Connector Sets

Linas Vepstas

June 11, 2017

Abstract

Extract from the language-learning diary, reporting on the first small dataset containing connector sets. This is the 11 June 2017 update of the original 11 May 2017 report. It includes more data and figures, and updates the figures for readability (legibility). It also revises notation slightly.

Introduction

This is a report on a small dataset of disjuncts and connector sets, extracted from MST parses of a batch of sentences. First, a recap of what these are, then a characterization of the database contents, and finally, a short report on the grammatical similarity of words in the dataset.

Summary of results

The primary results reported below are these:

* Most scores and metrics that can be assigned to connector sets give a (scale-free) Zipfian ranking distribution, and are thus fairly boring.

* The greater the average number of observations per disjunct, the more grammatically acceptable (accurate) the disjunct seems to be. This is good news: it means that the general technique is not generating ungrammatical garbage.

* Connector sets can be given a mutual information score. The ranking distribution for this is not at all Zipfian. The MI score seems to be quite good at identifying words that participate in idioms, set phrases and institutional phrases.

* The average number of connectors per disjunct, which should have indicated the part-of-speech that the word belongs to, fails to do this. This seems to be due to the small size of the dataset, the fact that it is polluted with lists and tables, masquerading as sentences (its a Wikipedia sample), and that there seem the be very few verbs in the sample (Wikipedia articles describe concepts and events, using the copula to describe them: “is”, “has”, “was”, and is almost devoid of narrative verbs: “ran” “jumped” “hit”, “ate” “thought” “took”.) The WP sources need to be supplemented with narrative texts, ideally from young-adult literature.

* Cosine similarity applied to connector sets seems to be an effective way of determining the grammatical similarity of words. This is despite the fact that closer examination reveals that the dataset is quite thin and poor, and the similarity is based on

scant evidence. Despite the scant evidence, the similarity scores do seem to distinguish different kinds of nouns quite well. A big surprise is that it also seems to cluster prepositions quite well. There seem to be few or no verbs in the dataset, as already noted. It seems that even small datasets are sufficient to get started with grammatical similarity clustering, but much larger samples are needed for discovering more sophisticated grammar.

Recap

The story so far: Starting from a large text corpus, the mutual information (MI) of word-pairs are counted. This MI is used to perform a maximum spanning-tree (MST) parse (of a different subset of) the corpus. From each parse, a pseudo-disjunct is extracted for each word. The pseudo-disjunct is like a real LG disjunct, except that each connector in the disjunct is the word at the far end of the link.

So, for example, in in idealized world, the MST parse of the sentence "Ben ate pizza" would produce the parse Ben <–> ate <–> pizza and from this, we can extract the pseudo-disjunct (Ben- pizza+) on the word "ate". Similarly, the sentence "Ben puked pizza" should produce the disjunct (Ben- pizza+) on the word "puked". Since these two disjuncts are the same, we can conclude that the two words "ate" and "puked" are very similar to each other. Considering all of the other disjuncts that arise in this example, we can conclude that these are the only two words that are similar.

Any given word will have many pseudo-disjuncts attached to it. Each disjunct has a count of the number of times it has been observed. Thus, this set of disjuncts can be imagined to be a vector in a high-dimensional vector space, which each disjunct being a single basis element. The similarity of two words can be taken to be the cosine-similarity between the disjunct-vectors.

Equivalently, the set of disjuncts can be thought of as a weighted set: each disjunct has a weight, corresponding to the number of times it has been observed. A weighted set is more or less the same thing as a vector, and these two are treated as the same, in what follows. Note that the disjunct vectors are sparse: for any given word, almost all coefficients will have a count of zero. For example, the dataset that will be examined next has over a quarter of a million different pseudo-disjuncts in it; most words have fewer than a hundred disjuncts on them.

Some terminology and notation are introduced next, followed by a characterization of the dataset. This is followed by a statistical analysis of the word-disjunct pairs, and is followed by an analysis of the resulting word-similarity.

Terminology

It is useful to introduce some notation for counting words, disjuncts, and connectors. Let $N(w)$ be the number of times that the word w has been observed, in the dataset. Let $N(w,d)$ be the number of times that the disjunct d has been observed on word w. The pair (w,d) is referred to as a "connector set" or "cset" in the text below. Thus, for a word w, there is a set $(w,*) = \{(w,d)|N(w,d) > 0\}$ of associated csets, called the "support" of the word. The size of this set can be written using the standard notation for set-sizes

as $|(w,*)|$. Similarly, a disjunct d, is supported by the set $(*,d) = \{(w,d)|N(w,d) > 0\}$ of associated csets.

The primary contents of the database are the counts $N(w,d)$ and everything else of interest in this section can be obtained from this. Note that $N(w,d)$ can be understood as a matrix, where the disjuncts identify columns, and the words identify rows. In general, this is a very sparse matrix: the number of non-zero entries $|(*,*)|$ is far less than the number of rows times the number of columns.

Every time a word is observed in an MST parse, a disjunct is extracted for it; thus, word observations and disjunct observations are on one-to-one correspondence. In notation:

$$\sum_d N(w,d) = N(w,*) = N(w)$$

Similarly, the total number of times that a disjunct was observed is just

$$N(*,d) = \sum_w N(w,d)$$

Frequencies can be obtained by dividing by the total number of observations, so that $p(w,d) = N(w,d)/N(*)$ and $p(w) = N(w)/N(*)$ with $N(*) = \sum_w N(w)$ the total number of observations of words.

A single disjunct is always composed of a fixed number of connectors, independently of any observations; let $C(d,c)$ be the number of times that connector c appears in disjunct d. Note that $C(d,c)$ is almost always either zero or one; however, a connector can appear more than once in a disjunct, so this count can rise to 2 or 3 or very rarely higher. The wild-card sum $C(d,*) = \sum_c C(d,c)$ is the total number of connectors in the disjunct; it is the vertex degree of all edges connecting to that disjunct. It is also useful to define $C(d,+)$ and $C(d,-)$ as the total number of right-linking and left-linking connectors.

Dataset characterization

The following charts and analyses are derived from a single dataset, a relatively small dataset, taken as a snapshot during counting. Its called 'en_pairs_sim'. It contains 37413 words that have disjuncts attached to them. These words have been observed a total of 661104 times, for an average of 661104/37413 = 17.6 observations per word. This dataset contains 291637 different, unique disjuncts, for an average of 661104/291637 = 2.27 observations per disjunct. The period appears 32536 times, suggesting that this many sentences were observed. Each sentence thus has an average of 661104/32536 = 20.3 words per sentence. The dataset contains 446204 unique connector-sets, for an average of 661104/446204 = 1.48 observations per cset. It is this last number that makes this dataset feel thin and sparse.

The dataset is sparse in a completely different sense: viewing $N(w,d)$ as a matrix whose size is 37413×291637, but only a very small number of these is non-zero: this is $446204/(37413 \times 291637) = 4.089 \times 10^{-5}$. The sparsity of this matrix can be defined as $-\log_2$ of this number, which is 14.58.

