A Nonparametric Method for Extraction of Candidate Phrasal Terms

Paul Deane
Center for Assessment, Design and Scoring
Educational Testing Service
pdeane@ets.org

Abstract

This paper introduces a new method for identifying candidate phrasal terms (also known as multiword units) which applies a nonparametric, rank-based heuristic measure. Evaluation of this measure, the *mutual rank ratio* metric, shows that it produces better results than standard statistical measures when applied to this task.

1 Introduction

The ordinary vocabulary of a language like English contains thousands of *phrasal terms* -- multiword lexical units including compound nouns, technical terms, idioms, and fixed collocations. The exact number of phrasal terms is difficult to determine, as new ones are coined regularly, and it is sometimes difficult to determine whether a phrase is a fixed term or a regular, compositional expression. Accurate identification of phrasal terms is important in a variety of contexts, including natural language parsing, question answering systems, information retrieval systems, among others.

Insofar as phrasal terms function as lexical units, their component words tend to cooccur more often, to resist substitution or paraphrase, to follow fixed syntactic patterns, and to display some degree of semantic noncompositionality (Manning, 1999:183-186). However, none of these characteristics are amenable to a simple algorithmic interpretation. It is true that various term extraction systems have been developed, such as Xtract (Smadja 1993), Termight (Dagan & Church 1994), and TERMS (Justeson & Katz 1995) among others (cf. Daille 1996, Jacquemin & Tzoukermann 1994, Jacquemin, Klavans, & Toukermann 1997, Boguraev & Kennedy 1999, Lin 2001). Such systems typically rely on a combination of linguistic knowledge and statistical association measures. Grammatical patterns, such as adjective-noun or noun-noun sequences are selected then ranked statistically, and the resulting ranked list is either used directly or submitted for manual filtering.

The linguistic filters used in typical term extraction systems have no obvious connection with the criteria that linguists would argue define a phrasal term (noncompositionality, fixed order, nonsubstitutability, etc.). They function, instead, to reduce the number of a priori improbable terms and thus improve precision. The association measure does the actual work of distinguishing between terms and plausible nonterms. A variety of methods have been applied, ranging from simple frequency (Justeson & Katz 1995), modified frequency measures such as c-values (Frantzi, Anadiou & Mima 2000, Maynard & Anadiou 2000) and standard statistical significance tests such as the t-test, the chi-squared test, and log-likelihood (Church and Hanks 1990, Dunning 1993), and information-based methods, e.g. pointwise mutual information (Church & Hanks 1990).

Several studies of the performance of lexical association metrics suggest significant room for improvement, but also variability among tasks.

One series of studies (Krenn 1998, 2000; Evert & Krenn 2001, Krenn & Evert 2001; also see Evert 2004) focused on the use of association metrics to identify the best candidates in particular grammatical constructions, such as adjective-noun pairs or verb plus prepositional phrase constructions, and compared the performance of simple frequency to several common measures (the log-likelihood, the t-test, the chi-squared test, the dice coefficient, relative entropy and mutual information). In Krenn & Evert 2001, frequency outperformed mutual information though not the t-test, while in Evert and Krenn 2001, log-likelihood and the t-test gave the best results, and mutual information again performed worse than frequency. However, in all these studies performance was generally low, with precision falling rapidly after the very highest ranked phrases in the list.

By contrast, Schone and Jurafsky (2001) evaluate the identification of phrasal terms without grammatical filtering on a 6.7 million word extract from the TREC databases, applying both WordNet and online dictionaries as gold standards. Once again, the general level of performance was low, with precision falling off rapidly as larger portions

Proceedings of the 43rd Annual Meeting of the ACL, pages 605–613,
Ann Arbor, June 2005.

of the n-best list were included, but they report better performance with statistical and information theoretic measures (including mutual information) than with frequency. The overall pattern appears to be one where lexical association measures in general have very low precision and recall on unfiltered data, but perform far better when combined with other features which select linguistic patterns likely to function as phrasal terms.

