A Re-examination of Lexical Association Measures

Hung Huu Hoang
Dept. of Computer Science
National University
of Singapore
hoanghuu@comp.nus.edu.sg

Su Nam Kim
Dept. of Computer Science
and Software Engineering
University of Melbourne
snkim@csse.unimelb.edu.au

Min-Yen Kan
Dept. of Computer Science
National University
of Singapore
kanmy@comp.nus.edu.sg

Abstract

We review lexical Association Measures (AMs) that have been employed by past work in extracting multiword expressions. Our work contributes to the understanding of these AMs by categorizing them into two groups and suggesting the use of rank equivalence to group AMs with the same ranking performance. We also examine how existing AMs can be adapted to better rank English verb particle constructions and light verb constructions. Specifically, we suggest normalizing (Pointwise) Mutual Information and using marginal frequencies to construct penalization terms. We empirically validate the effectiveness of these modified AMs in detection tasks in English, performed on the Penn Treebank, which shows significant improvement over the original AMs.

1 Introduction

Recently, the NLP community has witnessed a renewed interest in the use of lexical association measures in extracting Multiword Expressions (MWEs). Lexical Association Measures (hereafter, AMs) are mathematical formulas which can be used to capture the degree of connection or association between constituents of a given phrase. Well-known AMs include *Pointwise Mutual Information* (*PMI*), *Pearson's* χ^2 and the *Odds Ratio*. These AMs have been applied in many different fields of study, from information retrieval to hypothesis testing. In the context of MWE extraction, many published works have been devoted to comparing their effectiveness. Krenn and Evert (2001) evaluate *Mutual Information* (*MI*), *Dice*, *Pearson's* χ^2, *log-likelihood ratio* and the *T score*. In Pearce (2002), AMs such as *Z score*, *Pointwise MI*, *cost reduction*, *left and right context entropy*, *odds ratio* are evaluated. Evert (2004) discussed a wide range of AMs, including exact hypothesis tests such as *the binomial test* and *Fisher's exact test*s, various coefficients such as *Dice* and *Jaccard*. Later, Ramisch *et al.* (2008) evaluated *MI*, *Pearson's* χ^2 and *Permutation Entropy*. Probably the most comprehensive evaluation of AMs was presented in Pecina and Schlesinger (2006), where 82 AMs were assembled and evaluated over Czech collocations. These collocations contained a mix of idiomatic expressions, technical terms, light verb constructions and stock phrases. In their work, the best combination of AMs was selected using machine learning.

While the previous works have evaluated AMs, there have been few details on why the AMs perform as they do. A detailed analysis of why these AMs perform as they do is needed in order to explain their identification performance, and to help us recommend AMs for future tasks. This weakness of previous works motivated us to address this issue. In this work, we contribute to further understanding of association measures, using two different MWE extraction tasks to motivate and concretize our discussion. Our goal is to be able to predict, *a priori*, what types of AMs are likely to perform well for a particular MWE class.

We focus on the extraction of two common types of English MWEs that can be captured by bigram model: Verb Particle Constructions (VPCs) and Light Verb Constructions (LVCs). VPCs consist of a verb and one or more particles, which can be prepositions (e.g. *put on*, *bolster up*), adjectives (*cut short*) or verbs (*make do*). For simplicity, we focus only on bigram VPCs that take prepositional particles, the most common class of VPCs. A special characteristic of VPCs that affects their extraction is the

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 31–39,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

mobility of noun phrase complements in transitive VPCs. They can appear after the particle (*Take off your hat*) or between the verb and the particle (*Take your hat off*). However, a pronominal complement can only appear in the latter configuration (*Take it off*).

In comparison, LVCs comprise of a verb and a complement, which is usually a noun phrase (*make a presentation*, *give a demonstration*). Their meanings come mostly from their complements and, as such, verbs in LVCs are termed semantically light, hence the name *light verb*. This explains why modifiers of LVCs modify the complement instead of the verb (*make a serious mistake* vs. **make a mistake seriously*). This phenomenon also shows that an LVC's constituents may not occur contiguously.

