Term Dependence: A Basis for Luhn and Zipf Models

Journal of the American Society for Information Science and Technology, 52 (12), 1019-1025, 2001

Robert M. Losee *
Manning Hall, CB#3360
U. of North Carolina
Chapel Hill, NC 27599-3360
losee@ils.unc.edu

September 25, 2001

Abstract

There are regularities in the statistical information provided by natural language terms about neighboring terms. We find that when phrase rank increases, moving from common to less common phrases, the value of the expected mutual information measure (EMIM) between the terms regularly decreases. Luhn's model suggests that mid-range terms are the best index terms and relevance discriminators. We suggest reasons for this principle based on the empirical relationships shown here between the rank of terms *within* phrases and the average mutual information between terms, which we refer to as the *Inverse Representation–EMIM* principle. We also suggest an

*The authors wishes to acknowledge the helpful suggestions on earlier drafts of this manuscript by Miles Efron and an anonymous referee.

Inverse EMIM term weight for indexing or retrieval applications that is consistent with Luhn's distribution. An information theoretic interpretation of Zipf's Law, a power law, is provided. Using the regularity noted above, we suggest that Zipf's Law is a consequence of the statistical dependencies that exist between terms, described here using information theoretic concepts.

1 Introduction

When we communicate with one another, the words that we use occur in groups: phrases, sentences, or larger socio-linguistic structures, such as conversations or speeches. Are there regularities found in term co-occurrence relationships? Can these regularities help explain the existing principles found in linguistics (Yngve, 1986) and information retrieval (Losee, 1998), such as Zipf's Law (Zipf, 1949), which relates term frequencies and term ranks, or Luhn's model, which predicts the relative "resolving power of significant words" (Luhn, 1958), given the relative term frequency?

We suggest here relationships between the frequency-based characteristics of neighboring terms in natural language and the rank or frequency of the terms. Given the term rank or frequency, we can infer the entropy, or average information, of a term or a group of terms. The amount of information that one term has about another depends on the rank of one of the terms and of the rank or frequency of the term pair. Using these relationships, we can offer a partial explanation of why Luhn's distribution and Zipf's Law occur as they do (Simon, 1955; Mandelbrot, 1961; Rapoport, 1982; Naranan & Balasubrahmanyan, 1998; Egghe, 1999). We propose descriptions of the relationships between terms that hold across all languages.

The frequency of a term may be taken as the raw term frequency or as the probability of the term occurring. The rank of a term is the position of the term in a list of n term types ordered by their probability of occurrence, with the most frequent term having rank 1. The rarest term will have rank n. When describing the ranks of terms, we will treat a rank of x as *greater than* a rank of $x - 1$. For the data analysis below, terms with equal frequencies will be assigned the mean rank for the set of equi-frequent terms.

Note that Zipf's Law takes a number of forms, including one suggesting that the probability of a term multiplied by the term's rank equals a constant of about 0.1. This value is relatively stable over a wide range of the total number of terms. The general argument below is independent of the exact form of Zipf's Law that is adopted.

2 Information Theoretic Measures

Information theory is based on a range of measures that have developed beginning with the work in the preceding century of Nyquist, Hartley, and Shannon (Nyquist, 1924; Hartley, 1928; Shannon & Weaver, 1949; Aczél & Daróczy, 1975; Losee, 1990; Cover & Thomas, 1991). The information in a single event is referred to as self-information. Measured as the amount of information provided by knowledge of the probability of event x, $I(x) = -\log \Pr(x)$, self-information serves as the basis for more elaborate concepts. We denote the probability of event x as $\Pr(x)$. The average self-information, or entropy, is computed as

$$H(X) = -\sum_{i=1}^{n} \Pr(x_i) \log \Pr(x_i). \tag{1}$$

A single random variable or event may be studied with a second variable or event to produce a *joint distribution*, the probability of the two variables occurring in a certain way. If I flip two coins, C_1 and C_2, with heads denoted as H and tails as T, we may have one of four results: HH, HT, TH, or TT. If we make the reasonable assumption that the two coins C_1 and C_2 are statistically independent of each other with regards to how they land, we may compute the probability of a certain result $\Pr(C_1, C_2)$ as the product of the probabilities for each of the individual coins: $\Pr(C_1, C_2) = \Pr(C_1) \Pr(C_2)$.

In many circumstances, joint probabilities aren't derived from statistically independent events. Knowledge about dependent events, where one variable or event provides information about the chances of the other event occurring, can be very useful in many forms of analysis. For example, the probability that one point on earth is under water is highly dependent on the probability that a point one foot to the north of it is under water. While these are separate locations, the probability that one location is underwater and the probability that the other location is underwater are clearly not independent. In this situation, we cannot use the product rule above to compute the joint probability.

The expected mutual information measure (EMIM), which computes the amount of information that random variables Y and X provide about each other, is measured as

$$I(X;Y) = \sum_{i,j} \Pr(x_i, y_j) \log \frac{\Pr(x_i, y_j)}{\Pr(x_i) \Pr(y_j)}. \tag{2}$$

In addition, this may be computed as (Cover & Thomas, 1991):

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y). \qquad (3)$$

These random variables may be term frequencies, such as whether a term occurs or not in a given document, or how often terms occur. We may describe the term relations as the mutual information or as the statistical dependence relating one term to another (Chow & Liu, 1968; Yu, Buckley, Lam, & Salton, 1983; Losee, 1994; Croft, 1986; Gey, 1993; Cox & Wermuth, 1993). Below, we will consider methods to estimate the EMIM between terms, given other factors.