The relatively low precision of lexical association measures on unfiltered data no doubt has multiple explanations, but a logical candidate is the failure or inappropriacy of underlying statistical assumptions. For instance, many of the tests assume a normal distribution, despite the highly skewed nature of natural language frequency distributions, though this is not the most important consideration except at very low *n* (cf. Moore 2004, Evert 2004, ch. 4). More importantly, statistical and information-based metrics such as the log-likelihood and mutual information measure significance or informativeness relative to the assumption that the selection of component terms is statistically independent. But of course the possibilities for combinations of words are anything but random and independent. Use of linguistic filters such as "attributive adjective followed by noun" or "verb plus modifying prepositional phrase" arguably has the effect of selecting a subset of the language for which the standard null hypothesis -- that any word may freely be combined with any other word -- may be much more accurate. Additionally, many of the association measures are defined only for bigrams, and do not generalize well to phrasal terms of varying length.

The purpose of this paper is to explore whether the identification of candidate phrasal terms can be improved by adopting a heuristic which seeks to take certain of these statistical issues into account. The method to be presented here, the *mutual rank ratio*, is a nonparametric rank-based approach which appears to perform significantly better than the standard association metrics.

The body of the paper is organized as follows: Section 2 will introduce the statistical considerations which provide a rationale for the mutual rank ratio heuristic and outline how it is calculated. Section 3 will present the data sources and evaluation methodologies applied in the rest of the paper. Section 4 will evaluate the mutual rank ratio statistic and several other lexical association measures on a larger corpus than has been used in previous evaluations. As will be shown below, the mutual rank ratio statistic recognizes phrasal terms more effectively than standard statistical measures.

2 Statistical considerations

2.1 Highly skewed distributions

As first observed e.g. by Zipf (1935, 1949) the frequency of words and other linguistic units tend to follow highly skewed distributions in which there are a large number of rare events. Zipf's formulation of this relationship for single word frequency distributions (Zipf's first law) postulates that the frequency of a word is inversely proportional to its rank in the frequency distribution, or more generally if we rank words by frequency and assign rank *z*, where the function $f_z(z,N)$ gives the frequency of rank z for a sample of size N, Zipf's first law states that:

$$f_z(z,N) = \frac{C}{z^{\alpha}}$$

where C is a normalizing constant and α is a free parameter that determines the exact degree of skew; typically with single word frequency data, α approximates 1 (Baayen 2001: 14). Ideally, an association metric would be designed to maximize its statistical validity with respect to the distribution which underlies natural language text -- which is if not a pure Zipfian distribution at least an LNRE (large number of rare events, cf. Baayen 2001) distribution with a very long tail, containing events which differ in probability by many orders of magnitude. Unfortunately, research on LNRE distributions focuses primarily on unigram distributions, and generalizations to bigram and n-gram distributions on large corpora are not as yet clearly feasible (Baayen 2001:221). Yet many of the best-performing lexical association measures, such as the t-test, assume normal distributions, (cf. Dunning 1993) or else (as with mutual information) eschew significance testing in favor of a generic information-theoretic approach. Various strategies could be adopted in this situation: finding a better model of the distribution,or adopting a nonparametric method.

2.2 The independence assumption

Even more importantly, many of the standard lexical association measures measure significance (or information content) against the default assumption that word-choices are statistically independent events. This assumption is built into the highest-performing measures as observed in Evert & Krenn 2001, Krenn & Evert 2001 and Schone & Jurafsky 2001.

This is of course untrue, and justifiable only as a simplifying idealization in the absence of a better model. The actual probability of any sequence of words is strongly influenced by the base grammatical and semantic structure of language, particularly since phrasal terms usually conform to

the normal rules of linguistic structure. What makes a compound noun, or a verb-particle construction, into a phrasal term is not deviation from the base grammatical pattern for noun-noun or verb-particle structures, but rather a further pattern (of meaning and usage and thus heightened frequency) superimposed on the normal linguistic base. There are, of course, entirely aberrant phrasal terms, but they constitute the exception rather than the rule.

This state of affairs poses something of a chicken-and-the-egg problem, in that statistical parsing models have to estimate probabilities from the same base data as the lexical association measures, so the usual heuristic solution as noted above is to impose a linguistic filter on the data, with the association measures being applied only to the subset thus selected. The result is in effect a constrained statistical model in which the independence assumption is much more accurate. For instance, if the universe of statistical possibilities is restricted to the set of sequences in which an adjective is followed by a noun, the null hypothesis that word choice is independent -- i.e., that any adjective may precede any noun -- is a reasonable idealization. Without filtering, the independence assumption yields the much less plausible null hypothesis that any word may appear in any order.