2 Classification of Association Measures

Although different AMs have different approaches to measuring association, we observed that they can effectively be classified into two broad classes. *Class I* AMs look at the degree of institutionalization; i.e., the extent to which the phrase is a semantic unit rather than a free combination of words. Some of the AMs in this class directly measure this association between constituents using various combinations of co-occurrence and marginal frequencies. Examples include *MI*, *PMI* and their variants as well as most of the association coefficients such as *Jaccard*, *Hamann*, *Brawn-Blanquet*, and others. Other Class I AMs estimate a phrase's MWE-hood by judging the significance of the difference between observed and expected frequencies. These AMs include, among others, statistical hypothesis tests such as *T score*, *Z score* and *Pearson's* χ^2 *test*.

Class II AMs feature the use of context to measure non-compositionality, a peculiar characteristic of many types of MWEs, including VPCs and idioms. This is commonly done in one of the following two ways. First, non-compositionality can be modeled through the diversity of contexts, measured using entropy. The underlying assumption of this approach is that non-compositional phrases appear in a more restricted set of contexts than compositional ones. Second, non-compositionality can also be measured through context similarity between the phrase and its constituents. The observation here is that non-compositional phrases have different semantics from those of their constituents. It then follows that contexts in which the phrase and its constituents appear would be different (Zhai, 1997). Some VPC examples include *carry out*, *give up*. A close approximation stipulates that contexts of a non-compositional phrase's constituents are also different. For instance, phrases such as *hot dog* and *Dutch courage* are comprised of constituents that have unrelated meanings. Metrics that are commonly used to compute context similarity include *cosine* and *dice similarity*; distance metrics such as *Euclidean* and *Manhattan norm*; and probability distribution measures such as *Kullback-Leibler divergence* and *Jensen-Shannon divergence*.

Table 1 lists all AMs used in our discussion. The lower left legend defines the variables *a, b, c*, and *d* with respect to the raw co-occurrence statistics observed in the corpus data. When an AM is introduced, it is prefixed with its index given in Table 1(e.g., [M2] Mutual Information) for the reader's convenience.

3 Evaluation

We will first present how VPC and LVC candidates are extracted and used to form our evaluation data set. Second, we will discuss how performances of AMs are measured in our experiments.

3.1 Evaluation Data

In this study, we employ the *Wall Street Journal* (WSJ) section of one million words in the Penn Tree Bank. To create the evaluation data set, we first extract the VPC and LVC candidates from our corpus as described below. We note here that the mobility property of both VPC and LVC constituents have been used in the extraction process.

For VPCs, we first identify particles using a pre-compiled set of 38 particles based on Baldwin (2005) and Quirk *et al.* (1985) (Appendix A). Here we do not use the WSJ particle tag to avoid possible inconsistencies pointed out in Baldwin (2005). Next, we search to the left of the located particle for the nearest verb. As verbs and particles in transitive VPCs may not occur contiguously, we allow an intervening NP of up to 5 words, similar to Baldwin and Villavicencio (2002) and Smadja (1993), since longer NPs tend to be located after particles.