3 Data and Methods

We use a standard set of texts for our analysis of the term occurrences and relationships found in natural language. We compute statistics largely from the fulltext of 392 documents in the CF database (Moon, 1993). This set of documents represents the fulltext of every other document containing an abstract and the subject heading "Cystic Fibrosis" included in the National Library of Medicine's database between 1974 and 1979.

The exact choice of text database has relatively little impact on the kinds of results obtained for this type of study. Earlier work by Smith and Devine (1985) using several different datasets shows that relationships between term frequency, rank, and phrase size are relatively robust across a range of document types. To informally see this similarity, the reader might wish to briefly glance ahead at Figures 5 and 6, which compute the same data for the CF medical documents and for English literature. A detailed interpretation of this work is described later in the article. However, the reader might note the visual similarities in the results for these two very different kinds of literature.

The literature describing Zipf's Law suggests that there is a strong relationship between rank and frequency. One way of examining the character of this database vis á vis rank and frequency studies is to look at the relationship between rank and frequency. We find that the correlation between the logarithm of the rank and the logarithm of the frequency in the CF database is $-.995$. This suggests that there is a strong relationship between rank and frequency and that the data tends to cluster tightly around the line suggested by Zipf's Law.

All graphs show the average y value for the given x value (rank or frequency.) On the other hand, correlations in this work are based on each individual value and not on the averages.

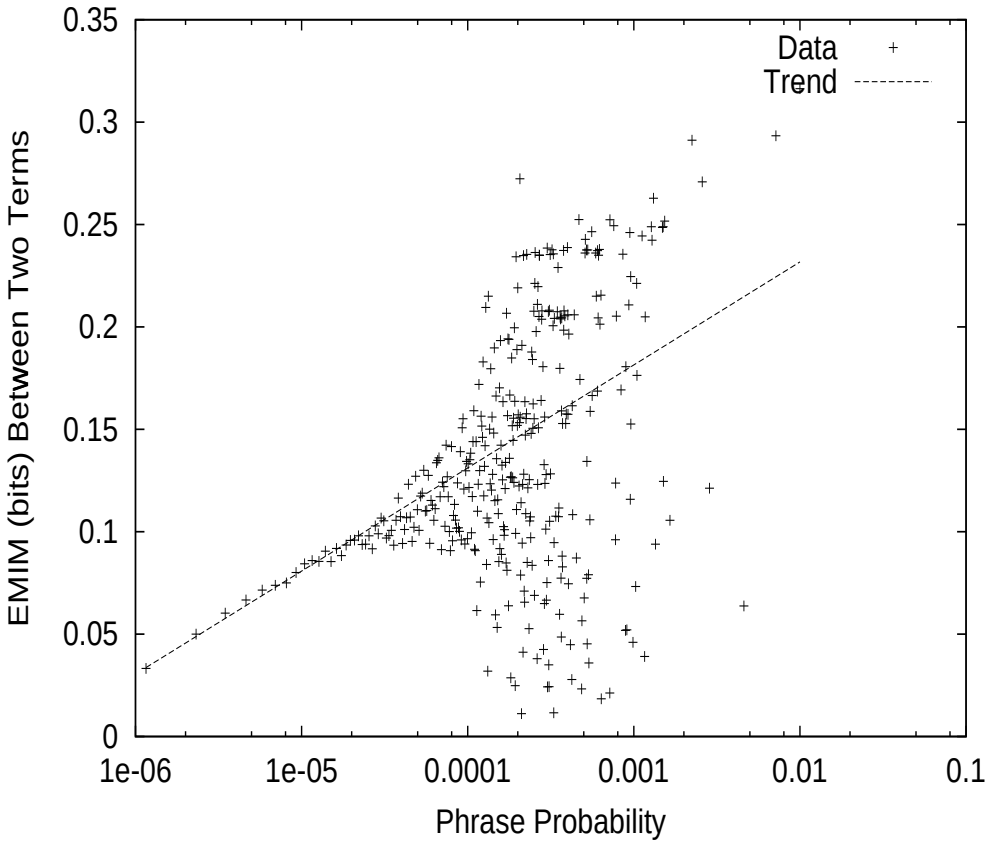

Figure 1: The relationship between the probability of a two word phrase occurring and the EMIM between the two terms.

4 EMIM as a Function of Phrase Frequency and Rank

Information theory may be applied to the study of phrase occurrences in order to help us describe and understand existing relationships between terms. The relationships between the terms in a term-pair may be examined by looking at the information that terms in the pairs provide about each other. By considering the empirical relationships found between the EMIM and the rank and the frequency of either term-pairs or of the individual terms in the pairs, we can discern trends that are both interesting by themselves and that enable us to explain other phenomena, such as the Zipf or Luhn models.

Figure 1 shows the relationship between the EMIM of terms in term-pairs and the frequency of the term pairs in the CF database. Clearly, as we move from

rare term-pairs to more common term-pairs, on the average, the EMIM increases. A second observation that can be made is that the lowest frequency term-pairs closely follow the trend line in Figure 1 and then begin to diverge from the trend line when the term-pairs become as frequent as about one in ten thousand. This frequency roughly corresponds to rank 1000. The diversity increases for more common terms, those that are less likely to be topic-carrying terms.