It is thus worth considering whether there are any ways to bring additional information to bear on the problem of recognizing phrasal terms without presupposing statistical independence.

2.3 Variable length; alternative/overlapping phrases

Phrasal terms vary in length. Typically they range from about two to six words in length, but critically we cannot judge whether a phrase is lexical without considering both shorter and longer sequences.

That is, the statistical comparison that needs to be made must apply in principle to the entire set of word sequences that must be distinguished from phrasal terms, including longer sequences, subsequences, and overlapping sequences, despite the fact that these are not statistically independent events. Of the association metrics mentioned thus far, only the C-Value method attempts to take direct notice of such word sequence information, and then only as a modification to the basic information provided by frequency.

Any solution to the problem of variable length must enable normalization allowing direct comparison of phrases of different length. Ideally, the solution would also address the other issues -- the independence assumption and the skewed distributions typical of natural language data.

2.4 Mutual expectation

An interesting proposal which seeks to overcome the variable-length issue is the *mutual expectation* metric presented in Dias, Guilloré, and Lopes (1999) and implemented in the SENTA system (Gil and Dias 2003a). In their approach, the frequency of a phrase is normalized by taking into account the relative probability of each word compared to the phrase.

Dias, Guilloré, and Lopes take as the foundation of their approach the idea that the cohesiveness of a text unit can be measured by measuring how strongly it resists the loss of any component term. This is implemented by considering, for any n-gram, the set of [continuous or discontinuous] (n-1)-grams which can be formed by deleting one word from the n-gram. A *normalized expectation* for the n-gram is then calculated as follows:

$$\frac{p([w_1, w_2 \ldots w_n])}{FPE([w_1, w_2 \ldots w_n])}$$

where $[w_1, w_2 \ldots w_n]$ is the phrase being evaluated and $FPE([w_1, w_2 \ldots w_n])$ is:

$$\frac{1}{n}\left(p([w_1, w_2 \ldots w_n]) + \sum_{i=1}^{n} p\left([w_1 \ldots \hat{w}_i \ldots . w_n] \right) \right)$$

where w_i is the term omitted from the n-gram.

They then calculate mutual expectation as the product of the probability of the n-gram and its normalized expectation.

This statistic is of interest for two reasons: first, it provides a single statistic that can be applied to n-grams of any length; second, it is not based upon the independence assumption. The core statistic, normalized expectation, is essentially frequency with a penalty if a phrase contains component parts significantly more frequent than the phrase itself.

It is of course an empirical question how well mutual expectation performs (and we shall examine this below) but mutual expectation is not in any sense a significance test. That is, if we are examining a phrase like the *east end*, the conditional probability of *east* given *[__ end]* or of *end* given *[__ east]* may be relatively low (since other words can appear in that context) and yet the phrase might still be very lexicalized if the association of both words with this context were significantly stronger than their association for

other phrases. That is, to the extent that phrasal terms follow the regular patterns of the language, a phrase might have a relatively low conditional probability (given the wide range of alternative phrases following the same basic linguistic patterns) and thus have a low mutual expectation yet still occur far more often than one would expect from chance.

In short, the fundamental insight -- assessing how tightly each word is bound to a phrase -- is worth adopting. There is, however, good reason to suspect that one could improve on this method by assessing relative statistical significance for each component word without making the independence assumption. In the heuristic to be outlined below, a nonparametric method is proposed. This method is novel: not a modification of mutual expectation, but a new technique based on ranks in a Zipfian frequency distribution.

2.5 Rank ratios and mutual rank ratios

This technique can be justified as follows. For each component word in the n-gram, we want to know whether the n-gram is more probable for that word than we would expect given its behavior with other words. Since we do not know what the expected shape of this distribution is going to be, a nonparametric method using ranks is in order, and there is some reason to think that frequency rank regardless of n-gram size will be useful. In particular, Ha, Sicilia-Garcia, Ming and Smith (2002) show that Zipf's law can be extended to the combined frequency distribution of n-grams of varying length up to rank 6, which entails that the relative rank of words in such a combined distribution provide a useful estimate of relative probability. The availability of new techniques for handling large sets of n-gram data (e.g. Gil & Dias 2003b) make this a relatively feasible task.