AM Name	Formula	AM Name	Formula
M1. Joint Probability	$f(xy)/N$	M2. Mutual Information	$\frac{1}{N}\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$
M3. Log likelihood ratio	$2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$	M4. Pointwise MI (PMI)	$\log \frac{P(xy)}{P(x*)P(*y)}$
M5. Local-PMI	$f(xy) \times \text{PMI}$	M6. PMI^k	$\log \frac{Nf(xy)^k}{f(x*)f(*y)}$
M7. PMI^2	$\log \frac{Nf(xy)^2}{f(x*)f(*y)}$	M8. Mutual Dependency	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
M9. Driver-Kroeber	$\frac{a}{\sqrt{(a+b)(a+c)}}$	M10. Normalized expectation	$\frac{2a}{2a+b+c}$
M11. Jaccard	$\frac{a}{a+b+c}$	M12. First Kulczynski	$\frac{a}{b+c}$
M13. Second Sokal-Sneath	$\frac{a}{a+2(b+c)}$	M14. Third Sokal-Sneath	$\frac{a+d}{b+c}$
M15. Sokal-Michiner	$\frac{a+d}{a+b+c+d}$	M16. Rogers-Tanimoto	$\frac{a+d}{a+2b+2c+d}$
M17. Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$	M18. Odds ratio	$\frac{ad}{bc}$
M19. Yule's ω	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	M20. Yule's Q	$\frac{ad-bc}{ad+bc}$
M21. Brawn-Blanquet	$\frac{a}{\max(a+b, a+c)}$	M22. Simpson	$\frac{a}{\min(a+b, a+c)}$
M23. S cost	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$	M24*. Adjusted S Cost	$\log(1+\frac{\max(b,c)}{a+1})^{-\frac{1}{2}}$
M25. Laplace	$\frac{a+1}{a+\min(b, c)+2}$	M26*. Adjusted Laplace	$\frac{a+1}{a+\max(b, c)+2}$
M27. Fager	$[\text{M9}]-\frac{1}{2}\max(b,c)$	M28*. Adjusted Fager	$[\text{M9}]-\frac{1}{\sqrt{aN}}\max(b,c)$
M29*. Normalized PMIs	PMI / NF(α) PMI / NFMax	M30*. Simplified normalized PMI for VPCs	$\frac{\log(ad)}{\alpha \times b + (1-\alpha) \times c}$
M31*. Normalized MIs	MI / NF(α) MI / NFMax	NF(α) = $\alpha P(x*) + (1-\alpha)P(*y)$ $\alpha \in [0, 1]$ NFMax = $\max(P(x*), P(*y))$	

$a = f_{11} = f(xy)$ $b = f_{12} = f(x\bar{y})$ $c = f_{21} = f(\bar{x}y)$ $d = f_{22} = f(\bar{x}\bar{y})$	$f(x*)$ $f(\bar{x}*)$
$f(*y)$ $f(*\bar{y})$	N

Contingency table of a bigram (x y), recording co-occurrence and marginal frequencies; $\bar{w}$ stands for all words except w; * stands for all words; N is total number of bigrams. The expected frequency under the independence assumption is $\hat{f}(xy) = f(x*)f(*y)/N$.

Table 1. Association measures discussed in this paper. Starred AMs (*) are developed in this work.

Extraction of LVCs is carried out in a similar fashion. First, occurrences of light verbs are located based on the following set of seven frequently used English light verbs: *do*, *get*, *give*, *have*, *make*, *put* and *take*. Next, we search to the right of the light verbs for the nearest noun,

permitting a maximum of 4 intervening words to allow for quantifiers (*a*/*an*, *the*, *many*, etc.), adjectival and adverbial modifiers, etc. If this search fails to find a noun, as when LVCs are used in the passive (e.g. *the presentation was made*), we search to the right of the light verb, also allowing a maximum of 4 intervening words. The above extraction process produced a total of 8,652 VPC and 11,465 LVC candidates when run on the corpus. We then filter out candidates with observed frequencies less than 6, as suggested in Pecina and Schlesinger (2006), to obtain a set of 1,629 VPCs and 1,254 LVCs.

Separately, we use the following two available sources of annotations: 3,078 VPC candidates extracted and annotated in (Baldwin, 2005) and 464 annotated LVC candidates used in (Tan *et al.*, 2006). Both sets of annotations give both positive and negative examples.

Our final VPC and LVC evaluation datasets were then constructed by intersecting the gold-standard datasets with our corresponding sets of extracted candidates. We also concatenated both sets of evaluation data for composite evaluation. This set is referred to as "Mixed". Statistics of our three evaluation datasets are summarized in Table 2.

	VPC data	LVC data	Mixed
Total (*freq* ≥ 6)	413	100	513
Positive instances	117 (28.33%)	28 (28%)	145 (23.26%)

Table 2. Evaluation data sizes (type count, not token).

While these datasets are small, our primary goal in this work is to establish initial comparable baselines and describe interesting phenomena that we plan to investigate over larger datasets in future work.