The total word-entropy for the dataset is defined as

$$H_{word} = -\sum_{w} p(w) \log_2 p(w)$$

and was measured to be $H_{word} = 10.28$ bits.[1] The connector-set entropy is much higher. It is defined as

$$H_{cset} = -\sum_{w,d} p(w,d) \log_2 p(w,d)$$

and is measured to be $H_{cset} = 18.30$ bits. The disjunct entropy is dual to the word entropy:

$$H_{disjunct} = -\sum_{d} p(*,d) \log_2 p(*,d)$$

and is measured to be $H_{disjunct} = 16.01$ bits. The total mutual information between the words and disjuncts is then

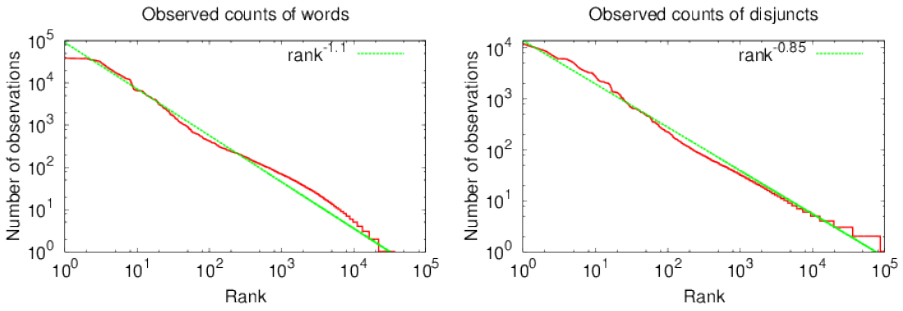
$$MI_{cset} = -\sum_{w,d} p(w,d) \log_2 \frac{p(w,d)}{p(*,d)p(w,*)} = H_{cset} - H_{word} - H_{disjunct}$$

and is measured to be $MI_{cset} = -7.985$ bits.

Connector-set distribution

Some connector-sets will be observed far more often than others. Likewise for the two sides of the connector-set: some words will have far more observations, and some disjuncts will be seen more often.

Two graphs, dual to one-another. The one on the left shows $N(w,*)$, ranked by count. The one on the right shows $N(*,d)$, also ranked.[2]. The first follows the canonical Zipf distribution. The green line is an eyeballed, approximate fit, of exponent -1.1. The second has an exponent of -0.85.



[1] This and the following entropies were measured with the word-entropy-bits, disjunct-entropy-bits, etc. functions in disjunct-stats.scm

[2] Obtained by running (print-ts-rank sorted-word-obs outport) from the disjunct-stats.scm file, on the en_pairs_sim database. The second one prints sorted-dj-obs.

The first ten words in the word ranking are: "the" "," "." "of" "and" "in" "to" "a" "was". This is the ranking of how often these words appear, overall, in the MST-parsed corpus. The number of periods should be equal to the number of sentences in the corpus; commas and the word “the” can appear more than once in a sentence. This list is repeated in the table below. The frequency is just the count divided by 661104.

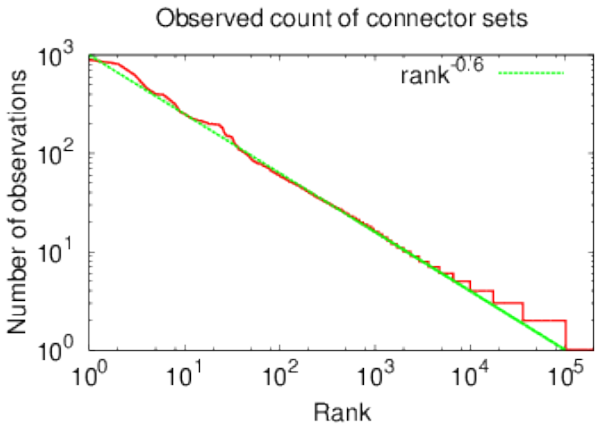
word	count	frequency	$-\log_2$frequency
the	38977	0.0589	4.084
,	37524	0.0568	4.139
.	32536	0.0489	4.353
of	22654	0.0343	4.867
and	17708	0.0268	5.222
in	14900	0.0225	5.471
to	12825	0.0194	5.688
a	11882	0.0180	5.798
was	6970	0.0105	6.568

The main point of this table is to demonstrate the log-likelihood column. At this point, these numbers won’t seem to have much meaning; however, they provide an overall scale that will be seen, repeatedly, in the analysis below. The range of magnitudes – 4 to 7 – is no accident, and similar ranges will be seen later.

The first ten pseudo-disjuncts in the disjunct-ranking are ".+ " "the- " "the+ " ",+ " ",- " "of+ " "of- " "a- " "and+ " "and- ". The meaning of the plus and minus signs was explained above; but to recap: the disjunct “.+” means that there are many words that expect to be followed by a period (on the right). The disjunct “the-” means that there are many words that want to link to the word “the” on the left. This is grammatically correct: “the” is a determiner, and it is always the dependent of some noun. The disjunct “the+” is grammatical garbage/nonsense: it states that there are many words that want to link to the word “the” on the right. This is never correct for English; determiners always precede the noun that they modify. The rest of the disjunct look reasonable: it’s legitimate to link to commas on both the left and right, and likewise to the preposition “of” and conjunction “and”. The disjunct “a-” is correct, and the insane disjunct “a+” doesn’t appear until the 14th slot.

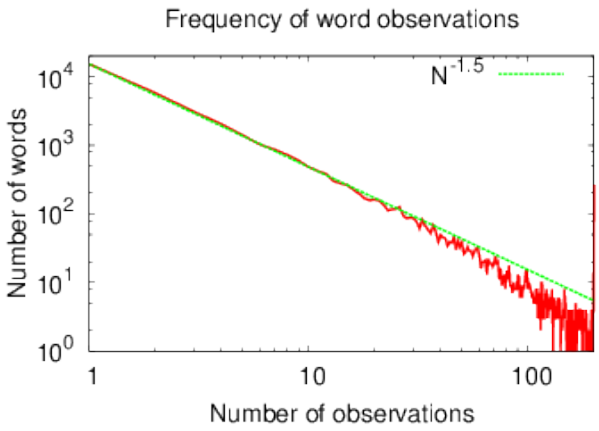
It’s not clear at this point why the insane disjuncts “the+” and “a+” are so highly ranked. The hope is that these appear due to the poor quality of the dataset (as will become more clear, below), rather than a deficiency in the theory. Based on decades-old research on MST parsing, its reasonable to hope that these disjuncts will disappear from a larger, better corpus.

The ranking of connector sets is shown below. It’s a graph of the ranked counts $N(w,d)$.



The top-ten connector-sets are ".:)-" "and: ,-" "in: the+" "(:)+" ",: and+" "the: of-" "of: the+" "): .+" "He: was+". These are hard to read, so, decoded: the first is a parenthesis followed by a period. The second is a comma, followed by "and". The third is the word-pair "in the". All the others are pairs, also. Written in order: "()" (presumably with other text between the parens) ", and" "of the" "of the" ") ." and "He was". Curiously the 1st and 8th give the same pair, as do 2nd and 5th, and the 6th and 7th. This makes sense: if one word-disjunct pair is observed a lot, the converse should be also. This duality might be behind the slope of slightly less than -0.6 in the graph.