The increasing variation with higher probability terms may be due in part to the aggregation that has taken place in the dataset. Common terms often have their own unique frequencies, so that some points representing more common terms represent only a single term type. Those terms that are rare occur with low frequencies such as one or two or three occurrences. There are many terms with each of these low probabilities of occurrence. The aggregation of those terms with a given rare probability may dampen fluctuations that might be found with more common term types that occur with much lower numbers of term types. This may explain, in part, the greater variation seen in the center and right parts of Figure 1

Using all individual values in the data set, the correlation between the log of the term frequency and the EMIM is $.28$.

Similarly, Figure 2 shows the same kind of relationship that exists between term-pair rank and the EMIM. While there is a moderate amount of variation in the EMIM for highly ranked terms, terms ranked greater than one thousand follow the overall trend rather well, with the variation decreasing as the term-pair rank increases. The correlation between EMIM and the log rank of phrases is $-.28$, while the correlation between EMIM and the term rank is $-.25$.

We can conclude that there is a relationship between term frequency or rank and the information between terms. If we limit ourselves to the less common terms, those above rank 1000, the relationship is relatively strong and seems to capture a fundamental underlying regularity.

5 EMIM as a Function of Frequency and Rank of Individual Terms in a Pair

The occurrences of terms in natural language are often dependent on the occurrences of the terms immediately around them. This dependence may be measured with the EMIM, computed from knowledge of the frequencies of the two terms in a term-pair. Data in Figures 1 and 2 show that, when examining pairs of sequential terms from documents in the CF database, the EMIM generally increases as the phrase frequency increases. However, Figure 3 shows that when using the term frequency from the first term in a pair, the EMIM decreases as a term moves

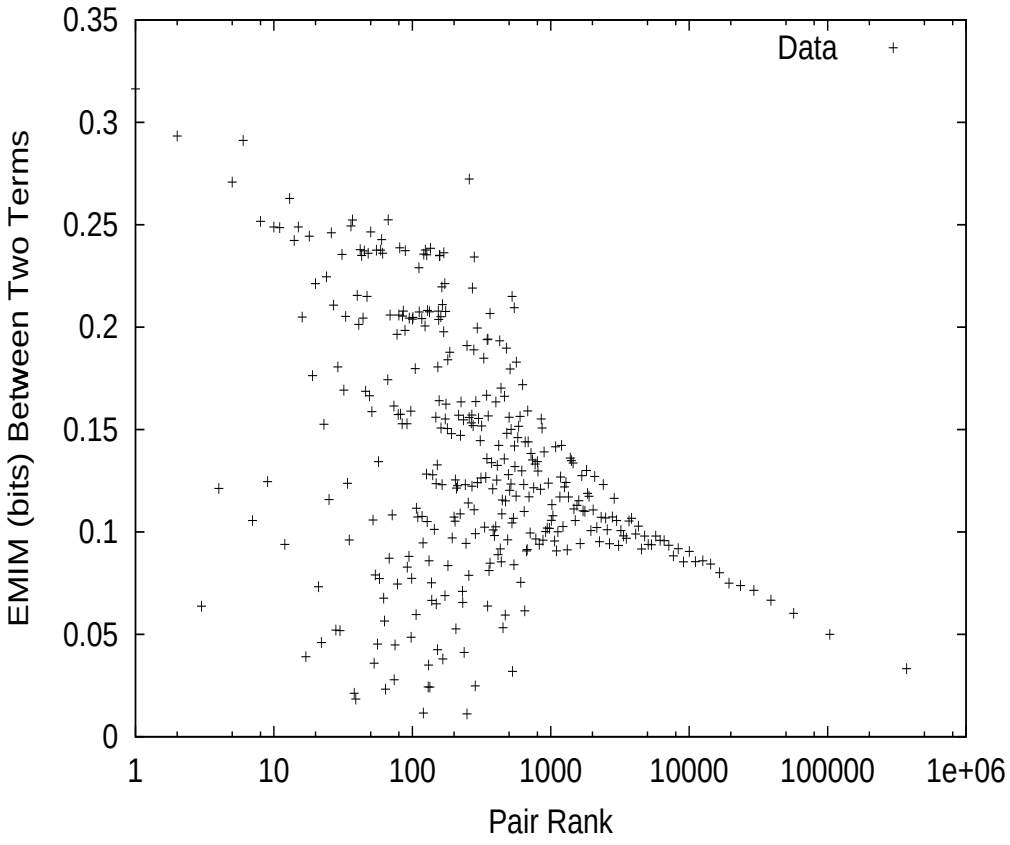

Figure 2: Two term phrase rank and the EMIM between the terms.

