THE DISTRIBUTION OF ENGLISH LANGUAGE WORD PAIRS

LINAS VEPSTAS

ABSTRACT. This short note presents some empirical data on the distribution of word pairs obtained from English text. Particular attention is paid to the distribution of mutual information.

1. INTRODUCTION

The mutual information of word pairs observed in text is commonly used in statistical corpus linguistics for a variety of reasons: to help identify idioms and collocations, to create dependency parsers, such as minimum spanning tree parsers, and many other uses. This is because the mutual information of a word-pair co-occurrence is a good indicator of how often two words 'go together'. However, little seems to have been written on the distribution of word pairs, or the distribution of mutual information. The current work is a short note reporting on the empirical distribution of word pairs seen in English text.

Mutual information is a measure of the relatedness of two words. For example "Northern Ireland" will have high mutual information, since the words "Northern" and "Ireland" are commonly used together. By contrast, "Ireland is" will have negative mutual information, mostly because the word "is" is used with many, many other words besides "Ireland"; there is no special relationship between these two words. High mutual information word pairs are typically noun phrases, often idioms and collocations, and almost always embody some concept (so, for example, "Northern Ireland" is the name of a place — the name of the conception of a particular country).

2. DEFINITIONS

In what follows, a "pair" will always mean an "ordered word pair"; an ordered pair is simply a pair (x,y) where both words x and y occur in the same sentence, and the word x occurs to the left of the word y in that sentence. The pairs are taken to be ordered because, in English, word order matters. Some of the graphs below show pairs of neighboring words; others show graphs of pairs simply taken from the same sentence (so that other words may intervene in the middle of the pair).

The frequency of an ordered word pair (x,y) is defined as

$$P(x,y) = \frac{N(x,y)}{N(*,*)}$$

where $N(x,y)$ is the number of times that the pair (x,y) was observed, with word x to the left of word y. The * represents 'any', so that $N(*,*)$ is the total number of word pairs observed.

Date: 13 March 2009.

Key words and phrases. statistical linguistics, Zipf's law, power-law distribution.

The mutual information $M(x,y)$ associated with a word pair (x,y) is defined as

$$M(x,y) = log_2 \frac{P(x,y)}{P(x,*)P(*,y)} \tag{2.1}$$

Here, $P(x,*)$ is the marginal probability of observing a pair where the left word is x; similarly, $P(*,y)$ is the marginal probability of observing a pair where the right word is y.

In the following, a selection of 925 thousand sentences consisting of 13.4 million words were observed, from a corpus consisting of Voice of America shows, Simple English Wikipedia, and a selection of books from Project Gutenberg, including Tolstoy's 'War and Peace'.

3. Unigram counts, Zipf's law

Individual words, ranked by their frequency of occurrence, are known to be distributed according to the Zipf power law; this was noted by Zipf in the 1930's[4, 3, 1], and possibly by others even earlier. The Zipf distribution takes the form

$$P_s(k) \sim k^{-s}$$

where s is the characteristic power defining the distribution. Here, k is the k'th ranked word, and $P(k)$ is the probability of observing that word. The power law is illustrated in figure3, and is generated from the current corpus. As may be seen, power laws become straight lines in a log-log graph.

The Zipfian distribution for texts appears to be entirely due to a side effect of the ranking of words [2]. That is, if one generates random texts, one can observe the very same Zipf law, with the same parameters. The way that this may be demonstrated is straight-forward: one begins with an alphabet of N letters, one of which represents a space (thus, $N = 26 + 1 = 27$ for English, not counting letter cases). One then picks letters randomly from this set, with a uniform distribution, to generate random strings. If the space character is picked, then one has a complete word. After ranking and plotting such randomly generated words, one obtains a Zipfian distribution with the same power law exponent as above. More precisely, it may be shown that the exponent is

$$s = \frac{\log(N+1)}{\log N}$$

for an alphabet of N letters [2]. This begs a question: what is the mutual information content of a random text?