Thus, given a phrase like *east end*, we can rank how often __ *end* appears with *east* in comparison to how often other phrases appear with *east*.That is, if *{__ end, __side, the __, toward the __, etc.}* is the set of (variable length) n-gram contexts associated with *east* (up to a length cutoff), then the ***actual rank*** of __ *end* is the rank we calculate by ordering all contexts by the frequency with which the actual word appears in the context.

We also rank the set of contexts associated with *east* by their overall corpus frequency. The resulting ranking is the ***expected rank*** of __ *end* based upon how often the competing contexts appear regardless of which word fills the context.

The rank ratio (RR) for the word given the context can then be defined as:

$$\text{RR(word,context)} = \frac{ER(word, context)}{AR(word, context)}$$

where ER is the expected rank and AR is the actual rank. A normalized, or mutual rank ratio for the n-gram can then be defined as

$$\sqrt[n]{RR(w_{1,[\_\_\ w2....wn]}) * RR(w_{2,[w1\ \_\_\ ...wn]}) ... * RR(w_{n,[w1,\ w2...\ \_]})}$$

The motivation for this method is that it attempts to address each of the major issues outlined above by providing a nonparametric metric which does not make the independence assumption and allows scores to be compared across n-grams of different lengths.

A few notes about the details of the method are in order. Actual ranks are assigned by listing all the contexts associated with each word in the corpus, and then ranking contexts by word, assigning the most frequent context for word n the rank 1, next next most frequent rank 2, etc. Tied ranks are given the median value for the ranks occupied by the tie, e.g., if two contexts with the same frequency would occupy ranks 2 and 3, they are both assigned rank 2.5. Expected ranks are calculated for the same set of contexts using the same algorithm, but substituting the unconditional frequency of the (n-1)-gram for the gram's frequency with the target word.[1]

3 Data sources and methodology

The Lexile Corpus is a collection of documents covering a wide range of reading materials such as a child might encounter at school, more or less evenly divided by Lexile (reading level) rating to cover all levels of textual complexity from kindergarten to college. It contains in excess of 400 million words of running text, and has been made available to the Educational Testing Service under a research license by Metametrics Corporation.

This corpus was tokenized using an in-house tokenization program, *toksent*, which treats most punctuation marks as separate tokens but makes single tokens out of common abbreviations, numbers like *1,500*, and words like *o'clock*. It should be noted that some of the association measures are known to perform poorly if punctuation marks and common stopwords are

[1] In this study the rank-ratio method was tested for bigrams and trigrams only, due to the small number of WordNet gold standard items greater than two words in length. Work in progress will assess the metrics' performance on n-grams of orders four through six.

included; therefore, n-gram sequences containing punctuation marks and the 160 most frequent word forms were excluded from the analysis so as not to bias the results against them. Separate lists of bigrams and trigrams were extracted and ranked according to several standard word association metrics. Rank ratios were calculated from a comparison set consisting of all contexts derived by this method from bigrams and trigrams, e.g., contexts of the *form word1__, ___word2, ___word1 word2, word1 ___ word3*, and *word1 word2 ___*.[2]

Table 1 lists the standard lexical association measures tested in section four[3].

The logical evaluation method for phrasal term identification is to rank n-grams using each metric and then compare the results against a gold standard containing known phrasal terms. Since Schone and Jurafsky (2001) demonstrated similar results whether WordNet or online dictionaries were used as a gold standard, WordNet was selected. Two separate lists were derived containing two- and three-word phrases. The choice of WordNet as a gold standard tests ability to predict general dictionary headwords rather than technical terms, appropriate since the source corpus consists of nontechnical text.