It is also interesting to turn this graph "on it's side". So, in this dataset, there were 15270 words observed exactly once (out of a total of 661104 observations of 37413 words). This is quite something: of all the words observed, almost half were seen only once. More than half were seen twice, or less. These are presumably rare typos, foreign words, IPA pronunciation guides: any word that appears only once must be unusual; and yet, there are a lot of them! There are 5790 words that appear twice, 3093 that appear exactly 3 times, *etc.* These counts are graphed below.[3]



[3]graph of binned-word-counts.dat, generated in disjunct-stats.scm

This graph indicates that most (almost all) words were observed less than 100 times. In this dataset, there were only 630 words that were observed more than 100 times, and 267 words that were observed 200 or more times.

Writing N for the number of times that some word was observed, it appears that there are approximately $15270 \times N^{-3/2}$ words observed that many times. In formulas, the size of the set of words $\{w|N(w) = N\}$is given by

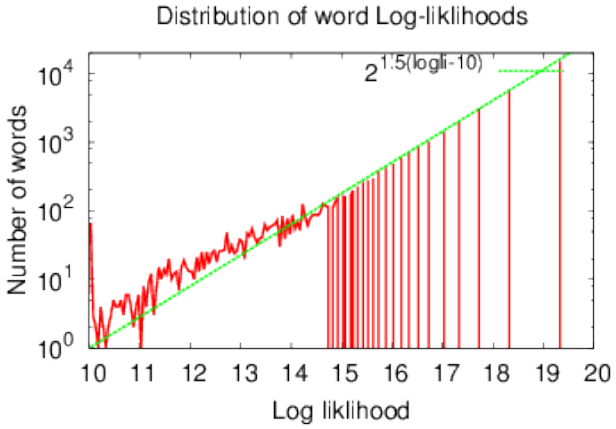
$$|\{w|N(w) = N\}| \sim N^{-3/2}$$

where $\{w|\text{cond}\}$ is a set of words (subject to the condition cond) and $|\{w|\text{cond}\}|$ denotes the size of that set of words.

The next graph belabors the point, and yet it's important.[4] It shows nothing new, but it does show it in a format that will be recur frequently, later. Thus, its worth understanding now. This graph shows exactly the same data as the previous graph: it *is* the same graph, except that the x-axis is now labeled differently, and some of the counts have been binned together. So first: note that $-\log_2 1/661104 = 19.334$ and so this is the location of the first spike on the far-right. Next, $-\log_2 2/661104 = 18.334$ and $-\log_2 3/661104 = 17.750$ are the locations of the second and third spikes: these correspond to words that have been observed 1,2 and 3 times. The height of the spikes are

$$|\{w|N(w) = N\}| \sim 2^{-3/2 \times \log_2 N}$$

as clearly demonstrated by the straight green line.



One peculiar difference between this, and the previous graph is the use of binning. Notice here that the observed, red line bounces *above* the straight green line, while in the previous graph, it was strictly below. If these both show the same data, how can this be? This seems to be a bit of a paradox. The difference can be understood as follows (and is important to keep in mind). This graph was generated by bin-counting. The x-axis was divided into 200 equal-sized bins, and whenever the log-likelihood of

[4] Generated from binned-word-logli.dat

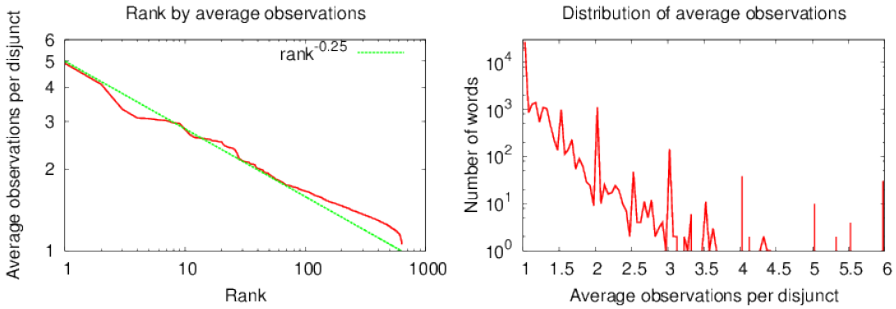
a word landed within a particular bin, the count was accumulated into that bin. Both this and the previous graph are distributions, but they use different measures. The two measures are related by the Jacobian determinant[5], which is given by the derivative of the log. This is a curious effect, to be kept in mind when comparing different graphs.

Ranked average observations per disjunct

A more interesting distribution arises by looking at the average number of observations per disjunct (per word). That is, a single word may have hundreds of disjuncts, observed thousands of times; what is the average number of times that a disjunct is observed? By "average", it is explicitly meant $N(w,*)/|(w,*)|$, the number of observations divided by the support for those observations.

This number gives a hint of how "narrow" the grammatical usage of a word is. If the average is high, it means that the word just does not have very many disjuncts on it; the few that it does have are observed a lot. Recall that these disjuncts (pseudo-disjuncts) connect to individual words, and not to word-classes. Thus, if a disjunct is seen a lot, it probably connects to another word, forming a high-MI pair. This can be explicitly seen in the example further below.

A graph of the ranked average number of observations, per disjunct, per word, is shown on the left, below.[6] Words with fewer than 100 observations have been excluded, so as to minimize the spurious noise. Note that although the line is straight in this log-log graph, the Zipf exponent is approximately -1/4. The graph on the right expresses an alternate view of the same idea: this one shows a bin-count of all of the words. Reaffirming the graph on the left, it indicates that almost all words have an average disjunct observation count of less than three. Both of these graphs emphasize just how thin and weak this dataset is: to be certain of a disjunct, it sure would be nice to observe it, in actual use, more than half-a-dozen times!



The spike on the far right on the graph on the right suggests that there are about 30 words that have an average number of observations of greater than 6! Since they are *not* showing up in the graph on the left, this shows that the underlying word have fewer than 100 observations.

[5] https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

[6] Computed with the sorted-avg list in disjunct-stats.scm

The first ten on the ranked list are "According" "Gestalt" "Cao" "Berger" "2015" "It" "U.S" "United" "However" "y". Unsurprisingly, most of these are all capitalized, suggesting that they are either part of a proper name, or that perhaps they begin a sentence.

For example, the appearance of "United" is almost surely due to the high-MI pair "United States", and so this probably dominates the disjunct count. This is confirmed by manual inspection[7], per table below.

disjunct	number of observations
the- States+	108
in- the- States+	41
States+	35
the- Kingdom+	16
the- the- States+	14

Only the first four disjuncts seem to be grammatically acceptable. In the first, second and fourth usage, "States" seems to be a head word, with "United" being the dependent. In the third case, it would appears that "United" is the head-word, and "States" the dependent.

The fifth most frequent disjunct is grammatical garbage: noise from bad MST parses. There are even more disjuncts, which occur with lower frequency: there are 156 in all. Almost all of these are presumably junk of some kind. We conclude that the signal/noise ratio here is about $10 \log_{10}(108/14)$ or about 9 dB. These first four disjuncts really pull up the average, which is quite low, given this skew: the average number of observations per disjunct for "United" is 2.96.