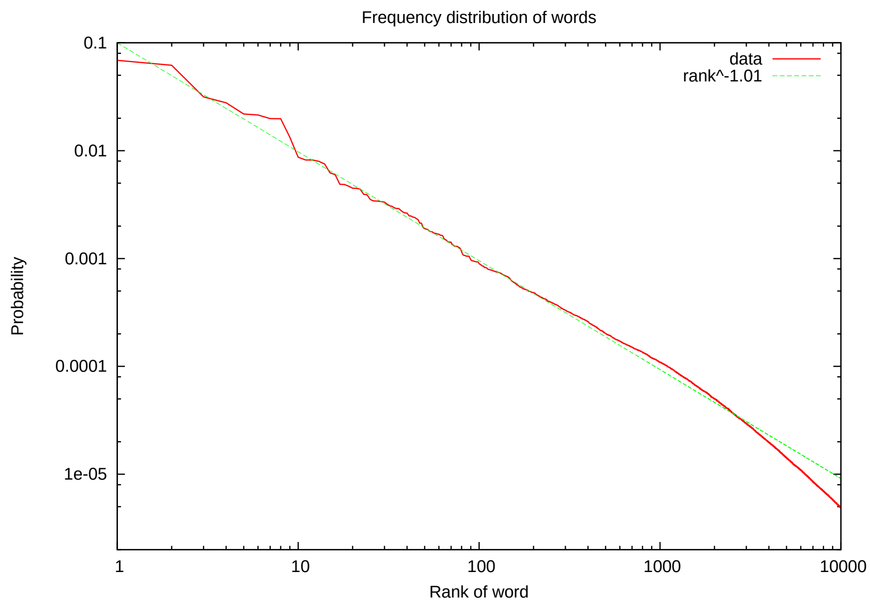
4. Distribution of Mutual Information

How is mutual information distributed in the English language? The graph 4.1 shows the (unweighted) distribution histogram of mutual information, while the graph 4.2 shows the weighted distribution.

The figures are constructed by means of "binning" or histograming. For a given value of mutual information M and bin width ΔM, the (unweighted) bin count is the total number of pairs (x,y) with $M < M(x,y) \leq M + \Delta M$. The weighted bin count is defined similarly, except that each pair contributes $P(x,y)$ to the bin count.

Notable in the first figure is a distinct triangular shape, with a blunted, lop-sided nose, and "fat" tails extending to either side. The triangular sides appear to be log-linear over many orders of magnitude, and seem to have nearly equal but opposite slopes.

FIGURE 3.1. Single Word Distribution



The graph above shows the frequency distribution of the most commonly occurring words in the sample corpus, ranked by their frequency of occurrence. The first point corresponds to the beginning-of-sentence maker; the average sentence length of about 14.4 words corresponds to a frequency of 1/14.4=0.07. This is followed by the words "the, of, in to, and", and so on. The distribution follows Zipf's power law. The thicker red line represents the data, whereas the thinner green line represents the power law:

$$P(k) = 0.1k^{-1.01}$$

where k is the rank or the word, and $P(k)$ is the frequency at which the k'th word is observed. Notice that after the first thousand words, the probability starts dropping at a greater rate than at first. This is bend is commonly seen such word frequency graphs; it is not covered by Zipf's distribution. To the author's best knowledge, there has not been any published analysis of this bend.

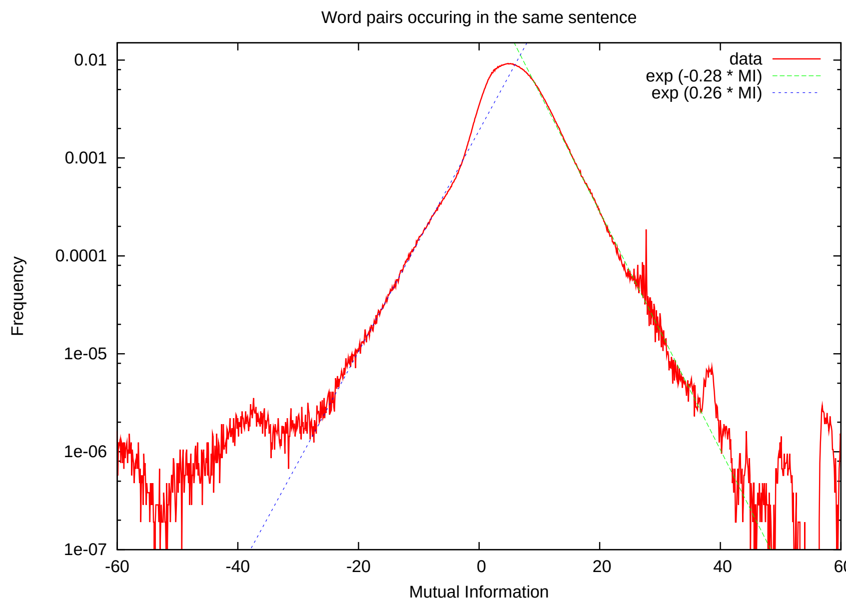
Triangular distributions are the hallmark of the sum of two uniformly distributed random variables: the convolution of two rectangles is a triangle. Taking equation 2.1 and propagating the logarithm through, one obtains

$$M(x,y) = \log_2 P(x,y) - \log_2 P(x,*) - \log_2 P(*,y)$$

which can be seen to be a sum of three random variables.

