present their results together. In Table 2, we present the definitions of these measures. While the motivation given for SCI in (Washtell and Markert, 2009) is slightly different, in light of the discussion in Section 2.1, we can assume that SCI is PMI adapted for statistical significance (multiplied by $\sqrt{\mathbf{f(y)}}$), where the site of random trial is taken to be the occurrences of the second word y, instead of the less frequent word, as in the case of PMI$_{sig}$.

When counting co-occurrences, we only consider the non-overlapping *span*-constrained occurrences. The span of a word-pair's occurrence is the direction-independent distance between the occurrences of the members of the pair. We consider only those co-occurrences where span is less than a given threshold. Therefore, span threshold is a parameter for all the co-occurrence measures being considered.

4 Performance Evaluation

Having introduced the revised measures PMIs and sPMId, we need to evaluate the performance of these measures compared to PMI and the original measures introducing significance. In addition, we also wish to compare the performance of these measures with other co-occurrence measures. To compare the performance of these measures with more resource heavy non co-occurrence based measures, we have chosen those tasks and datasets on which published results exist for distributional similarity and knowledge based word association measures.

4.1 Task Details

We evaluate these measures on three tasks: Sentence Similarity(65 sentence-pairs from (Li et al., 2006)), Synonym Selection(50 questions ESL (Turney, 2001) and 80 questions TOEFL (Landauer and Dutnais, 1997) datasets), and, Semantic Relatedness(353 words Wordsim (Finkelstein et al., 2002) dataset).

For each of these tasks, gold standard human judgment results exist. For sentence similarity, following (Li et al., 2006), we evaluate a measure by the Pearsons correlation between the ranking produced by the measure and the human ranking. For synonym selection, we compute the percentage of correct answers, since there is a unique answer for each challenge word in the datasets. Semantic relatedness has been evaluated by Spearman's rank correlation with human judgment instead of Pearsons correlation in literature and we follow the same practice to make results comparable.

For sentence similarity detection, the algorithm used by us (Li et al., 2006) assumes that the association scores are between 0 and 1. Hence we normalize the value produced by each measure using

Challenge x	Option y (correct)	Option z (incorrect)	$f(x)$	$f(y)$	$f(z)$	$f(x,y)$	$f(x,z)$	PMI$_{sig}$ (x,y)	PMI$_{sig}$ (x,z)	PMIs (x,y)	PMIs (x,z)
brass	metal	plastic	15923	125088	24985	228	75	14	24	40	30
twist	intertwine	curl	11407	153	2047	1	9	7	17	61	41
saucer	dish	frisbee	2091	12453	1186	5	1	9	14	21	18
mass	lump	element	90398	1595	43321	14	189	4	10	29	15
applause	approval	friends	1998	19673	11689	8	6	9	11	29	28
confession	statement	plea	7687	47299	5232	76	12	18	22	45	26
swing	sway	bounce	33580	2994	4462	13	17	7	8	24	21
sheet	leaf	book	20470	20979	586581	20	194	7	2	7	12

Table 4: Details of ESL word-pairs, correctness of whose answers changes between PMI$_{sig}$ and PMIs. Except for the gray-row, for all other questions, incorrect answers becomes correct on using PMIs instead of PMI$_{sig}$, and vice-versa for the gray-row. The association values have been suitably scaled for readability. To save space, of the four choices, options not selected by either of the methods have been omitted. These results are for a 10 word span.

max-min normalization:

$$v' = \frac{v - min}{max - min}$$

where *max and* min are computed over all association scores for the entire task for a given measure.

4.2 Experimental Results

We use a 1.24 Gigawords Wikipedia corpus for getting co-occurrence statistics. Since co-occurrence methods have span-threshold as a parameter, we follow the standard methodology of five-fold cross validation. Note that, in addition to span-threshold, cPMId and sPMId have an additional parameter δ.

In Table 3, we present the performance of all the co-occurrence measures considered on all the tasks. Note that, except GoogleDistance and LLR, all results for all co-occurrence measures are statistically significant at p = .05. For completeness of comparison, we also include the best known results from literature for different non co-occurrence based word association measures on these tasks.