Following Schone & Jurafsky (2001), the bigram and trigram lists were ranked by each statistic then scored against the gold standard, with results evaluated using a figure of merit (FOM) roughly characterizable as the area under the precision-recall curve. The formula is:

$$\frac{1}{K}\sum_{i=1}^{k} P_i$$

where P_i (precision at i) equals i/H_i, and H_i is the number of n-grams into the ranked n-gram list required to find the i^{th} correct phrasal term.

It should be noted, however, that one of the most pressing issues with respect to phrasal terms is that they display the same skewed, long-tail distribution as ordinary words, with a large proportion of the total displaying very low frequencies. This can be measured by considering

METRIC	FORMULA
Frequency (Guiliano, 1964)	f_{xy}
Pointwise Mutual Information [PMI] (Church & Hanks, 1990)	$\log_2\left(P_{xy} / P_x P_y\right)$
True Mutual Information [TMI] (Manning, 1999)	$P_{xy}\log_2\left(P_{xy} / P_x P_y\right)$
Chi-Squared (χ^2) (Church and Gale, 1991)	$\sum_{\substack{i\in\{x,\bar{x}\}\\ j\in\{y,\bar{y}\}}} \frac{(f_{ij}-\zeta_{ij})^2}{\zeta_{ij}}$
T-Score (Church & Hanks, 1990)	$\frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
C-Values[4] (Frantzi, Anadiou & Mima 2000)	$\begin{cases} \log_2\lvert\alpha\rvert\cdot f(\alpha)_a \text{ is not nested} \\ \log_2\lvert\alpha\rvert\cdot f(\alpha) - \frac{1}{P(T_a)}\sum_{b\in T_a} f(b) \end{cases}$ where α is the candidate string f(α) is its frequency in the corpus Tα is the set of candidate terms that contain α P(Tα) is the number of these candidate terms

Table 1. Some Lexical Association Measures

the overlap between WordNet and the Lexile corpus. A list of 53,764 two-word phrases were extracted from WordNet, and 7,613 three-word phrases. Even though the Lexile corpus is quite large -- in excess of 400 million words of running text -- only 19,939 of the two-word phrases and

[2] Excluding the 160 most frequent words prevented evaluation of a subset of phrasal terms such as verbal idioms like *act up* or *go on*. Experiments with smaller corpora during preliminary work indicated that this exclusion did not appear to bias the results.

[3] Schone & Jurafsky's results indicate similar results for log-likelihood & T-score, and strong parallelism among information-theoretic measures such as Chi-Squared, Selectional Association (Resnik 1996), Symmetric Conditional Probability (Ferreira and Pereira Lopes, 1999) and the Z-Score (Smadja 1993). Thus it was not judged necessary to replicate results for all methods covered in Schone & Jurafsky (2001).

[4] Due to the computational cost of calculating C-Values over a very large corpus, C-Values were calculated over bigrams and trigrams only. More sophisticated versions of the C-Value method such as NC-values were not included as these incorporate linguistic knowledge and thus fall outside the scope of the study.

1,700 of the three-word phrases are attested in the Lexile corpus. 14,045 of the 19,939 attested two-word phrases occur at least 5 times, 11,384 occur at least 10 times, and only 5,366 occur at least 50 times; in short, the strategy of cutting off the data at a threshold sacrifices a large percent of total recall. Thus one of the issues that needs to be addressed is the accuracy with which lexical association measures can be extended to deal with relatively sparse data, e.g., phrases that appear less than ten times in the source corpus.

A second question of interest is the effect of filtering for particular linguistic patterns. This is another method of prescreening the source data which can improve precision but damage recall. In the evaluation bigrams were classified as N-N and A-N sequences using a dictionary template, with the expected effect. For instance, if the WordNet two word phrase list is limited only to those which could be interpreted as noun-noun or adjective noun sequences, N>=5, the total set of WordNet terms that can be retrieved is reduced to 9,757..