XXX pursue this idea.

FIGURE 4.1. Histogram of Mutual Information



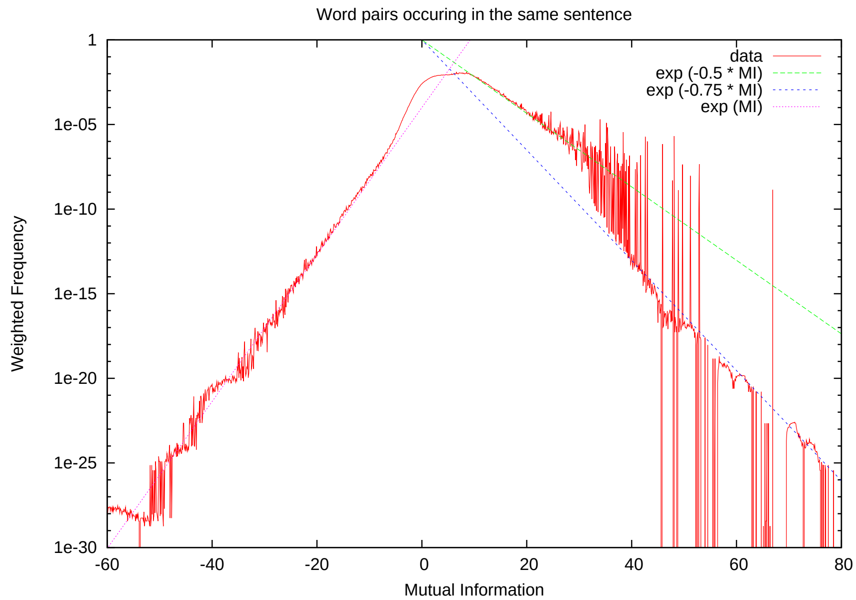
This graph shows the distribution histogram of the mutual information of word pairs taken from the text corpus. The figure shows approximately 10.4 million distinct word pairs observed, with each distinct word pair counted exactly once. The sides of the distribution are clearly log-linear, and are easily fit by straight lines. Shown are two such fits, one proportional to $e^{-0.28M}$ and the other to $e^{0.26M}$ where M is the mutual information.

5. Random Texts

As has been noted, the Zipfian distribution appears to be entirely explainable as a side-effect of choosing word rank as an index. The question naturally arises: which of the features of the distribution of mutual information are associated with the structure of language, and which are characteristic of random text? This question is explored with two different ways of generating random text.

In the first approach, random text is generated by generating random strings. An alphabet of 12 letters and one space was defined ("etaion shrdlu"). A random number generator was used to pick letters from this alphabet; when the space character was picked, the string up to that point is declared a "complete word", and added to the current sentence. A new word is then started, and the process is repeated, until a total of thirteen words were assembled into a sentence. Once such a randomly-generated sentence was at hand, it was handed over to the word-pair counting subsystem. This process was repeated to obtain a total of 4.1 million words and 20.5 million word pairs. The distribution of mutual information is shown in figure 5.1.