3.2 Evaluation Metric

To evaluate the performance of AMs, we can use the standard precision and recall measures, as in much past work. We note that the ranked list of candidates generated by an AM is often used as a classifier by setting a threshold. However, setting a threshold is problematic and optimal threshold values vary for different AMs. Additionally, using the list of ranked candidates directly as a classifier does not consider the confidence indicated by actual scores. Another way to avoid setting threshold values is to measure precision and recall of only the *n* most likely candidates (the *n*-best method). However, as discussed in Evert and Krenn (2001), this method depends heavily on the choice of *n*. In this paper, we opt for average precision (AP), which is the average of precisions at all possible recall values. This choice also makes our results comparable to those of Pecina and Schlesinger (2006).

3.3 Evaluation Results

Figure 1(a, b) gives the two average precision profiles of the 82 AMs presented in Pecina and Schlesinger (2006) when we replicated their experiments over our English VPC and LVC datasets. We observe that the average precision profile for VPCs is slightly concave while the one for LVCs is more convex. This can be interpreted as VPCs being more sensitive to the choice of AM than LVCs. Another point we observed is that a vast majority of Class I AMs, including PMI, its variants and association coefficients (excluding hypothesis tests), perform reasonably well in our application. In contrast, the performances of most of context-based and hypothesis test AMs are very modest. Their mediocre performance indicates their inapplicability to our VPC and LVC tasks. In particular, the high frequencies of particles in VPCs and light verbs in LVCs both undermine their contexts' discriminative power and skew the difference between observed and expected frequencies that are relied on in hypothesis tests.

4 Rank Equivalence

We note that some AMs, although not mathematically equivalent (i.e., assigning identical scores to input candidates) produce the same lists of ranked candidates on our datasets. Hence, they achieve the same average precision. The ability to identify such groups of AMs is helpful in simplifying their formulas, which in turn assisting in analyzing their meanings.

Definition: *Association measures M_1 and M_2 are rank equivalent over a set C, denoted by $M_1 \underset{C}{\overset{r}{\equiv}} M_2$, if and only if $M_1(c_j) > M_1(c_k) \Leftrightarrow M_2(c_j) > M_2(c_k)$ and $M_1(c_j) = M_1(c_k) \Leftrightarrow M_2(c_j) = M_2(c_k)$ for all c_j, c_k belongs to C where $M_k(c_i)$ denotes the score assigned to c_i by the measure M_k.*

As a corollary, the following also holds for rank equivalent AMs:

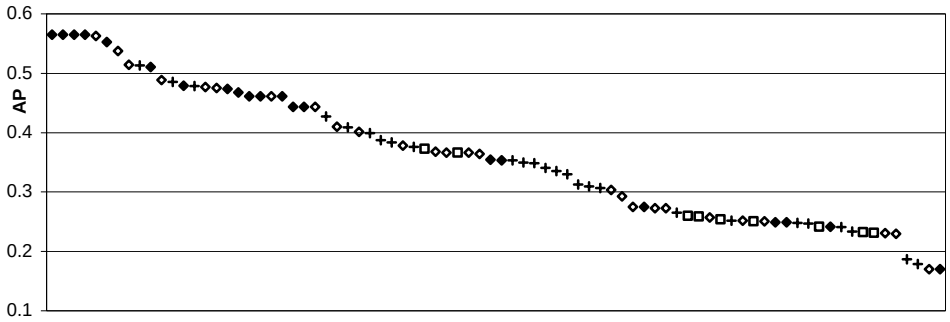

Figure 1a. AP profile of AMs examined over our VPC data set.