The corresponding table for "It" is

disjunct	number of observations
is+	216
was+	197
has+	55
was+ was+	12
is+ was+	10

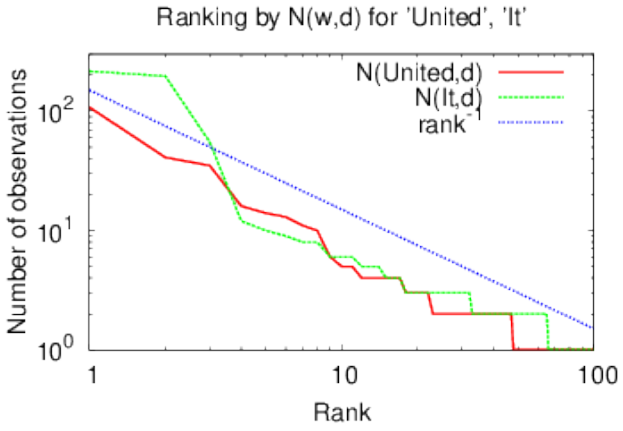
This also is much as expected: "It" is a sentence opener, and the three sentence types it opens are "It is", "It was" and "It has". The fourth and fifth disjuncts are grammatical garbage. The average number of disjunct observations is 3.03, almost identical to that for "United".

The skewness appears to be very sharp. This suggests that we should not waste time looking at mean-square variations in the average, although we'll do this anyway. But first, its worth graphing the skewness directly. Again, this is done on a log-log graph, in a Zipfian way.

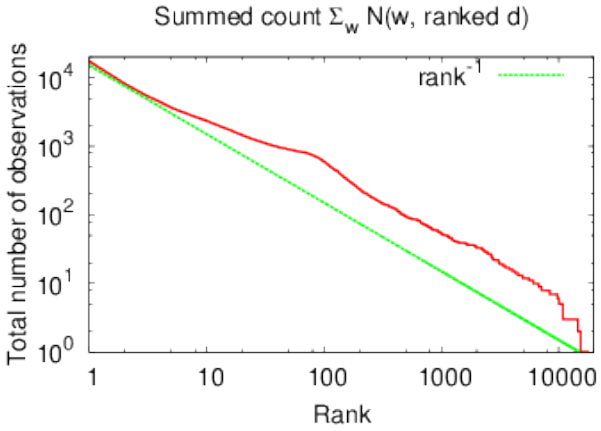
[7] View disjuncts by saying (filter (lambda (cset) (< 10 (get-count cset))) (cog-incoming-by-type (Word "foo") 'LgWordCset)) where 10 is the minimum number of counts.

Disjunct count distribution

The graph below shows the distribution of the disjunct observations on the two words "United" and "It". Indeed, it looks Zipfian; since we know that all but the first three or four disjuncts are noise, this graph illustrates "pink noise" or "1/f noise".[8]



Can one get a smoother distribution by summing together these two graphs? Sure... and one can sum together not just these two words, but all words (that have been observed at least 100 times). That graph is shown below.[9]



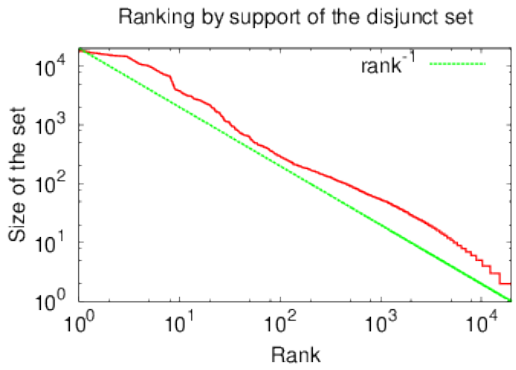
Its kind of a strange graph. Yes, the x-axis of this graph does imply that there are dozens, if not hundreds of words with more than a thousand disjuncts on them, and maybe half-a-dozen that have more than ten-thousand (unique, different) disjuncts on them! Exactly what does this mean? This is covered in the next section.

[8] Computed with the dj-united and dj-it arrays indisjunct-stats.scm
[9] Computed with the accum-dj-all function in disjunct-stats.scm

Disjunct Support Distribution

Is it possible that some words have a large number of disjuncts on them? Yes, it is. For example, the period was observed to have 18200 unique, different disjuncts associated with it. The word "the" has 15420 unique, different disjuncts, the comma has 14510. Rounding out this list are "of" "and" "in" "a" "to" "was". Its not clear what fraction of these disjuncts are grammatically valid, and what fraction are junk.

The graph below shows the distribution of the size of the support: the ranking of $|(w,*)|$. Again, the graph appears to be approximately Zipfian.[10]



Terminology: the "support" of a vector is the number of basis elements that have a non-zero coefficient. This is the set $(w,*)$ defined earlier. Equivalently, this is the size of the set of disjuncts associated with a word, when counted *without* multiplicity.

Ranked Euclidean length (RMS Size)

A different distribution arises by looking at the ranked RMS sizes of the disjunct sets[11]. Here, the RMS size[12] is computed by taking the root-mean-square of the counts on each disjunct in the set, that is, by computing $\sqrt{\sum_d N(w,d)^2}$ for each word w and then ranking. Interpreting d as a basis element of a vector space, this can be recognized as the Euclidean length of the count-vector.

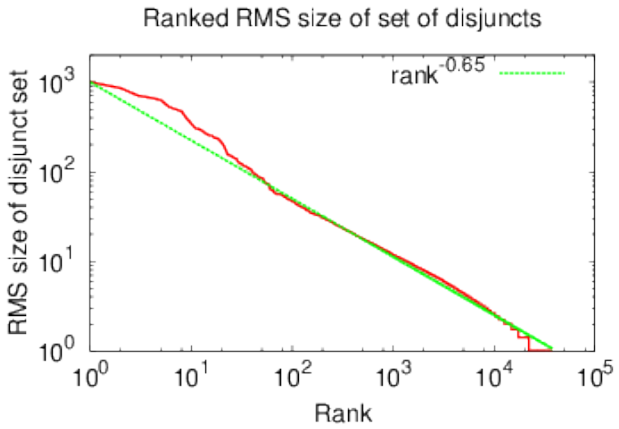
The RMS size of the set is thus larger not only when more disjuncts have been observed, but also when most of the observations are made of only a small handful of disjuncts. That is, the RMS size should be relatively larger, if the word is less grammatically flexible. So for example, prepositions tend to be very flexible; adjectives, not so much. Thus, we expect adjectives to appear higher-up on this ranking list, than

[10] Generated by sorted-support in disjunct-stats.scm

[11] Obtained by running (print-ts-rank sorted-lengths outport) from the disjunct-stats.scm file on the en_pairs_sim database.

[12] The word "length" can be used to describe the root-mean-square size of the set of disjuncts associated with a word. That is, each element of the set is a disjunct, and that disjunct has a count, the number of times it has been observed. The root-mean-square of these counts can be taken as the set-size. But this set can also be interpreted as a vector, and so the RMS size is the same thing as the Euclidean length of the vector. Thus, the word "length" is sometimes used for the RMS size; they're the same thing.

the observation-based list. And this might be true, relatively, but certainly not true absolutely.