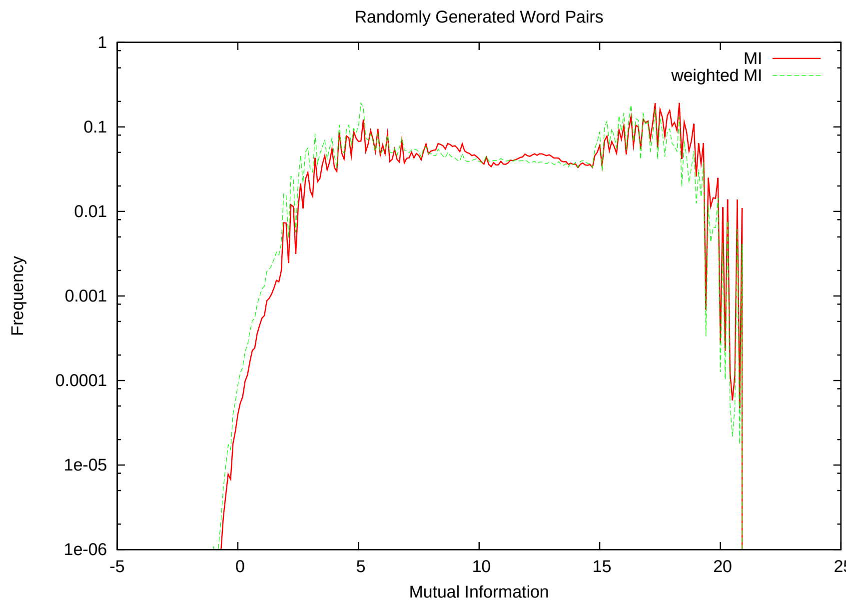
FIGURE 4.2. Weighted Histogram of Mutual Information



This graph shows the weighted frequency of mutual information. Unlike the previous figure 4.1, where each pair was counted with a weight of 1.0, here, each pair is counted with a weight of $P(x,y)$. That is, for a fixed value of mutual information, frequent pairs with that value of mutual information contribute more strongly than infrequent ones. Several different log-linear behaviors are visible; these have been hand-fit with the straight lines illustrated; these are given by $e^{-0.5M}$, $e^{-0.75M}$ and e^{M}. The two slopes on the right hand side require no further scaling: they intercept $P = 1$ at $M = 0$. The appearance of such simple fractions in the fit is curious: changing the slopes by as little as 5% leads to a noticeably poorer fit. Whether this is a pure coincidence, or the sign of something deeper, is unclear.

One problem immediately evident from these statistics is that the random generator produced far too many unique, distinct words. This is despite several efforts that were taken to limit word size: the likelihood of choosing a space was dramatically increased, so as to generate relatively short strings. Also, a hard-cutoff was forced to limit all strings to a length of 8 characters or less. Nonetheless, this means that there are, in principle, $12^8 + 12^7 + \cdots \approx 4.7 \times 10^8$ distinct "words" in the language, and 10^{17} observable word pairs. Thus, the 20 million observed word pairs are just a minuscule portion of all possible words pairs. Each observed word (and only a tiny fraction of all possible words were observed) participates in only a relatively small number of possible word pairs. Thus, simply due to sampling effects, one finds that almost all pairs have a high mutual information content: there have not been enough pairs observed to realize that the words are "maximally promiscuous" in their pairing.

FIGURE 5.1. Mutual Information in Random Text

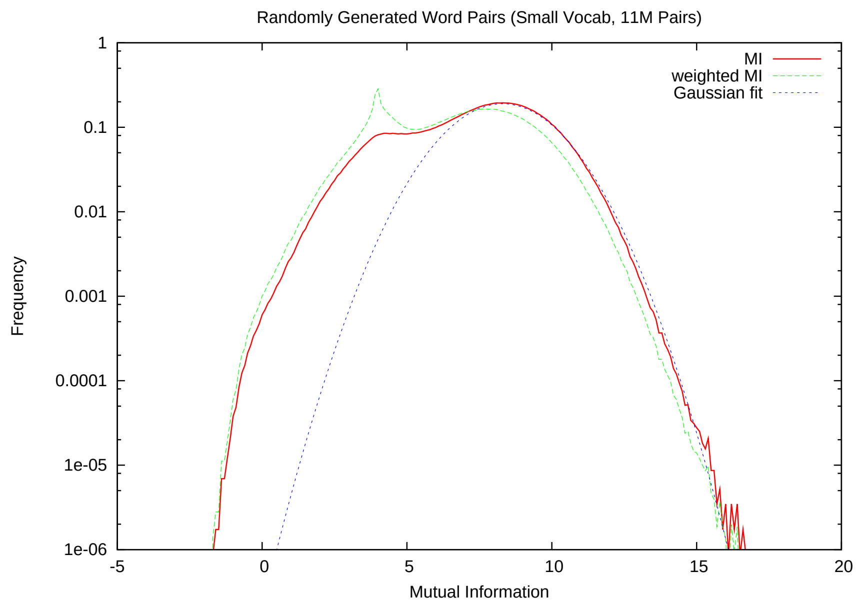


This figure shows the (weighted and unweighted) mutual information obtained from analyzing randomly generated text. While the overall shape is quite different than that for English text, shown in figures 4.1 and 4.2, it is perhaps surprising that the average mutual information is quite high. This is because, despite the small alphabet, and the relatively small word size, and the seemingly large sample of word pairs, only a tiny fraction of all possible word pairs was observed. This means that most words participate in only a relatively small number of pairs, and thus get a defacto high value of mutual information. Whether or not the left or right sides of the figure are best fit by Gaussian or plain exponentials is not clear. That perhaps the fall-off behavior is Gaussian is suggested by the following two graphs.