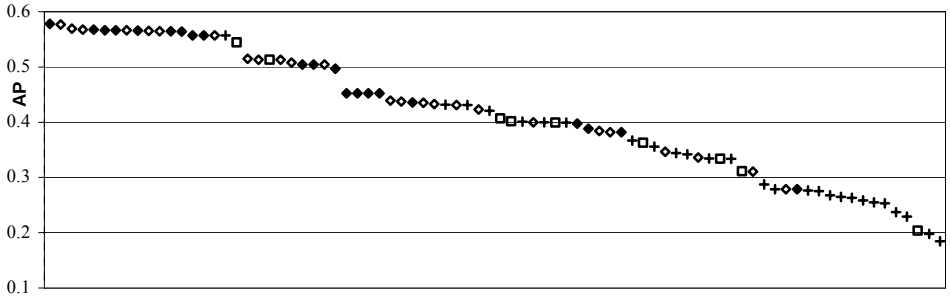

Figure 1b. AP profile of AMs examined over our LVC data set.

Figure 1. Average precision (AP) performance of the 82 AMs from Pecina and Schlesinger (2006), on our English VPC and LVC datasets. Bold points indicate AMs discussed in this paper.
□ Hypothesis test AMs ◊ Class I AMs, excluding hypothesis test AMs + Context-based AMs.

Corollary: *If* $M_1 \underset{C}{\overset{r}{\equiv}} M_2$ *then* $AP_C(M_1) = AP_C(M_2)$ *where* $AP_C(M_i)$ *stands for the average precision of the AM* M_i *over the data set C.*

Essentially, M_1 and M_2 are rank equivalent over a set C if their ranked lists of all candidates taken from C are the same, ignoring the actual calculated scores[1]. As an example, the following 3 AMs: *Odds ratio*, *Yule's ω* and *Yule's Q* (Table 3, row 5), though not mathematically equivalent, can be shown to be rank equivalent. Five groups of rank equivalent AMs that we have found are listed in Table 3. This allows us to replace the below 15 AMs with their (most simple) representatives from each rank equivalent group.

1) [M2] Mutual Information, [M3] Log likelihood ratio
2) [M7] PMI^2, [M8] Mutual Dependency, [M9] Driver-Kroeber (a.k.a. Ochiai)
3) [M10] Normalized expectation, [M11] Jaccard, [M12] First Kulczynski, [M13]Second Sokal-Sneath (a.k.a. Anderberg)
4) [M14] Third Sokal-Sneath, [M15] Sokal-Michiner, [M16] Rogers-Tanimoto, [M17] Hamann
5) [M18] Odds ratio, [M19] Yule's ω, [M20] Yule's Q

Table 3. Five groups of rank equivalent AMs.

5 Examination of Association Measures

We highlight two important findings in our analysis of the AMs over our English datasets. Section 5.1 focuses on MI and PMI and Section 5.2 discusses penalization terms.

5.1 Mutual Information and Pointwise Mutual Information

In Figure 1, over 82 AMs, PMI ranks 11th in identifying VPCs while MI ranks 35th in

[1] Two AMs may be rank equivalent with the exception of some candidates where one AM is undefined due to a zero in the denominator while the other AM is still well-defined. We call these cases *weakly rank equivalent*. With a reasonably large corpus, such candidates are rare for our VPC and LVC types. Hence, we still consider such AM pairs to be rank equivalent.

identifying LVCs. In this section, we show how their performances can be improved significantly.

Mutual Information (MI) measures the common information between two variables or the reduction in uncertainty of one variable given knowledge of the other. $\mathrm{MI}(U; V) = \sum_{u,v} p(uv) \log \frac{p(uv)}{p(u*)p(*v)}$. In the context of bigrams, the above formula can be simplified to [M2] $\mathrm{MI} = \frac{1}{\mathrm{N}} \sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$. While MI holds between random variables, [M4] Pointwise MI (PMI) holds between specific values: $\mathrm{PMI}(x, y) = \log \frac{P(xy)}{P(x*)P(*y)} = \log \frac{Nf(xy)}{f(x*)f(*y)}$. It has long been pointed out that PMI favors bigrams with low-frequency constituents, as evidenced by the product of two marginal frequencies in its denominator. To reduce this bias, a common solution is to assign more weight to the co-occurrence frequency $f(xy)$ in the numerator by either raising it to some power k (Daille, 1994) or multiplying PMI with $f(xy)$. Table 4 lists these adjusted versions of PMI and their performance over our datasets. We can see from Table 4 that the best performance of PMI^k is obtained at k values less than one, indicating that it is better to rely less on $f(xy)$. Similarly, multiplying $f(xy)$ directly to PMI reduces the performance of PMI. As such, assigning more weight to $f(xy)$ does not improve the AP performance of PMI.