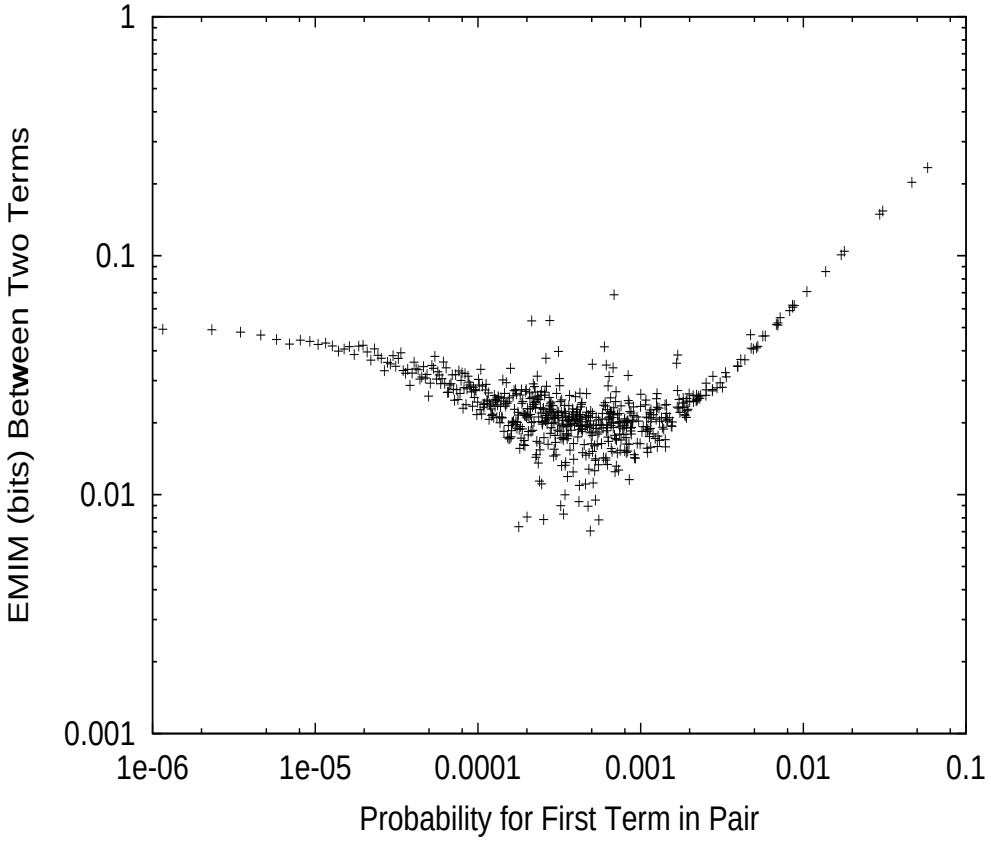

Figure 3: The relationship between the probability of the first term in a term pair and the EMIM between the two terms. The decrease in information between terms at mid-range frequencies results in decreased term dependence and an increased power of discrimination for each term in the mid-range.

toward the mid-frequency range, and then increases again. An almost identical plot is obtained when using the second term in the pair in place of the first term.

Informally, the EMIM might be expected to increase with individual term rank (and with lower frequency) because more-common terms might be expected to carry less subject-based information about the terms or their neighbors. The higher frequency terms might have less self-information and thus are informally less informative or less useful. This is similar to the assumption underlying the Inverse Document Frequency (IDF) term weighting that is often used in information retrieval applications (Sparck Jones, 1972; Losee, 1998).

However, we find that the EMIM between terms in a term-pair is minimized for mid-frequency terms, as is shown at the center of Figure 3. Phrases with more commonly occurring first terms carry more EMIM. Common phrases contain terms that co-occur more frequently and thus are more permanent constructs; the phrases may be treated as units of meaning rather than as two separate meaning units. When phrases occur with unusual first terms, they are less likely to be used as a unit, but are rather a temporary union developed for the purpose of conveying a meaning or idea.

In summary, three information-related phenomena are found in terms: (1) The IDF weighting system suggests that the less frequent is the occurrence of a term, the better the term will serve as an index term or as a representative of what the document is about; (2) Figure 1 shows that, in general, as phrase frequency decreases, the EMIM between terms in the phrases decreases; (3) Figure 3 suggests that as the first term in a pair of terms varies in frequency, the lowest EMIM between terms in the pair is found in the mid-frequency range. A similar set of relationships is observed when using the second term in the two-term phrase rather than the first term.

6 Entropy and Rank

The entropy of a term-pair is a function of the pair's frequency. Given the strong relationship between term frequency and rank described by Zipf's Law and the relationships between rank and frequency seen when comparing Figures 1 and 2, we may expect there to be a relationship between phrase and term rank, frequency, and entropy.

The data shown in Figure 4 is similar to that found elsewhere (Smith & Devine, 1985). Theoretical arguments for some of the variations as phrase sizes increase are provided by Egghe (1999). The relationship between rank and frequency shifts in a consistent way as one increases the size of the phrases being analyzed from