The first ten words of the RMS-size-list are: "." "and" "," "in" "the" ")" "(" "of" "as" "to". Not that interesting: these are all words that were observed a lot in the text. The RMS size is dominated by the total number of observations of a word in text. In and of itself, its insufficient to indicate how "concentrated" the disjuncts are, how grammatically narrow a word is. For this, some other quantity is needed.

Note that the graph above is Zipfian, but the slope is approximately -0.65. The green line is an "eyeball" fit. Why the slope is approximately 2/3rds of the canonical slope is not clear.

Mean-square to size ratio

More interesting is the ratio of mean-square size to the total size. In formulas, by ranking according to

$$\frac{\sqrt{\sum_d N^2(w,d)}}{N(w,*)} = \frac{\sqrt{\sum_d p^2(w,d)}}{p(w,*)}$$

This seems like the interesting ratio, because the Zipf exponent of -0.65 would be doubled, when working with mean-square sizes, thus making the two rankings comparable.

Words high in this score will be words that have relatively few disjuncts on them, or at least, few that matter much, that rise above the level of noise. The first ten on this list are (after excluding all words with fewer than 100 observations): "It" "According" "He" ")" "and" "(" "United" "as" "." "well". Note that most are capitalized: that means that these words appeared at the beginning of sentences; the way in which the can link to what follows is fairly constrained, and thus it is no surprise that these have only a few disjuncts on them.

Excluding the capitalized words, and punctuation, in this list what remains is quite surprising. It is shown in the table below.[13]

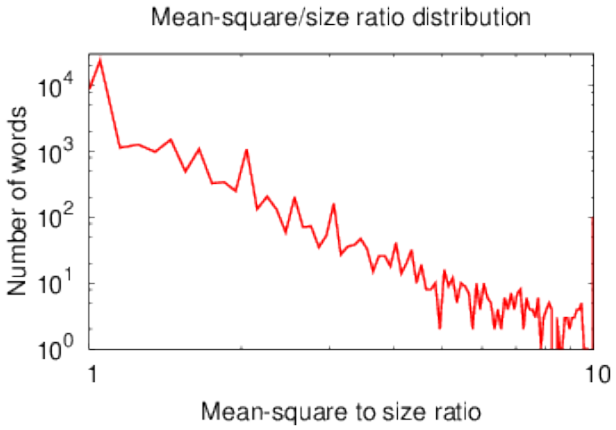
[13] Extracted from ranked-sqlen-norm.dat

rank	score	word
5	41.4	and
8	31.6	as
10	30.1	well
11	29.9	in
18	18.8	first
21	17.2	is
22	16.7	also
23	15.3	can
24	15.3	on
25	14.9	at

rank	score	word
26	14.4	been
31	13.2	but
32	13.1	part
34	12.8	with
35	12.8	old
36	12.7	same
38	11.9	one
39	11.6	which
43	10.8	not
49	10.2	the

It is surprising because, grammatically, we expect most of these words to have a large number of varied disjuncts attached to them. We expect them to be diffuse, not sharp: we expect that these would have a large number of observations smeared over a large variety of disjuncts, instead of having their weight concentrated in only one or two disjuncts, the way that "United" was, above. So what is going on, here?

The distribution is unusual; it is shown below. Note that the x-axis is drawn with a log-scale. It appears to be linear, with an exponent of -4.



Are there other interesting measures? One could contemplate the ratio of the mean-square size to the support $\sum_d N(w,d)^2 / |(w,*)|$. Another possibility would be this, minus the average-squared, which would give the second moment, aka, the mean-square deviation from the average, specifically

$$\frac{\sum_d N(w,d)^2}{|(w,*)|} - \left[\frac{N(w,*)}{|(w,*)|}\right]^2 = \frac{1}{|(w,*)|}\sum_d \left[N(w,d) - \frac{N(w,*)}{|(w,*)|}\right]^2$$

Neither of these variations seem promising; they seem to offer up more of the same, at least on this dataset. A larger and more refined dataset might reveal otherwise.

Mutual information

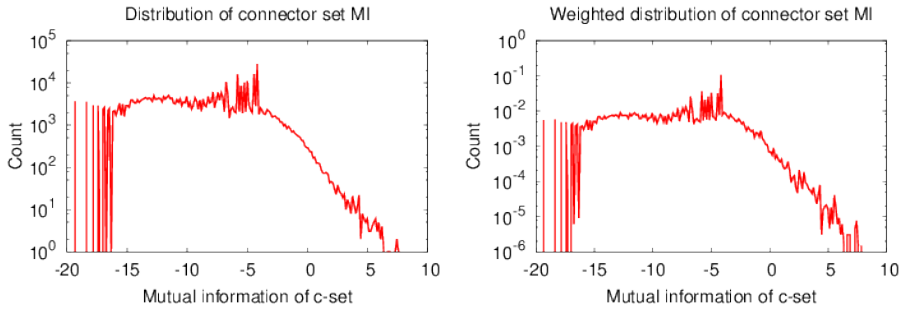
The concept of the "fractional mutual information" for a pair is interesting to explore. Define this as

$$MI_{pair}(w,d) = -\log_2 \frac{p(w,d)}{p(w,*)p(*,d)} \tag{1}$$

This is "fractional" in the sense that the total MI for the set of all pairs can be written as

$$MI_{cset} = \sum_{w,d} p(w,d) MI_{pair}(w,d)$$

Fractional MI is interesting because it usually has a reasonably nice distribution. For this particular dataset, it ranges from about -20 to +8. The distribution is shown in the graphs below.[14]



These graphs are generated by computing the value for $MI_{pair}(w,d)$ for each of the 446204 (w,d) pairs (aka 'connector sets'), and approximating it's distribution by bin counting. In each graph above, there are 200 bins, each of width of about 30/200, and each pair is assigned to one of the binds, according to it's MI value. The graph on the left then shows how many pairs there are in each bin. The graph on the left is similar, but not the same: it sums the frequencies for all the pairs in each bin. In formulas: the graph on the left shows the value of

$$\text{sizeof } \left\{(w,d) | MI_{pair}(w,d) \text{ is in bin}\right\}$$

while the graph on the right shows

$$\sum_{MI_{pair}(w,d) \text{ is in bin}} p(w,d)$$

where 'x is in bin' simply means $lo \le x < hi$ with the bin being the interval $[lo,hi)$.

Both of these graphs show "combs" in the left side. These combs are exactly the same combs as noted in the last figure in section Connector-set distribution on page 7. The combs are due to the large number of words that have been observed only a small handful of times. In essence, the combs attest that the bulk of the high-MI pairs have been observed more than just a few times; *i.e.* the high MI values are meaningful.

[14] These are graphed by binned-cset-mi and weighted-cset-mi in the disjunct-stats.scm file.