AM	VPCs	LVCs	Mixed
Best [M6] PMI^k	.547 (k = .13)	.573 (k = .85)	.544 (k = .32)
[M4] PMI	.510	.566	.515
[M5] Local-PMI	.259	.393	.272
[M1] Joint Prob.	.170	.28	.175

Table 4. AP performance of PMI and its variants. Best alpha settings shown in parentheses.

Another shortcoming of (P)MI is that both grow not only with the degree of dependence but also with frequency (Manning and Schütze, 1999, p. 66). In particular, we can show that $\mathrm{MI}(X; Y) \leq \min(H(X), H(Y))$, where $H(.)$ denotes entropy, and $\mathrm{PMI}(x,y) \leq \min(-\log P(x*), -\log P(*y))$.

These two inequalities suggest that the allowed score ranges of different candidates vary and consequently, MI and PMI scores are not directly comparable. Furthermore, in the case of VPCs and LVCs, the differences among score ranges of different candidates are large, due to high frequencies of particles and light verbs. This has motivated us to normalize these scores before using them for comparison. We suggest MI and PMI be divided by one of the following two normalization factors: $\mathrm{NF}(\alpha) = \alpha P(x*) + (1-\alpha)P(*y)$ with $\alpha \in [0, 1]$ and $\mathrm{NFmax} = \max(P(x*), P(*y))$. $\mathrm{NF}(\alpha)$, being dependent on alpha, can be optimized by setting an appropriate alpha value, which is inevitably affected by the MWE type and the corpus statistics. On the other hand, NFmax is independent of alpha and is recommended when one needs to apply normalized (P)MI to a mixed set of different MWE types or when sufficient data for parameter tuning is unavailable. As shown in Table 5, normalized MI and PMI show considerable improvements of up to 80%. Also, PMI and MI, after being normalized with NFmax, rank number one in VPC and LVC task, respectively. If one re-writes MI as = $(1/\mathrm{N}) \sum_{i,j} f_{ij} \times \mathrm{PMI}_{ij}$, it is easy to see the heavy dependence of MI on direct frequencies compared with PMI and this explains why normalization is a pressing need for MI.

AM	VPCs	LVCs	Mixed
MI / $\mathrm{NF}(\alpha)$	.508 (α = .48)	.583 (α = .47)	.516 (α = .5)
MI / NFmax	.508	.584	.518
[M2] MI	.273	.435	.289
PMI / $\mathrm{NF}(\alpha)$	.592 (α = .8)	.554 (α = .48)	.588 (α = .77)
PMI / NFmax	.565	.517	.556
[M4] PMI	.510	.566	.515

Table 5. AP performance of normalized (P)MI versus standard (P)MI. Best alpha settings shown in parentheses.

5.2 Penalization Terms

It can be seen that given equal co-occurrence frequencies, higher marginal frequencies reduce the likelihood of being MWEs. This motivates us to use marginal frequencies to synthesize *penalization terms* which are formulae whose values are inversely proportional to the likelihood of being MWEs. We hypothesize that incorporating such penalization terms can improve the respective AMs detection AP.

Take as an example, the AMs *[M21] Brawn-Blanquet* (a.k.a. *Minimum Sensitivity*) and *[M22] Simpson*. These two AMs are identical, except

for one difference in the denominator: *Brawn-Blanquet* uses max(*b, c*); *Simpson* uses min(*b, c*). It is intuitive and confirmed by our experiments that penalizing against the more frequent constituent by choosing max(*b, c*) is more effective. This is further attested in AMs *[M23] S Cost* and *[M25] Laplace*, where we tried to replace the min(*b, c*) term with max(*b, c*). Table 6 shows the average precision on our datasets for all these AMs.