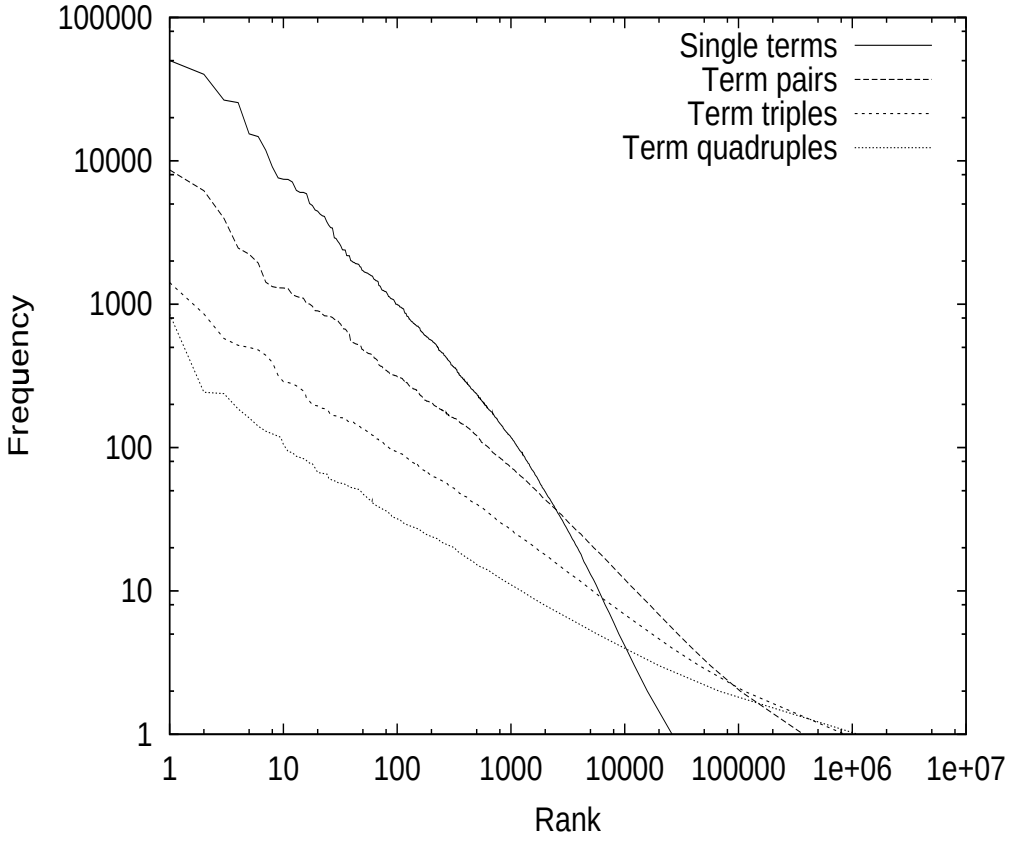

Figure 4: Phrase frequency versus phrase rank for phrases composed of one term (upper-right) to phrases composed of four adjacent terms (lower-left).

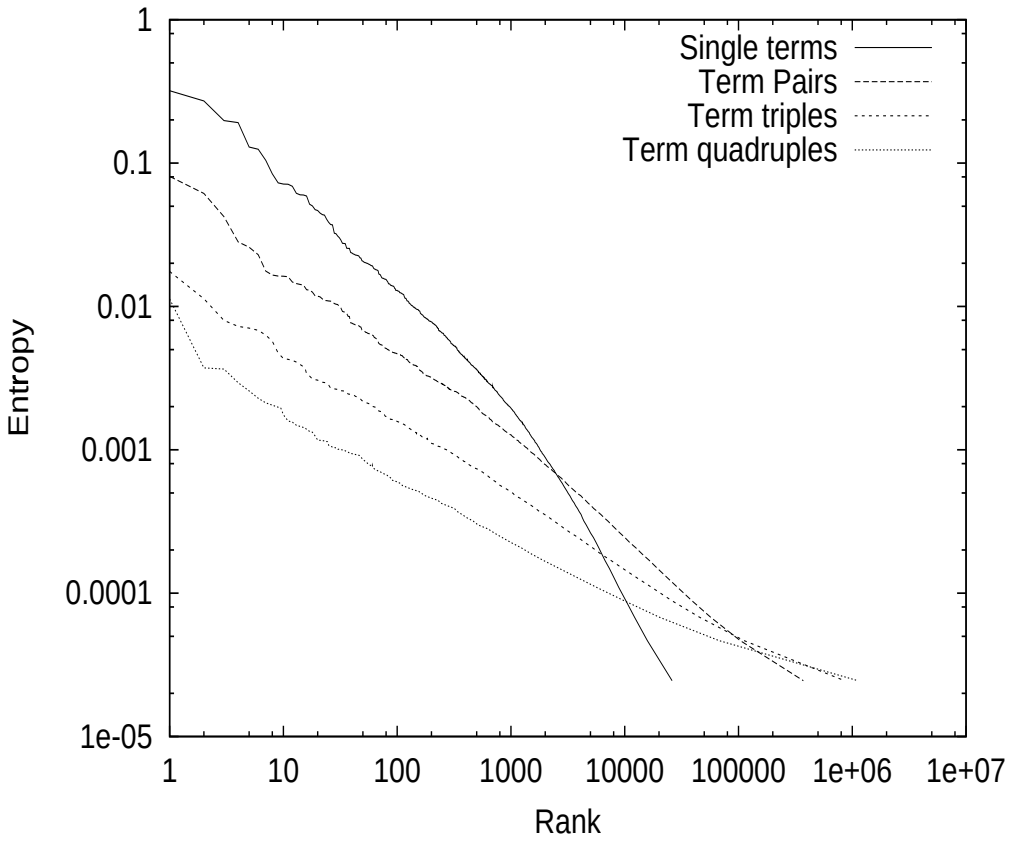

Figure 5: The entropy of phrases decreases as the phrase rank increases. Phrases of a single term are to the upper-right and phrases of 4 consecutive terms are to the lower-left.

being one term (toward the upper-right) to phrases containing four consecutive terms (toward the lower-left). Of interest is the crossover in Figure 4 that also occurs in Smith and Devine's datasets. Figure 5 similarly shows the relationship between entropy and rank for phrases consisting of from one to four consecutive terms. Following earlier work relating rank, frequency, and entropy for different sized phrases (Smith & Devine, 1985; Yavuz, 1974), the entropy for the phrases shows a strong regularity with the size of the phrase and the rank of the phrase.

Clearly, the data in Figure 5 suggests that Zipf's Law and its consequent regularities can be expressed as a relationship between term entropy and rank similar to the more traditional form of the Law as relationship between term frequency and rank.

Entropy is shown empirically to decrease regularly as the rank of terms in-

crease, and, similarly, the entropy decreases as the frequency of terms decreases. The definitional relationship between phrase probability and entropy is described by Equation 1.