4.3 Performance Analysis and Conclusions

We find that on average, PMI$_{sig}$ and cPMId, the recently introduced measures that incorporate significance in PMI, do not perform better than PMI on the given datasets. Both of them perform worse than PMI on three out of four datasets. By appropriately incorporating significance, we get new measures PMIs and sPMId that perform better than PMI(also PMI$_{sig}$ and cPMId respectively) on most datasets. PMIs improves performance over PMI on three out of four datasets, while sPMId improves performance on all four datasets.

The performance improvement of PMIs over PMI$_{sig}$ and of sPMId over cPMId, is not random. For example, on the ESL dataset, while the percentage of correct answers increases from 58 to 66 from PMI$_{sig}$ to PMIs, it is not the case that on moving from PMI$_{sig}$ to PMIs, several correct answers become incorrect and an even larger number of incorrect answers become correct. As shown in Table 4, only one correct answers become incorrect while seven incorrect answers get corrected. The same trend holds for most parameters values, and for moving from cPMId to sPMId. This substantiates the claim that the improvement is not random, but due to the appropriate incorporation of significance, as discussed in Section 2.1.

PMIs and sPMId perform better than not just PMI, but they perform better than all popular co-occurrence measures on most of these tasks. When compared with any other co-occurrence measure, on three out of four datasets each, both PMIs and sPMId perform better than that measure. In fact, PMIs and sPMId perform reasonably well compared with more resource intensive non co-occurrence based methods as well. Note that different non co-occurrence based measures perform well on different tasks. We are comparing the performance of a single measure (say sPMId or PMIs) against the best measure for each task.

Acknowledgements

We thank Dipak Chaudhari for his help with the implementation.

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 2009, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction, from form to meaning: Processing texts automatically. In *GSCL 2009, Proceedings of the Biennial International Conference of the German Society for Computational Linguistics and Language Technology*.

Hsin-Hsi Chen, Ming-Shun Lin, and Yu-Chuan Wei. 2006. Novel association measures using web search with double checking. In *COLING/ACL 2006, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL 1989, Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 76–83.

B. Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales etl-tres linguistiques*. Ph.D. thesis, Universitie Paris 7.

Om P. Damani. 2013. Improving pointwise mutual information (pmi) by incorporating significant co-occurrence. In *CoNLL 2013, Conference on Computational Natural Language Learning*.

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.

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

J. R. Firth. 1957. A synopsis of linguistics theory. *Studies in Linguistic Analysis*, pages 1930–1955.

N. Pham G. Dinu and M. Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *CVSC 2013, Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007, International Joint Conference on Artificial Intelligence*.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *EMNLP 2011, Conference on Empirical Methods on Natural Language Processing*, pages 1394–1404.

T Hughes and D Ramage. 2007. Lexical semantic relatedness with random graph walks. In *EMNLP 2007, Conference on Empirical Methods on Natural Language Processing*.

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.

Dimitrios Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *CoNLL 2013, Conference on Computational Natural Language Learning*.

Thomas K Landauer and Susan T. Dutnais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Rudi L.Cilibrasi and Paul M.B. Vitany. 2007. The google similarity distance. *Psychological review*, 19(3).

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, August.

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

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *ACL 2008, Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Jeffrey Mitchell. 2011. *Composition in Distributional Models of Semantics*. Ph.D. thesis, The University of Edinburgh.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *ACL 2006, Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Gabriel Recchia and Michael N. Jones. 2009. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 3(41):647–656.

George G. Simpson. 1943. Mammals and the nature of continents. *American Journal of Science*, pages 1–31.

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

P. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *ECML 2001, European Conference on Machine Learning*.

T. Wandmacher, E. Ovchinnikova, and T. Alexandrov. 2008. Does latent semantic analysis reflect human associations? In *ESSLLI 2008, European Summer School in Logic, Language and Information*.

Christian Wartena. 2013. Hsh: Estimating semantic similarity of words and short phrases with frequency normalized distance measures. In *SemEval 2013, International Workshop on Semantic Evaluation*.

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

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