4 Evaluation

Schone and Jurafsky's (2001) study examined the performance of various association metrics on a corpus of 6.7 million words with a cutoff of N=10. The resulting n-gram set had a maximum recall of 2,610 phrasal terms from the WordNet gold standard, and found the best figure of merit for any of the association metrics even with linguistic filterering to be 0.265. On the significantly larger Lexile corpus N must be set higher (around N=50) to make the results comparable. The statistics were also calculated for N=50, N=10 and N=5 in order to see what the effect of including more (relatively rare) n-grams would be on the overall performance for each statistic. Since many of the statistics are defined without interpolation only for bigrams, and the number of WordNet trigrams at N=50 is very small, the full set of scores were only calculated on the bigram data. For trigrams, in addition to rank ratio and frequency scores, extended pointwise mutual information and true mutual information scores were calculated using the formulas log $(P_{xyz}/P_xP_y\ P_z)$) and P_{xyz} log $(P_{xyz}/P_xP_y\ P_z)$). Also, since the standard lexical association metrics cannot be calculated across different n-gram types, results for bigrams and trigrams are presented separately for purposes of comparison.

The results are are shown in Tables 2-5. Two points should should be noted in particular. First, the rank ratio statistic outperformed the other association measures tested across the board. Its best performance, a score of 0.323 in the part of speech filtered condition with N=50, outdistanced

METRIC	POS Filtered	Unfiltered
RankRatio	0.323	0.196
Mutual Expectancy	0.144	0.069
TMI	0.209	0.096
PMI	0.287	0.166
Chi-sqr	0.285	0.152
T-Score	0.154	0.046
C-Values	0.065	0.048
Frequency	0.130	0.044

Table 2. Bigram Scores for Lexical Association Measures with N=50

METRIC	POS Filtered	Unfiltered
RankRatio	0.218	0.125
MutualExpectation	0.140	0.071
TMI	0.150	0.070
PMI	0.147	0.065
Chi-sqr	0.145	0.065
T-Score	0.112	0.048
C-Values	0.096	0.036
Frequency	0.093	0.034

Table 3. Bigram Scores for Lexical Association Measures with N=10

METRIC	POS Filtered	Unfiltered
RankRatio	0.188	0.110
Mutual Expectancy	0.141	0.073
TMI	0.131	0.063
PMI	0.108	0.047
Chi-sqr	0.107	0.047
T-Score	0.098	0.043
C-Values	0.084	0.031
Frequency	0.081	0.021

Table 4. Bigram Scores for Lexical Association Measures with N=5

METRIC	N=50	N=10	N=5
RankRatio	0.273	0.137	0.103
PMI	0.219	0.121	0.059
TMI	0.137	0.074	0.056
Frequency	0.089	0.047	0.035

Table 5. Trigram scores for Lexical Association Measures at N=50, 10 and 5 without linguistic filtering.

the best score in Schone & Jurafsky's study (0.265), and when large numbers of rare bigrams were included, at N=10 and N=5, it continued to outperform the other measures. Second, the results were generally consistent with those reported in the literature, and confirmed Schone & Jurafsky's observation that the information-theoretic measures (such as mutual information and chi-squared) outperform frequency-based measures (such as the T-score and raw frequency.)[5]

4.1 Discussion

One of the potential strengths of this method is that is allows for a comparison between n-grams of varying lengths. The distribution of scores for the gold standard bigrams and trigrams appears to bear out the hypothesis that the numbers are comparable across n-gram length. Trigrams constitute approximately four percent of the gold standard test set, and appear in roughly the same percentage across the rankings; for instance, they consistute 3.8% of the top 10,000 ngrams ranked by mutual rank ratio. Comparison of trigrams with their component bigrams also seems consistent with this hypothesis; e.g., the bigram *Booker T.* has a higher mutual rank ratio than the trigram *Booker T. Washington*, which has a higher rank that the bigram *T. Washington*. These results suggest that it would be worthwhile to examine how well the method succeeds at ranking n-grams of varying lengths, though the limitations of the current evaluation set to bigrams and trigrams prevented a full evaluation of its effectiveness across n-grams of varying length.

The results of this study appear to support the conclusion that the Mutual Rank Ratio performs notably better than other association measures on this task. The performance is superior to the next-best measure when N is set as low as 5 (0.110 compared to 0.073 for Mutual Expectation and 0.063 for true mutual information and less than .05 for all other metrics). While this score is still fairly low, it indicates that the measure performs relatively well even when large numbers of low-probability n-grams are included. An examination of the n-best list for the Mutual Rank ratio at N=5 supports this contention.