The curve representing single terms crosses over the curve representing term pairs near the bottom of Figures 4 and 5. In addition, other crossovers occur. The crossovers are due primarily to the different terms that occur in phrases of a given length or size. The number of unique phrases of length n is always greater than or equal to the number of unique phrases of length $n-1$. Given x unique phrases of length $n-1$, there must be at least x unique phrases of length n because there would be x unique phrases of length n beginning with the same x phrases composed of $n-1$ terms. If some of the x unique phrases of length $n-1$ were followed by more than one different term when made into a unique phrase of length n, there would be more than x unique phrases of length n. Clearly, phrases with a given term will almost always have different ranks in phrases of different sizes. Given the increased number of unique phrases as the length increases, phrases with some of the same terms will not have the same rank, may shift in rank in a regular manner, and won't be above and below each other in the Figures. Put diferently, as we move toward longer phrases, the number of types increases and the probability of each type decreases, changing the characteristics of each line to yield the slope and crossover seen in the Figures. This loosely explains the different plots for different phrase lengths and suggests why the crossover phenomenon exists.

A similar analysis of entropy and rank has been performed on other types of textual data, which are described more fully in Losee (1996), including sets of documents in computational linguistics, theoretical and experimental physics, and classics of English language literature. The results are similar to those found in Figure 5. Minor differences are sometimes found. For example, note the change found in the entropy for longer, common phrases in English literature, as shown in Figure 6.

7 Luhn's Model of Term Aboutness

Hans Peter Luhn (1958) suggested that mid-frequency terms are the best indicators of topicality, and that very common and very rare terms are weaker discriminators. This is in contrast to the Inverse Document Frequency (IDF) based model that implicitly suggests that terms continue to increase in worth as index terms as the relative frequency of documents with the term decreases (Sparck Jones, 1972). This is similar to ranking by overall term rarity in the database. The same arguments below apply to the constrast that can be drawn between Luhn's model

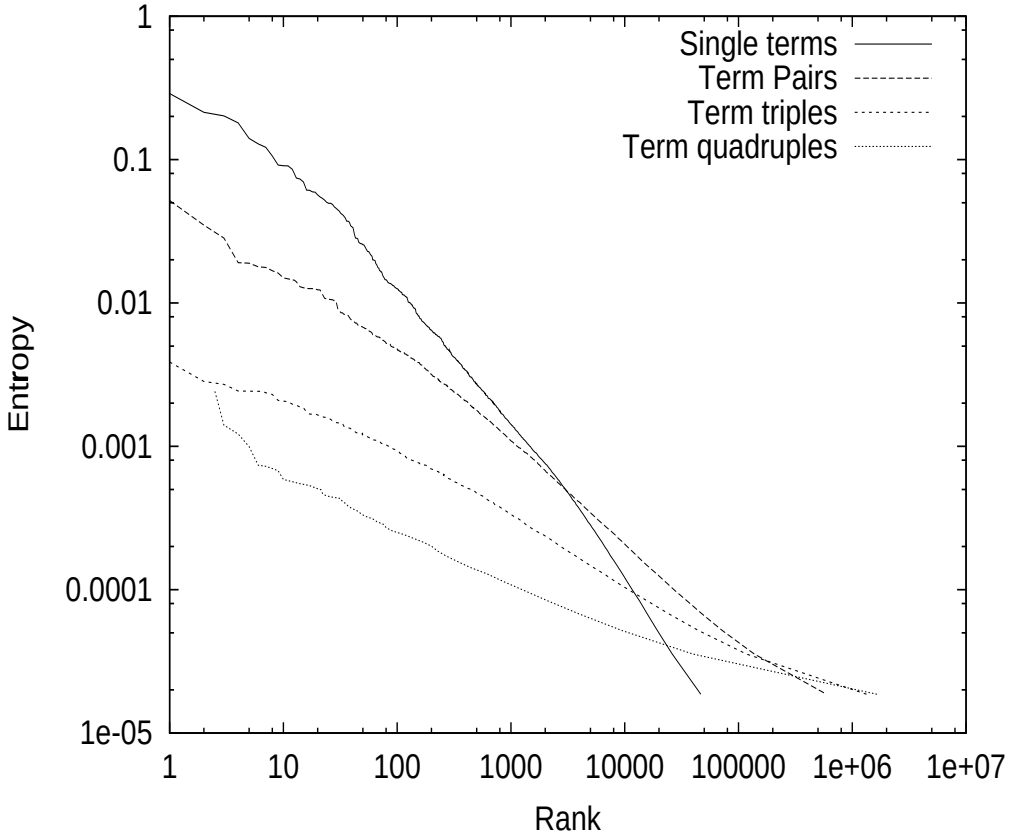

Figure 6: For documents representing English literature, the entropy of phrases decreases as the phrase rank increases. Phrases of a single term are to the upper-right and phrases of 4 consecutive terms are to the lower-left.

and other probabilistic models of text, including entropy based models (Shannon, 1951; Salton & McGill, 1983; Losee, 1990) and term discrimination models (Robertson & Sparck Jones, 1976; Salton & McGill, 1983; Losee, 1998).

Figure 3 shows that as one approaches the mid-frequency terms from either the higher or lower frequency terms, the EMIM between the terms in a two term phrase decreases. The minimum EMIM occurs between terms with probabilities ranging from .001 to .0001.