Mutual Information of words

The mutual information of a single word can be defined by summing the (fractional) mutual information between a word, and all of it's disjuncts:

$$MI_{word}(w) = \frac{1}{p(w)} \sum_d p(w,d) MI_{pair}(w,d)$$

This is also written in the "fractional" style, so that, again, the total MI of the entire dataset can be written as

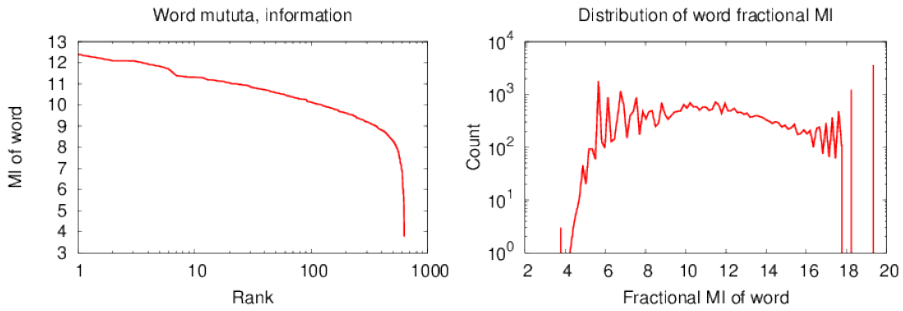
$$MI_{cset} = \sum_w p(w) MI_{word}(w)$$

That is, $MI_{word}(w)$ is the fractional contribution of the word to the total MI. The fractional MI is very convenient for comparing different words, since it factors out the frequency of how often a word is observed: the MI of two words with two very different frequencies can be directly compared. As can be seen from the graph below, the fractional MI ranges between +3 and +19 for this dataset.

The total MI for the dataset is measured to be 7.985 bits.

The distribution can be visualized in two different ways. The graph on the left, below[15] shows the ranked MI of the 630 words that have been observed more than 100 times in this dataset. Note that it is a semi-log plot; it is NOT Zipfian. Note that the MI seems to decay logarithmically, for a good long ways, and then drops off a cliff.

The graph on the right shows the distribution, for all 37413 words, bin-counted into 100 bins. Unfortunately, this sampling remains small and noisy, so the suggestion of a log-linear decay to both left and right is hidden by rather noisy spikes. The combs on the far right are again the same combs as noted in the last figure in section Connector-set distribution on page 7.



What re the disjuncts like at either end of this distribution? The first ten words in the ranking are: "Cao" "Award" "x" "y" "Prime" "we" "per" "Division" "League" "Game". The word "United", examined previously, is 86th in this rank. Lets look at some of these.

The word "Award" has only two disjuncts that were observed more than twice:

[15] This is graphed by sorted-word-mi-hi-p in disjunct-stats.scm

count	disjunct
8	Grammy- for+ Best+
7	Academy- for+ Best+

It is already clear from this one example that the hi-MI words will be those that take part in idioms, “set phrases” or “institutional phrases”, and that the disjunct identifies the words taking part in the setting.

count	disjunct for “Prime”
23	Minister+
5	the- Minister+
3	as- Minister+
3	Governor-General- Minister+

Only one disjunct for “Division” was observed 4 times; those with 3 observations or less suggest that they were extracted from tables listing WWII military battles and equipment.

count	disjunct for “Division”
4	NCAA-

The disjuncts for “League” correspond to word-pairs, and these will clearly be high-MI word-pairs.

count	disjunct for “League”
8	Baseball+
5	Premier-
4	Justice-

This ranking also captures a fair amount of the “garbage” in Wikipedia (which is the source for the dataset): the disjuncts on “x” and “y” indicate that these occur almost exclusively in mathematical formulas of some sort, involving single-letter variables, plus signs and equal signs. The most frequent disjunct on “per” was observed 5 times, and was “100,000+ 100,000+ 100,000+ 100,000+ 100,000+” suggesting that it arose in some parse of a table listing. None of the disjuncts on “we” were observed more than 4 times, and seemed to consist of some sort of grammatical word-salad. This is not surprising: Wikipedia has a severe shortage of pronouns; style guidelines prohibit their use.

At the other extreme, the ten words with the lowest MI are: "The" "a" "to" "in" "of" "and" "the" "," ".". These are already familiar from previous rankings: they occur with very high frequency; the disjunct lists on them will be lengthly variable, diffuse.

Mutual information of disjuncts

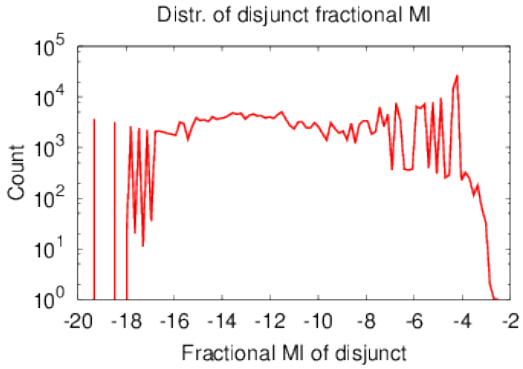
Symmetrically, one also has the mutual information of a disjunct, in comparison to all of the words it connects to:

$$MI_{disjunct}(d) = \frac{1}{p(*,d)} \sum_w p(w,d) MI_{pair}(w,d)$$

Again, this is presented in the "fractional" style, so that the total MI of the entire dataset can be written as before:

$$MI = \sum_d p(*,d) MI_{disjunct}(d)$$

The distribution is shown below.[16] The combs on the far left are again the same combs as noted in the last figure in section Connector-set distribution on page 7.



So, at first, this looks much like the word-MI graph, until one realizes its really almost a mirror-image. Also very notable is that the MI is negative, in all cases! What does this mean? In this case, it means that entropy is dominating over any "attractive" effect: basically, that the various disjuncts are getting used for a very large and widespread set of words, and that any one disjunct is not favoring any particular word or word-set; indeed, any given disjunct is mostly dis-favoring any word or small set of words, and is instead appearing with just about all words. This is explored a bit more in the next section.

Fractional Entropy

There is a simpler variant than the mutual information, that is also worth understanding: the fractional contribution to the total entropy. This is given by the sum

$$H_{word}(w) = -\frac{1}{p(w)} \sum_d p(w,d) \log_2 p(w,d)$$

[16] Use binned-dj-mi to get this.

This is written in the "fractional" style, so that the total entropy of the entire dataset can be written as

$$H_{cset} = \sum_w p(w) H_{word}(w)$$

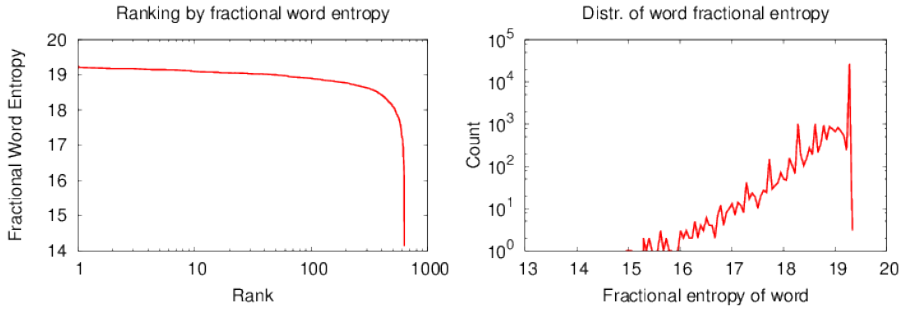
Analogously, one also has the fractional contribution of the disjuncts:

$$H_{disjunct}(d) = -\frac{1}{p(*,d)} \sum_w p(w,d) \log_2 p(w,d)$$

where, again, one has that

$$H_{cset} = \sum_d p(d) H_{disjunct}(d)$$

The ranked fractional entropy is shown in the left graph below.[17] It only shows those words that have been observed 100 times, or more. It resembles the graph for the ranked fractional MI, above. The graph on the right shows the distribution of the entropy, for all of the words. This affirms (or explains?) the sharp knee in the graph on the left: the knee occurs because almost all words have a large disjunct-entropy. This is explained below.