AM	VPCs	LVCs	Mixed
[M21]Brawn-Blanquet	.478	.578	.486
[M22] Simpson	.249	.382	.260
[M24] Adjusted S Cost	.485	.577	.492
[M23] S cost	.249	.388	.260
[M26] Adjusted Laplace	.486	.577	.493
[M25] Laplace	.241	.388	.254

Table 6. Replacing *min*() with *max*() in selected AMs.

In the *[M27] Fager* AM, the penalization term max(*b, c*) is subtracted from the first term, which is no stranger but rank equivalent to [M7] PMI^2. In our application, this AM is not good since the second term is far larger than the first term, which is less than 1. As such, *Fager* is largely equivalent to just –½ max(*b, c*). In order to make use of the first term, we need to replace the constant ½ by a scaled down version of max(*b, c*). We have approximately derived $1/\sqrt{aN}$ as a lower bound estimate of max(*b, c*) using the independence assumption, producing *[M28] Adjusted Fager*. We can see from Table 7 that this adjustment improves Fager on both datasets.

AM	VPCs	LVCs	Mixed
[M28] Adjusted Fager	.564	.543	.554
[M27] Fager	.552	.439	.525

Table 7. Performance of *Fager* and its adjusted version.

The next experiment involves *[M14] Third Sokal Sneath*, which can be shown to be rank equivalent to *–b –c*. We further notice that frequencies *c* of particles are normally much larger than frequencies *b* of verbs. Thus, this AM runs the risk of ranking VPC candidates based on only frequencies of particles. So, it is necessary that we scale *b* and *c* properly as in [M14'] $-\alpha \times b - (1-\alpha) \times c$. Having scaled the constituents properly, we still see that [M14'] by itself is not a good measure as it uses only constituent frequencies and does not take into consideration the co-occurrence frequency of the two constituents. This has led us to experiment with [MR14''] $\frac{\text{PMI}}{\alpha \times b + (1-\alpha) \times c}$. The denominator of [MR14''] is obtained by removing the minus sign from [MR14'] so that it can be used as a penalization term. The choice of PMI in the numerator is due to the fact that the denominator of [MR14''] is in essence similar to $\text{NF}(\alpha) = \alpha P(x*) + (1-\alpha)P(*y)$, which has been successfully used to divide PMI in the normalized PMI experiment. We heuristically tried to simplify [MR14''] to the following AM [M30] $\frac{\log(ad)}{\alpha \times b + (1-\alpha) \times c}$. The setting of alpha in Table 8 below is taken from the best alpha setting obtained the experiment on the normalized PMI (Table 5). It can be observed from Table 8 that [MR14'''], being computationally simpler than normalized PMI, performs as well as normalized PMI and better than *Third Sokal-Sneath* over the VPC data set.

AM	VPCs	LVCs	Mixed
PMI / $\text{NF}(\alpha)$	.592 (α=.8)	.554 (α=.48)	.588 (α=.77)
[M30] $\frac{\log(ad)}{\alpha \times b + (1-\alpha) \times c}$	.600 (α=.8)	.484 (α=.48)	.588 (α=.77)
[M14] Third Sokal Sneath	.565	.453	.546

Table 8. AP performance of suggested VPCs' penalization terms and AMs.

With the same intention and method, we have found that while addition of marginal frequencies is a good penalization term for VPCs, the product of marginal frequencies is more suitable for LVCs (rows 1 and 2, Table 9). As with the linear combination, the product *bc* should also be weighted accordingly as $b^{\alpha}c^{(1-\alpha)}$. The best alpha value is also taken from the normalized PMI experiments (Table 5), which is nearly .5. Under this setting, this penalization term is exactly the denominator of the *[M18] Odds Ratio*. Table 9 below show our experiment results in deriving the penalization term for LVCs.

AM	VPCs	LVCs	Mixed
–b –c	.565	.453	.546
1/bc	.502	.532	.502
[M18] Odds ratio	.443	.567	.456

Table 9. AP performance of suggested LVCs' penalization terms and AMs.