Terms with lower EMIM values have more capability to provide self-information or "aboutness," because less of the information the term carries is determined by the neighboring terms. The data in Figure 3 suggest that as one moves from the most common terms, at the right side of the graph, toward the center, the amount of information that a term carries by itself becomes less influenced by surrounding terms. Once the center of the figure is passed and one moves further toward the left side, terms of lower frequency than the mid-frequency terms are encountered and the EMIM increases slightly, although not to the level found with the most common terms. This suggests that the very rare terms are not as good at discriminating as are the mid-frequency terms, but they are better than the very common terms.

Luhn's distribution and the EMIM-based explanation for it can be used to select natural language components for indexing or term weighting purposes, based on the likelihood that a term would be subject-bearing. For example, terms with lower EMIM are more likely to be subject bearing terms. By choosing phrases with the lowest EMIM values, we obtain phrases and individual terms that can be used for indexing. Sentences with higher rates than normal of low EMIM terms might be assumed to be more subject bearing than other sentences. These sentences might be included in an automatically generated abstract or an automatic summary of a document.

The relative amount of information assumed by IDF weighting supports a different model of term valuation than do the EMIM results. The IDF weighting system is motivated by the desire to increase the term weight as we increase the specificity with less common terms (Sparck Jones, 1972). Thus, terms occurring in few documents have higher weights than common terms. However, rare phrases may have lower EMIM values between terms than do mid-frequency phrases.

Terms that receive less information about neighboring terms are those terms that Luhn predicts will be the best discriminators. We refer to the empirical relationships discussed above that support Luhn's model as the *Inverse Representation—EMIM* rule. Those terms that best represent the subject of a document or text fragment are those producing the lowest EMIM values. This rule holds, on the average, over the full range of term frequencies.

Given this rule, we propose that terms may be weighted based on either term frequency (IDF) or based on the average EMIM between the term and its neighbors (Inverse EMIM Weighting). Because EMIM information, like IDF information, can be obtained without knowledge of what the user considers relevant, weighting a term by noting its average EMIM with neighbors may be computed by a system before being presented with a query. The use of this weighting allows for high processing speeds on search engines and other retrieval systems that depend on rapid responses to queries.

The Inverse EMIM weighting for term t is proportional to the logarithm of $1/E(EMIM_{t,z})$, where term t is the first term in the phrase, and where the average is computed over the set of Z values. Because a term has a neighbor on both sides, we might weight a term by taking the sum or the average of the weights from both directions. Unlike the IDF weight, mid-frequency terms have the highest weight, the least frequent terms have mid-level weights, and common terms have the lowest weights.

Clearly, this weighting is different than IDF weighting. Like the IDF weight, the Inverses EMIM weight can be pre-computed for each term before the system accepts any queries. The Inverse EMIM may be computed a number of ways, and the exact nature of the averaging may have a significant impact on the effectiveness of the weight in retrieval applications (Losee, 2001). Which averaging method performs better, and under what circumstances, remains an open question.

8 Term Relationships as a Partial Cause for the Zipfian Distribution

What causes rank-frequency relations as suggested by Zipf's Law, and what determines the statistical frequency or the entropy of a term? In earlier sections, we considered some of the observable relationships that exist between term and phrase characteristics. Using information theoretic measures, we have been able to show relationships between information (entropy and EMIM), term frequencies, and ranks. While there may be other factors that contribute to the development of power laws such as Zipf's Law (Anderson & Tweney, 1997), the model we suggest below can be used to partially account for observed term frequencies and thus for some of the form taken by Zipf's Law.

How does the Zipfian model arise from terms in the set of natural language statements? Let us begin with the complete set of statements in a language. The terms in this set have an associated joint entropy. Beginning with this joint entropy and by the repeated application of the dependencies between terms, we can arrive

at individual term characteristics, and thus at some characteristics of Zipf's Law. We believe that this is the primary causal factor for Zipf's Law, and that the inverse isn't true, that Zipf's Law is the cause of the joint entropy for natural language, taken as a whole. Instead, the initial distribution of all languages, combined with term dependencies that reflect naturally occurring dependencies in the physical world, are the first causes. These ultimately cause the occurrences of individual terms, and thus Zipf's Law.

In the world, there are dependencies between objects and their characteristics. *Rocks* and *soil* co-occur with some frequency, as do *blizzards*, *northern latitudes*, and the *atmosphere*. The corresponding terms that represent these concepts have similar dependencies. Term relationships indicative of phenomenological dependencies are a frequent and regular component of all natural languages, and probably of synthetic languages as well.

The derivation of single term characteristics from term pairs is as follows. Beginning with Equation 3, we find that for random term variables T_1 and T_2,

$$H(T_1) = I(T_1; T_2) + H(T_1, T_2) - H(T_2).$$

The entropy of the first term in a pair is totally determined by the sum of the entropy of the pair, the EMIM between the terms, and the negation of the entropy of the second term. Let us assume that $H(T_2)$ approximates $H(T_1)$. As an equality, this happens infrequently. However, as an approximation, it can simplify our argument so we may more clearly see the relationship between other factors and term entropy. Thus, if they are similarly ranked terms, we may estimate

$$H(T_1) \approx \Big(I(T_1; T_2) + H(T_1, T_2)\Big)/2.$$

Here the value of $H(T_1)$ is completely determined by knowledge of $I(T_1; T_2)$ and $H(T_1, T_2)$, the EMIM between the terms and the joint entropy of the term pair.

For a fixed $H(T_1, T_2)$, increasing $I(T_1; T_2)$ can be said to increase $H(T)$. Therefore, the single term entropy can be understood as a function of the joint entropy and of the information the terms provide about each other. Using similar relationships, one can move from the entropy of language as a whole toward the characteristics of individual terms.