To overcome this problem, the experiment is repeated, with a word length limit of 4 characters. The total vocabulary then consists of $12^4+12^3+12^2+12=22620$ words. Given a vocabulary of this size, there are about 512 million word pairs possible. After observing 11.5 million distinct pairs (or about 2.2% of all possible pairs), the mutual information is graphed in figure 5.2. Note that the emphasis is on "distinct" pairs: at least 25% of the pairs were observed at least twice. Because many of the generated words are short (two or three letters long), and are generated fairly often, many of the pairs were observed dozens, hundreds or thousands of times.

Taking the small-vocabulary idea to an extreme, a third experiment is conducted, where the vocabulary consists of all words of length 3 or less, constructed from 12 letters: a total vocabulary of $12^3+12^2+12=1884$ words. There are a total number of 3.5 million possible word pairs. Generating random sentences, as described previously, a total of

FIGURE 5.2. Small Random Vocabulary



This figure shows the distribution of mutual information, obtained by sampling randomly created sentences. Unlike the graph shown in figure 5.1, this figure shows a much smaller vocabulary: that consisting 22.6 thousand words. The generated text contained 11.5 million distinct observed word pairs, or about 2.2% of all possible random word pairs. This graph retains the strong offset to positive MI, but is far more gently humped; the flat top has all but disappeared. The right hand side of the curve is clearly Gaussian; the dashed blue line is given by $0.19\exp -0.2(MI-8.3)^2$.

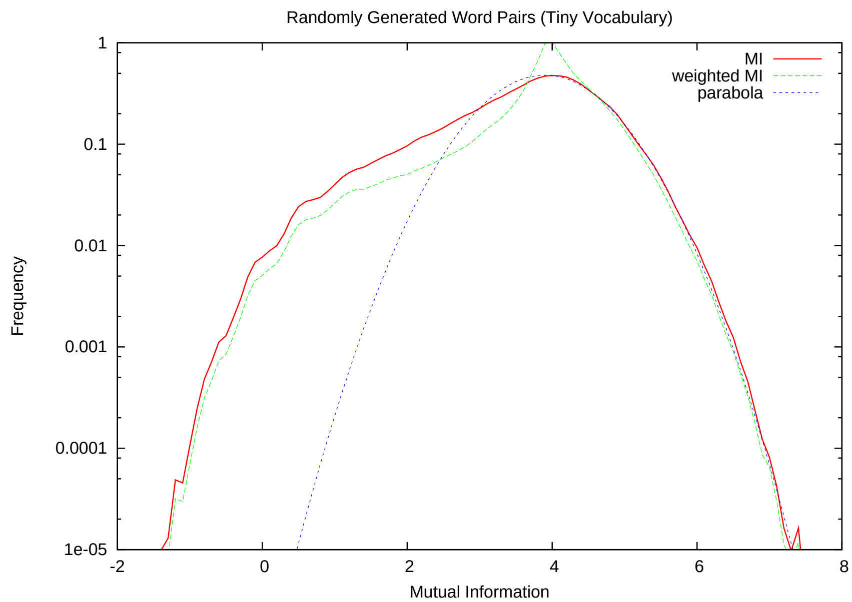
16.6 million word pairs were generated, with a total of 3.07 unique pairs observed. The distribution of mutual information is graphed in figure 5.3.

6. Linked Words

Instead of considering all possible word pairs in a sentence, one might instead limit consideration to words that are in some way related to one-another. To that end, the same corpus was parsed by means of the Link Grammar parser (need ref), to obtain pairs of words linked by Link Grammar linkages.

The basic insight of Link Grammar is that every word can be treated as a puzzle-piece, in that a word can only connect to other words that have matching connectors. The connector types are labeled; the natural output of parsing is a set of connected word pairs. It is an important coincidence to notice that, usually, the word pairs that the parser generates also have positive, and usually a large positive value of mutual information. This observation provides insight into why the Link Grammar "works" or is a linguistically successful theory of parsing: its rule set captures word associations.