The top 10 bigrams are:

Julius Caesar, Winston Churchill, potato chips, peanut butter, Frederick Douglass, Ronald Reagan, Tia Dolores, Don Quixote, cash register, Santa Claus

At ranks 3,000 to 3,010, the bigrams are:

Ted Williams, surgical technicians, Buffalo Bill, drug dealer, Lise Meitner, Butch Cassidy, Sandra Cisneros, Trey Granger, senior prom, Ruta Skadi

At ranks 10,000 to 10,010, the bigrams are:

egg beater, sperm cells, lowercase letters, methane gas, white settlers, training program, instantly recognizable, dried beef, television screens, vienna sausages

In short, the n-best list returned by the mutual rank ratio statistic appears to consist primarily of phrasal terms far down the list, even when N is as low as 5. False positives are typically: (i) morphological variants of established phrases; (ii) bigrams that are part of longer phrases, such as *cream sundae* (from *ice cream sundae*); (iii) examples of highly productive constructions such as *an artist, three categories* or *January 2.*

The results for trigrams are relatively sparse and thus less conclusive, but are consistent with the bigram results: the mutual rank ratio measure performs best, with top ranking elements consistently being phrasal terms.

Comparison with the n-best list for other metrics bears out the qualitative impression that the rank ratio is performing better at selecting phrasal terms even without filtering. The top ten bigrams for the true mutual information metric at N=5 are:

a little, did not, this is, united states, new york, know what, a good, a long, a moment, a small

Ranks 3000 to 3010 are:

waste time, heavily on, earlier than, daddy said, ethnic groups, tropical rain, felt sure, raw materials, gold medals, gold rush

Ranks 10,000 to 10,010 are:

quite close, upstairs window, object is, lord god, private schools, nat turner, fire going, bering sea,little higher, got lots

The behavior is consistent with known weaknesses of true mutual information -- its tendency to overvalue frequent forms.

Next, consider the n-best lists for log-likelihood at N=5. The top ten n-grams are:

sheriff poulson, simon huggett, robin redbreast, eric torrosian, colonel hillandale, colonel sapp, nurse leatheran, st. catherines, karen torrio, jenny yonge

N-grams 3000 to 3010 are:

comes then, stuff who, dinner get, captain see, tom see, couple get, fish see, picture go, building go, makes will, pointed way

[5] Schone and Jurafsky's results differ from Krenn & Evert (2001)'s results, which indicated that frequency performed better than the statistical measures in almost every case. However, Krenn and Evert's data consisted of n-grams preselected to fit particular collocational patterns. Frequency-based metrics seem to be particularly benefited by linguistic prefiltering.

N-grams 10000 to 10010 are:

sayings is, writ this, llama on, undoing this, dwahro did, reno on, squirted on, hardens like, mora did, millicent is, vets did

Comparison thus seems to suggest that if anything the quality of the mutual rank ratio results are being understated by the evaluation metric, as the metric is returning a large number of phrasal terms in the higher portion of the n-best list that are absent from the gold standard.

Conclusion

This study has proposed a new method for measuring strength of lexical association for candidate phrasal terms based upon the use of Zipfian ranks over a frequency distribution combining n-grams of varying length. The method is related in general philosophy of Mutual Expectation, in that it assesses the strenght of connection for each word to the combined phrase; it differs by adopting a nonparametric measure of strength of association. Evaluation indicates that this method may outperform standard lexical association measures, including mutual information, chi-squared, log-likelihood, and the T-score.

References

Baayen, R. H. (2001) *Word Frequency Distributions*. Kluwer: Dordrecht.

Boguraev, B. and C. Kennedy (1999). Applications of Term Identification Technology: Domain Description and Content Characterization. *Natural Language Engineering* 5(1):17-44.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocation expressions in large textual databases. *Proceedings of the RIAO*, pages 38-43.

Church, K.W., and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics 16(1):*22-29.

Dagan, I. and K.W. Church (1994). Termight: Identifying and translating technical terminology**.** *ACM International Conference Proceeding Series: Proceedings of the fourth conference on Applied natural language processing***,** pages 39-40.