6 Conclusions

We have conducted an analysis of the 82 AMs assembled in Pecina and Schlesinger (2006) for the tasks of English VPC and LVC extraction over the *Wall Street Journal* Penn Treebank data. In our work, we have observed that AMs can be divided into two classes: ones that do not use context (Class I) and ones that do (Class II), and find that the latter is not suitable for our VPC and LVC detection tasks as the size of our corpus is too small to rely on the frequency of candidates' contexts. This phenomenon also revealed the inappropriateness of hypothesis tests for our detection task. We have also introduced the novel notion of rank equivalence to MWE detection, in which we show that complex AMs may be replaced by simpler AMs that yield the same average precision performance.

We further observed that certain modifications to some AMs are necessary. First, in the context of ranking, we have proposed normalizing scores produced by MI and PMI in cases where the distributions of the two events are markedly different, as is the case for light verbs and particles. While our claims are limited to the datasets analyzed, they show clear improvements: normalized PMI produces better performance over our mixed MWE dataset, yielding an average precision of 58.8% compared to 51.5% when using standard PMI, a significant improvement as judged by paired T test. Normalized MI also yields the best performance over our LVC dataset with a significantly improved AP of 58.3%.

We also show that marginal frequencies can be used to form effective penalization terms. In particular, we find that $\alpha \times b + (1-\alpha) \times c$ is a good penalization term for VPCs, while $b^{\alpha} c^{(1-\alpha)}$ is suitable for LVCs. Our introduced alpha tuning parameter should be set to properly scale the values b and c, and should be optimized per MWE type. In cases where a common factor is applied to different MWE types, max(b, c) is a better choice than min(b, c). In future work, we plan to expand our investigations over larger, web-based datasets of English, to verify the performance gains of our modified AMs.

Acknowledgement

This work was partially supported by a National Research Foundation grant "Interactive Media Search" (grant # R 252 000 325 279).

References

Baldwin, Timothy (2005). The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Baldwin, Timothy and Villavicencio, Aline (2002). Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 98–104, Taipei, Taiwan.

Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques.* PhD thesis, Université Paris 7.

Evert, Stefan (2004). Online repository of association measures http://www.collocations.de/, a companion to *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. dissertation, University of Stuttgart.

Evert, Stefan and Krenn, Brigitte (2001) Methods for qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics,* pages 188-915, Toulouse, France.

Katz, Graham and Giesbrecht, Eugenie (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12-19, Sydney, Australia.

Krenn, Brigitte and Evert, Stefan (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, pages 39–46, Toulouse, France.

Manning D. Christopher and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Pearce, Darren (2002). A comparative evaluation of collocation extraction techniques. In *Proc. of the*

3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, pages 1530-1536, Canary Islands.

Pecina, Pavel and Schlesinger, Pavel (2006). Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 651-658, Sydney, Australia.

Quirk Randolph, Greenbaum Sidney, Leech Geoffrey and Svartvik Jan (1985). *A Comprehensive Grammar of the English Language*. Longman, London, UK.

Ramisch Carlos, Schreiner Paulo, Idiart Marco and Villavicencio Aline (2008). An Evaluation of Methods for the extraction of Multiword Expressions. In *Proceedings of the LREC-2008 Workshop on Multiword Expressions: Towards a Shared Task for Multiword Expressions*, pages 50-53, Marrakech, Morocco.

Smadja, Frank (1993). Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143–77.

Tan, Y. Fan, Kan M. Yen and Cui, Hang (2006). Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*, pages 49–56, Trento, Italy.

Zhai, Chengxiang (1997). Exploiting context to identify lexical atoms – A statistical view of linguistic context. In *International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97)*, pages 119-129, Rio de Janeiro, Brazil.

Appendix A. List of particles used in identifying verb particle constructions.

about, aback, aboard, above, abroad, across, adrift, ahead, along, apart, around, aside, astray, away, back, backward, backwards, behind, by, down, forth, forward, forwards, in, into, off, on, out, over, past, round, through, to, together, under, up, upon, without.