FIGURE 5.3. Mutual Information for a Tiny Vocabulary



This figure shows the distribution of mutual information for a tiny vocabulary, consisting of 1884 possible words, and 3.5 million word pairs. Enough sentences were generated to count a total of 16.6 million word pairs, of which 3.07 million were unique. The graph is still clearly shifted: some word pairs were observed far more frequently than others. This time, the right-hand side of the curve is well-fitted by a Gaussian: the dashed blue line, labeled as "parabola" in the image, is given by $0.48 \exp -0.92(MI-3.9)^2$.

The figures 6.1 and 6.2 reproduce the previous two graphs, except that this time, the set of word pairs is limited to those that have been identified by Link Grammar. This set consists of 2.37 million distinct word pairs.

7. MUTUAL INFORMATION SCATTERPLOTS

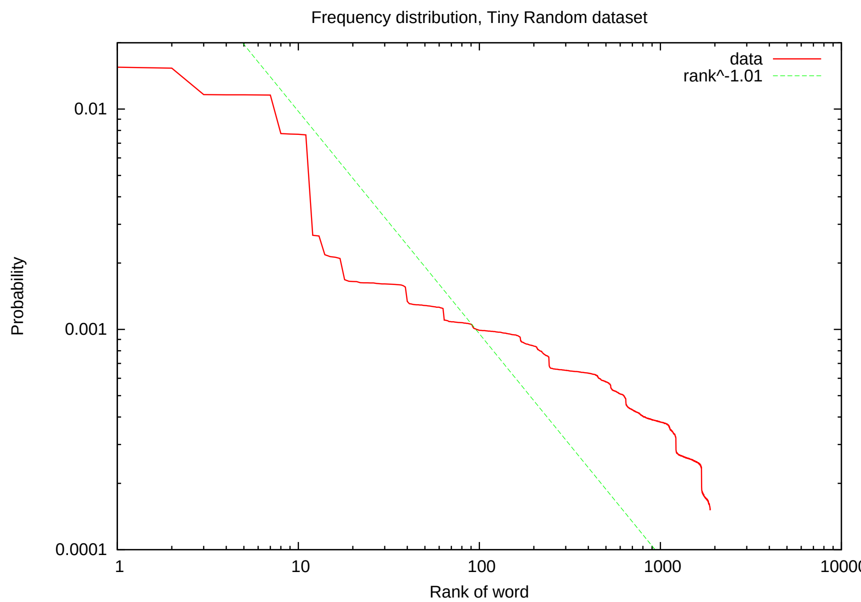
The previous sections explored the histogram distribution of the mutual information of word pairs. Alternately, one might wish to understand how mutual information correlates with the probability of observing word pairs. A scatterplot showing such a correlation is shown in figure 7.1. Slices through this scatterplot are shown in figure 7.2.

8. CONCLUSION

Some conclusion to be written.

One thing that I've learned from these experiments is that the rightward skew is not at all something that is "capturing actual associative structure in the language", but is arguably yet another statistical artifact.

FIGURE 5.4. Zipf for tiny vocab



Distribution of words in the tiny data set. For comparison, the straight dashed green line represents the Zipf law distribution $0.1k^{1.01}$.

Simply by drawing random pairs out of a bag, one will, for smaller sample sizes, see a rightward shift in the MI distribution, simply because one has not yet seen *all* possible random word pairs. Thus, the small sample size make it look like some words are associated with others, even though the association is purely accidental.

How big a sample size is this? I tried a "tiny" vocabulary of 1884 words, for a total of 3.5M possible word pairs. After randomly drawing 16M pairs (and thus observing 3.07M unique pairs), I still saw a strong rightward shift in the MI.

I'm now guessing that, in order to have this sample-size effect wash out, one needs to observe at least N^3 pairs for a vocabulary of N words. How big is this number? for English, its huge: assuming an "average" speaker vocabulary of 30K words, one would have to sample (30K)^3=27 trillion word pairs before one could argue that a right-ward shift in the MI graph is due to the "associative structure of language". Yikes!