The top-ranked words (which have been observed 100 times or more) are these:

Entropy	Word
19.22	possible
19.18	education
19.18	lost
19.15	days
19.15	children
19.14	players
19.14	available
19.12	every
19.12	remained
19.10	above

[17] Generated from sorted-word-ent in disjunct-stats.scm.

This is not a list that has been exposed with other statistics. These words are more "interesting" than any of the previous lists, which seemed to be filled with quirks associated with the dataset. By contrast, these are "normal" words: mostly nouns, some modifiers, a verb, a preposition. Why these?

The answer lies in looking at the disjunct set, shown in the table below. All of these words have a large support, but were observed very infrequently. For example, "possible" has only one disjunct that was observed 3 times, 4 that were observed twice, and 101 that were observed only once. A quick examination shows that many, maybe most, are grammatically reasonable, for example: (investigate- routes+) (is- also- for+) (is- only- to+) (made- through+) (many- as-) and so on. But all of these were observed precisely once. The table below illustrates this.

Entropy	Word	$\|(w,*)\|_2$	$\|(w,*)\|_1$	$\|(w,*)\|_0$
19.22	possible	1	5	106
19.18	education	1	7	100
19.18	lost	1	7	100
19.15	days	2	7	98
19.15	children	1	9	101

A bit of notation. Recall the definition for the set that supports a word:

$$(w,*) = \{(w,d)|N(w,d) > 0\}$$

together with the notation $|(w,*)|$ to indicate the size of this set. Extend this notation as

$$|(w,*)|_k = \text{sizeof } \{(w,d)|N(w,d) > k\}$$

That is, $|(w,*)|_k$ is the size of the support for when the disjuncts on word w have been seen more than k times.

From this table, it is now clear why "large entropy" can be intuitively understood to mean "many possibilities". Each of these words was seen in a very broad setting of possibilities: in a sense, the broadest possible. Each of these words was observed with a vary large set of different disjuncts, and this set was as spread-out as possible: the vast majority of disjuncts were observed exactly once.

More terms

Some curious terms show up in relating the fractional mutual information to the fractional entropy. Expanding out the above summations, one obtains

$$MI_{word}(w) = H_{word}(w) + \log_2 p(w,*) + \frac{1}{p(w)} \sum_d p(w,d) \log_2 p(*,d)$$

The last term is bizarre...

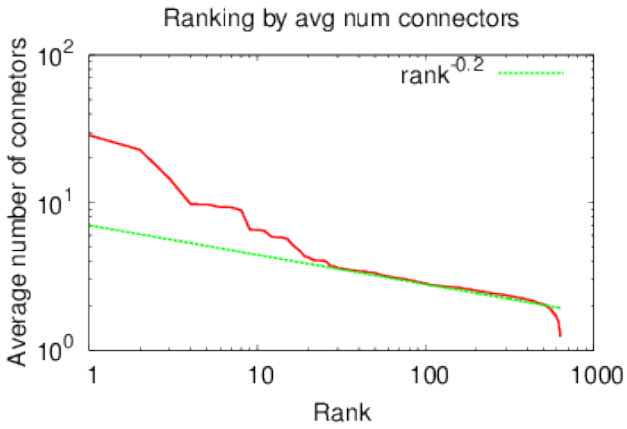
Vertex degrees and hubiness

Vertex degrees can be defined as the average number of connectors per disjunct. In principle, the vertex degree is an excellent indicator of the part of speech. For example, determiners, adjectives and adverbs typically have a degree of one: they have one connector, which is modifying the noun (or verb) that they act on. By contrast, nouns typically have a degree of two: one connector to attach to a verb, another to a modifier, and that's it. Verbs have a degree of three: one connector to a subject, one to a direct object, a third to an indirect object or a modifier. Of course, nouns might have two or more modifiers, or maybe zero modifiers; verbs are also quite variable, but the general concept of vertex degree is appealing. Closely related to this is the idea of "hubiness", which can be defined as the second moment of the degree.

Thus, its worth looking at this. Define the average degree as

$$K(w) = \frac{\sum_{d,c} N(w,d)C(d,c)}{N(w,*)}$$

This is graphed, below, for all words that have at least 100 observations.[18]



This graph is unexpected; in part, having an average number of connectors that exceed 5 or 6 is intuitively surprising. What's happening here? The first ten items in the ranking are: "-" "de" "y" ":" "(" ")" "General" "Department" "x" "Act".

Consider "de". There are 12 observations of the disjunct "Janeiro+". There are 9 observations of the disjunct "la+". There are 51 observations of a disjunct that has 117 connectors on it!! This starts out as "Diego- Francisco- Francisco- Alonso- Carlos- Fernández- Carlos-" and ends with "Figueroa+ (+ (+ y+" suggesting that there were possibly 51 really bad parses of a very long table of Spanish kings, which was mistaken for being a single sentence. Clearly, its junk; its frequently-occurring junk, which suggests that the table was repeatedly included in maybe 51 different Wikipedia pages.

Similarly, "Department" has 18 observations of a disjunct with 41 connectors on it. It starts with "Education- Education- Health- Services- Services- Immigration-" and

[18] Computed with sorted-avg-connectors in disjunct-stats.scm

ends with “Veterans+ of+ Treasury+ Treasury+”, again suggesting a bad parse of a table mistaken for a sentence, and included in 18 different Wikipedia pages.

The list continues in a similar way, for quite a while. The green line suggests that if some 30 or so pathological cases are ignored, the system settles down to a more respectable behavior. Entries 30 through 50 in the rankings are "Bay" "Street" "Island" "of" "century" "right" "Game" "Georgian" "or" "a" ";" "near" "Party" "team" "law" "Australia" "her" "research" "Church" "east" "Government". Notable is a preponderance of capitalized words, suggesting more tables of various sorts, and a complete lack of verbs. A spot-check of words like “team” and “law” shows that the pathological behavior continues. Several conclusions are possible.

One conclusion is that there is a severe shortage of verbs in Wikipedia articles, and this makes sense: its primarily descriptive, rather than active: running, jumping, hitting, putting, mixing, giving, setting are not the kinds of verbs that are required to describe a typical encyclopedia topic.

Another conclusion is that perhaps the number of observations of pairs are insufficient to get deep, reliable MST parsing. Junk links get used because there were not enough appropriate word-pairs seen to give a good-quality MST parse. A related conclusion is that the connector-set dataset is also too thin: The grammatically reasonable connectors are observed not even a few dozen times, barely pushing them out of the noise-floor of onesie-twosie observations of junk.