Daille, B. 1996. "Study and Implementation of Combined Techniques from Automatic Extraction of Terminology". Chap. 3 of "The Balancing Act": *Combining Symbolic and Statistical Approaches to Kanguage* (Klavans, J., Resnik, P. (eds.)), pages 49-66.

Dias, G., S. Guilloré, and J.G. Pereira Lopes (1999), Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *TALN*, p. 333-338.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19(1):* 65-74.

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Phd Thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, S. and B. Krenn. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 188-195.

Ferreira da Silva, J. and G. Pereira Lopes (1999). A local maxima method and a fair dispersion normalization for extracting multiword units from corpora. *Sixth Meeting on Mathematics of Language*, pages 369-381.

Frantzi, K., S. Ananiadou, and H. Mima. (2000). Automatic recognition of multiword terms: the C-Value and NC-Value Method. *International Journal on Digital Libraries* 3(2):115-130.

Gil, A. and G. Dias. (2003a). Efficient Mining of Textual Associations. *International Conference on Natural Language Processing and Knowledge Engineering.* Chengqing Zong (eds.) pages 26-29.

Gil, A. and G. Dias (2003b). Using masks, suffix array-based data structures, and multidimensional arrays to compute positional n-gram statistics from corpora. In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics, pages 25-33.

Ha, L.Q., E.I. Sicilia-Garcia, J. Ming and F.J. Smith. (2002), "Extension of Zipf's law to words and phrases", *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2002)*, pages 315-320.

Jacquemin, C. and E. Tzoukermann. (1999). NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax. *Natural Language Processing Information Retrieval*, pages 25-74. Kuwer, Boston, MA, U.S.A.

Jacquemin, C., J.L. Klavans and E. Tzoukermann (1997). Expansion of multiword terms for indexing and retrieval using morphology and syntax. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pages 24-31.

Johansson, C. 1994b, Catching the Cheshire Cat, In *Proceedings of COLING 94*, Vol. II, pages 1021 - 1025.

Johansson, C. 1996. Good Bigrams. In Proceedings from the 16th International Conference on Computational Linguistics (COLING-96), pages 592-597.

Justeson, J.S. and S.M. Katz (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1:9-27.

Krenn, B. 1998. Acquisition of Phraseological Units from Linguistically Interpreted Corpora. A Case Study on German PP-Verb Collocations. Proceedings of ISP-98, pages 359-371.

Krenn, B. 2000. Empirical Implications on Lexical Association Measures. Proceedings of The Ninth EURALEX International Congress.

Krenn, B. and S. Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. Proceedings of the ACL Workshop on Collocations, pages 39-46.

Lin, D. 1998. Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology*, pages 57-63

Lin, D. 1999. Automatic Identification of Non-compositional Phrases, In *Proceedings of The 37th Annual Meeting of the Association For Computational Lingusitics*, pages 317-324.

Manning, C.D. and H. Schütze. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, U.S.A.

Maynard, D. and S. Ananiadou. (2000). Identifying Terms by their Family and Friends. COLING 2000, pages 530-536.

Pantel, P. and D. Lin. (2001). A Statistical Corpus-Based Term Extractor. In: Stroulia, E. and Matwin, S. (Eds.) *AI 2001, Lecture Notes in Artificial Intelligence*, pages 36-46. Springer-Verlag.

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition 61*: 127-159.

Schone, P. and D. Jurafsky, 2001. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? Proceedings of *Empirical Methods in Natural Language Processing*, pages 100-108.

Sekine, S., J. J. Carroll, S. Ananiadou, and J. Tsujii. 1992. *Automatic Learning for Semantic Collocation.* Proceedings of the 3rd Conference on Applied Natural Language Processing, pages 104-110.

Shimohata, S., T. Sugio, and J. Nagata. (1997). Retrieving collocations by co-occurrences and word order constraints. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 476-481.

Smadja, F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics, 19:143-177.

Thanapoulos, A., N. Fakotakis and G. Kokkinkais. 2002. Comparaitve Evaluation of Collocation Extraction Metrics. *Proceedings of the LREC 2002 Conference*, pages 609-613.

Zipf, P. (1935). *Psychobiology of Language.* Houghton-Mifflin, New York, New York.

Zipf, P.(1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Cambridge, Mass.