REFERENCES

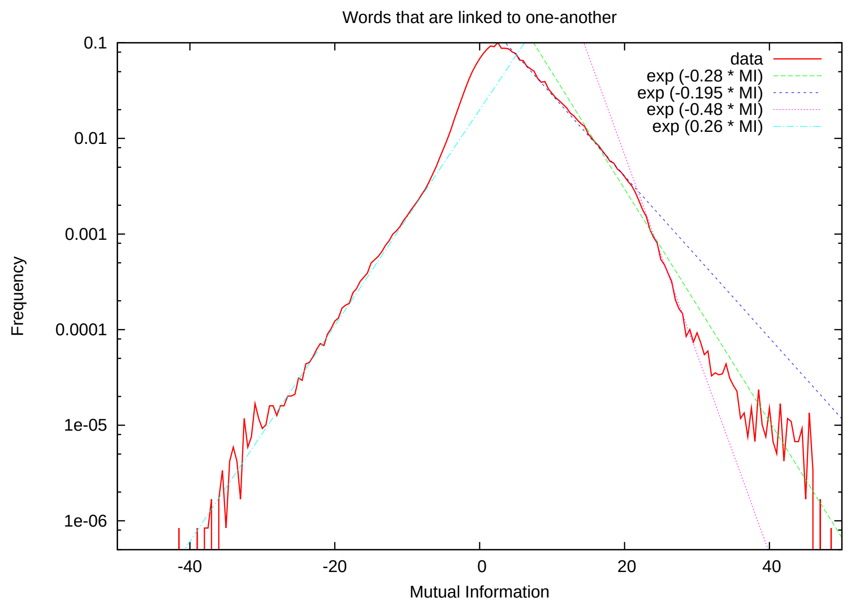
[1] Henry Kucera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University, 1967.

[2] Wentian Li. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845, 1992.

[3] George Kingsley Zipf. *Selective Studies and the Principle of Relative Frequency in Language (Cambridge, Mass, 1932)*. xx, 1932.

[4] George Kingsley Zipf. *The Psychobiology of Language*. Houghton-Mifflin, 1935.

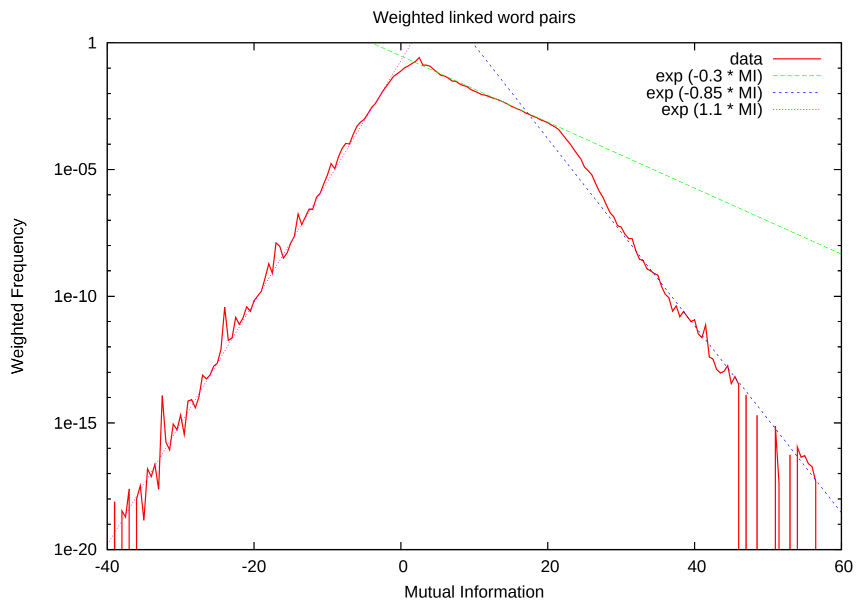
FIGURE 6.1. Histogram of Linked Word Pairs



This figure shows a histogram of word pairs that have been linked to one-another by means of the Link Grammar parser. The data set contains approximately 2.37 million such word pairs; each is counted precisely once (thus, this is the "unweighted" distribution). Several log-linear regimes are visible. That on the left has exactly the same slope as in figure 3, namely $e^{0.26M}$. The average slope on the right is as before, $e^{-0.28M}$, but seems to be composed of two distinct regions, first with a shallower, and then a steeper slope of $e^{-0.195M}$ and $e^{-0.48M}$.