Figure 5 shows that the entropy increases as the term or phrase's rank decreases. Similarly, Figure 2 suggests that the EMIM increases as the term rank decreases. The entropy for a single term is generally greater than the entropy for a phrase composed of multiple terms. Given the above relationships, we can see how the entropy associated with a single term may be derived from the lesser

amount of information (entropy) in a more complex grouping of terms, along with the EMIM associated with the phrase. In general, the entropy for a phrase of length n may be based in part on the entropy for a superset phrase of length $n + 1$ and the EMIM associated with the phrase of length $n + 1$. We suggest that the changes between phrases of different lengths in Figure 5 are due to the EMIM components shown in Figure 2.

9 Discussion and Conclusions

We have suggested here that relationships exist between term and phrase frequencies and ranks, and between them and information theoretic concepts, such as entropy and the expected mutual information measure (EMIM). In Figure 1 we showed the relationship between the probabilities of term-pairs and their EMIM, with Figure 2 similarly showing the relationship between EMIM and term-pair rank. As term-pairs increase in probability we find that their EMIM also increases. Interestingly, the regularity of this relationship holds most strongly for less frequent term-pairs, with more common term-pairs having a greater diversity of EMIMs.

In a manner somewhat similar to that shown with other studies of Zipf's Law (Smith & Devine, 1985), Figure 5 shows us how the entropy or average information in a phrase decreases as the phrases become less common. This relationship holds for a number of phrase sizes, with sizes of one through four being studied here. Figure 5 shows a downward shift in entropy values as the number of terms grouped together increases.

We can locate those terms that best represent the subject of a document using these information theoretic methods. Figure 3 shows how EMIM varies with individual terms frequencies from within term-pairs. This relationship can be used to explain Luhn's model of the interaction between term frequencies and their associated term discrimination values. Phrases with mid-frequency terms have lower EMIM about their neighbors than do high or low frequency terms. The mid-frequency terms are less dependent on their neighbors and are more informative about themselves. We refer to this as the Inverse Representation–EMIM principle. By finding the low EMIM point in Figure 3 and moving a short distance to the left and to the right, we can select those terms that best serve as indexing terms. An Inverse EMIM weight is suggested that weights terms for indexing and retrieval purposes and that is consistent with Luhn's model of the relationship between term discrimination capabilities and term frequencies.

We may choose to view Zipf's Law as a consequence of the relationships that

exist between terms. Given language taken as a whole, and with the repeated application of equations relating joint entropy, EMIM, and individual entropy, we arrive at term frequencies and ranks that can produce the term frequencies in Zipf's Law. It seems more natural to accept that Zipf's Law is caused by these more general concerns, rather than Zipf's Law arbitrarily causing relationships in language as a whole.

There are several regularities that exist in natural language. We have focused on those regularities that are based in term dependencies, suggesting practical applications for these relationships, including indexing and the automatic extraction of key sentences. Also, we have offered partial explanations for popular linguistic models, proposed by Zipf and Luhn, in terms of information theoretic and empirically based relationships.

References

Aczél, J., & Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. Academic Press, New York.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*(5), 724–730.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *IT-14*(3), 462–467.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley Interscience, New York.

Cox, D. R., & Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, *8*(3), 204–218.

Croft, W. B. (1986). Boolean queries and term dependencies in probabilistic retrieval models. *Journal of the American Society for Information Science*, *37*(2), 71–77.

Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society for Information Science*, *50*(3), 233–241.

Gey, F. C. (1993). *Probabilistic Dependence and Logistic Inference in Information Retrieval*. Ph.D. thesis, U. of California, Berkeley.

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal*, *7*, 535–563.

Losee, R. M. (1990). *The Science of Information: Measurement and Applications*. Academic Press, San Diego.

Losee, R. M. (1994). Term dependence: Truncating the Bahadur Lazarsfeld expansion. *Information Processing and Management*, *30*(2), 293–303.

Losee, R. M. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing and Management*, *32*(6), 747–767.

Losee, R. M. (1998). *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer, Boston.

Losee, R. M. (2001). Natural language processing in support of decision-making: Phrases and part-of-speech tagging. *Information Processing and Management, In Press.*

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159–165. The article is also included in *H. P. Luhn: Pioneer of Information Science, Selected Works*.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. In *Structure of Language and Its Mathematical Aspects: Proceedings of Symposia in Applied Mathematics, vol. XII*, pp. 190–219. American Mathematical Society.

Moon, S. B. (1993). *Enhancing Retrieval Performance of Full-Text Retrieval Systems Using Relevance Feedback*. Ph.D. thesis, U. of North Carolina, Chapel Hill, NC.

Naranan, S., & Balasubrahmanyan, V. K. (1998). Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics, 5*(1–2), 35–61.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Bell System Technical Journal, 3*, 324–346.

Rapoport, A. (1982). Zipf's law re-visited. *Quantitative Linguistics, 16*(1), 1–28.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 129–146.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30*, 50–64.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Ill.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425–440.

Smith, F. J., & Devine, K. (1985). Storing and retrieving word phrases. *Information Processing and Management, 21*(3), 215–224.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Yavuz, D. (1974). Zipf's law and entropy. *IEEE Transactions on Information Theory, IT-20*(5), 650.

Yngve, V. (1986). *Linguistics as a Science*. Indiana University Press.

Yu, C. T., Buckley, C., Lam, K., & Salton, G. (1983). A generalized term dependence model in information retrieval. *Information Technology: Research and Development, 2*(4), 129–154.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, Mass.