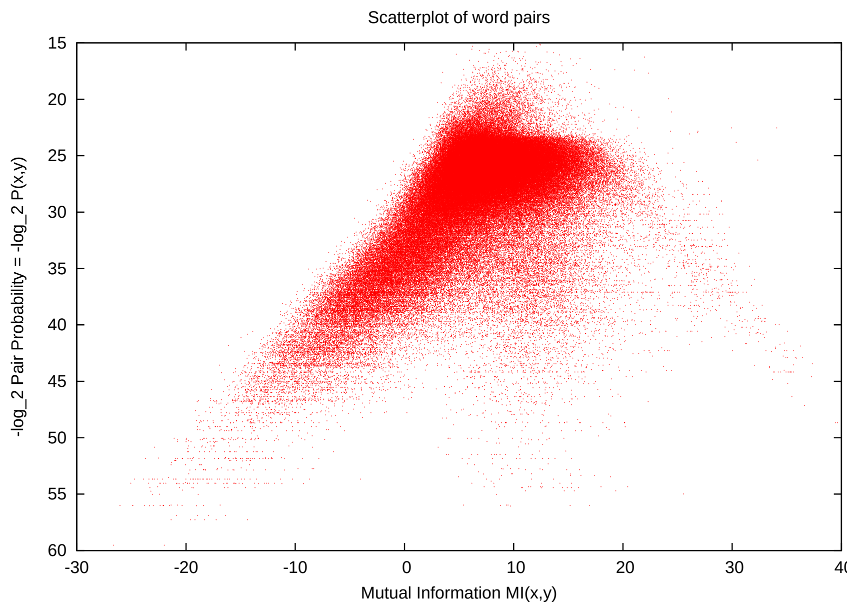
E-mail address: `linasvepstas@gmail.com`

FIGURE 6.2. Weighted Linked Words



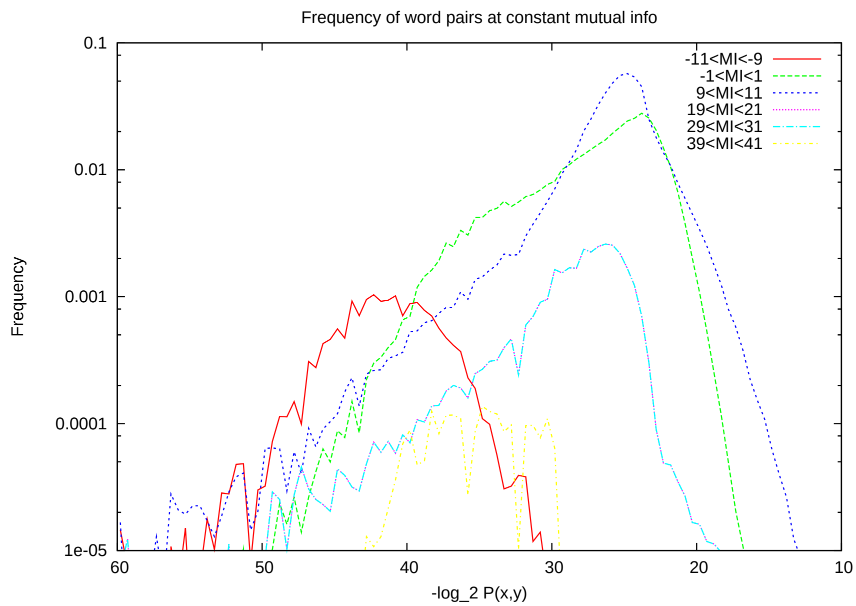
This figure shows weighted linked word pairs; it is the analogue to figure 4.2, but for linked word pairs. That is, the histogram bin counts are weighted by the probability $P(x,y)$ of seeing the linked word pair (x,y). Again, several sloped regions are visible; the linear fits are given by $e^{1.1M}$, $e^{-0.3M}$ and $e^{-0.85M}$.

FIGURE 7.1. Word Pair Scatterplot



This figure shows a scatterplot of the probability of occurrence $P(x,y)$ versus the mutual information $M(x,y)$ for word pairs (x,y). Shown are 200 thousand pairs (out of the sample set of 10.4 million pairs). Rather than plotting the probability $P(x,y)$ on a logarithmic axis, this plot uses a linear axis, labeled with $-log_2 P(x,y)$. This axis labeling gives a better feeling for the relative sizes of the horizontal and vertical scales.

FIGURE 7.2. Slices through the word-pair scatterplot



The above graph shows five slices through the mutual-info vs. pair-probability scatterplot. For each slice, the mutual information is held constant, and the frequency as a function of the log pair probability is graphed.