So: bigger datasets, and an urgent need for non-Wikipedia content. Fiction, and presumably teen fiction should be filled with the kinds of active verbs describing human motions and actions, and should be absent of tables and lists masquerading as sentences.

Hubiness

Similar to the above, hubiness can be defined as the second moment of the connector count:

$$hub(w) = \frac{\sum_{d,c} N(w,d) C^2(d,c)}{N(w,*)} - K^2(w)$$

Given the earlier Zipfian results on the average degree $K(w)$, it should be no surprise that a ranked listing of words by hubiness is very nearly identical to the listing for average degree. This is, after all, what the scale-free nature of the Zipfian distribution really means. Not only is the ranking nearly the same, but one also has the approximate equality $hub(w) \approx 2K(w)$ to some ten or twenty percent.

Disjunct Cosine Similarity

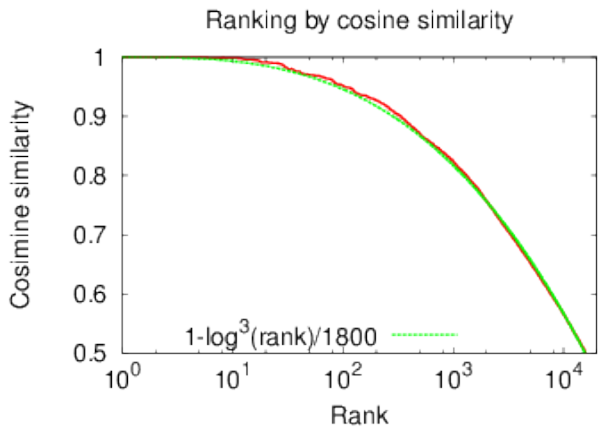
The cosine similarity between two vectors is simply their inner product. In this case, given two words w_1 and w_2, it is given by

$$\text{sim}(w_1, w_2) = \frac{\sum_d N(w_1,d) N(w_2,d)}{\text{len}(w_1)\text{len}(w_2)}$$

where $\text{len}(w)$ is the root-mean-square length (Euclidean length) of the connector-set vector:

$$\text{len}(w) = \sqrt{\sum_d N^2(w,d)}$$

The current dataset being analyzed contains 1985 words whose length is greater than 8; the ranking-by-length was already shown up above, in a previous graph. The similarity between all pairs of these was computed; this resulted in 15808 pairs with a cosine similarity of greater than 0.5. These can be sorted and ranked.[19] They are shown below.

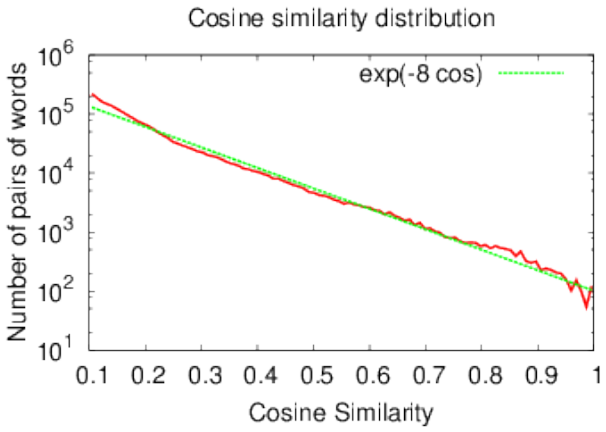


Well, that's new! The similarity ranking is very well fit by a cubic! In the above, the green line is given by $1 - \log^3(rank)/1800$ and it fits the tailing end of the distribution so well, that the red data line is almost completely hidden under the green line!

The graph below shows the distribution of cosine similarity.[20] There are 5544 words for which $4 < \text{len}(w)$. This shows the distribution of the 2.17 million word-pairs formed from these words, and for which $0.1 \leq \text{sim}(w_1, w_2)$. The eyeballed fit is for $\exp(-8\text{sim})$.

[19] See the good-sims and ranked-sims in disjunct-stats.scm file.

[20] Produced by binned-good-sims.



The top-ten similarity pairs are: 'Poir .. Perls' 'Poir .. Sevan' 'Sevan .. Perls' 'Sevan .. Ruins' 'Ruins .. Perls' 'Poir .. Ruins' 'Gongora .. Perls' 'Poir .. Gongora' 'Gongora .. Sevan' 'Gongora .. Ruins' 'Poir .. Gag'. The first three all have a cosine of 1.0, and the rest in the 0.999 range. The three words "Poir" "Perls" and "Sevan" all have exactly one disjunct, containing exactly one connector, linking them to a period, presumably at the end of a sentence. Its observed a few dozen times, and that's it. "Gongora" has three disjuncts; (.+) is observed 56 times, (,+) is observed twice and (,+ .+) is observed twice. So, yes, all these words have been discovered to behave in a grammatically similar fashion, however, this behavior seems quite boring: links to a period.

The next 300 entries are all links between capitalized words, most of them to the words above. After that, the list is just barely slightly more varied, with e.g. "tonight" showing up occasionally, presumably because there are many sentences that end with the word "tonight", followed by a period. There are very few links to lower-case words, but the ones that show up are.. interesting. The table below gives some hand-picked examples.

rank	cosine	word-pair
504	0.868	'in .. from'
633	0.855	'in .. at'
637	0.854	'In .. By'
643	0.854	'in .. After'
740	0.846	'canal .. swamp'
754	0.843	'at .. At'
757	0.843	'in .. In
818	0.838	'in .. on'
840	0.836	'creation .. existence'
878	0.833	'e.g .. i.e'
965	0.826	'and .. but'
969	0.826	'in .. With'
971	0.826	'in .. inside'

This table is surprising in several different ways. First, that the first entry in which two words are both not capitalized does not occur until about 3% into the list. Next, that this entry, and most of the others, involve prepositions! Not adjectives, adverbs, verbs; although we noted already that the source text is verb-poor. It is also pleasantly surprising that all of these equivalences appear to be correct, including the equivalence of upper-case prepositions that start sentences, with lower-case ones that appear mid-sentence. The identification of e.g. with i.e. seems like a bonus, as does the 'and .. but' equivalence.

Lets look at this gift-horse in the mouth. The equivalence 'canal .. swamp' is rather pathetic. So, 'canal' was observed with 68 different, unique disjuncts, but most have a count of only 1 or 2. The top-ranked disjunct on 'canal' is (the-) which is seen 20 times, followed by (the- .+) 8 times. Similarly, 'swamp' has 18 distinct disjuncts attached, most with counts of 1 or 2; the top-ranked disjunct is (the-) with 14 counts, and (the- .+) 3 times. So, yes, these are both nouns, evidenced by the connection to the word "the", but the evidence is otherwise very weak.

Likewise, 'creation' gets the disjunct (the- of+) 18 times, and no other disjuncts observed more than twice. 'existence' gets the same disjunct 10 times, and no other is observed more than 3 times. In total, 'creation' had 22 disjuncts on it, while 'exstence' had 26, with almost all having counts of 1 or 2. Again, the evidence is thin, with one notable point: 'creation' is not at all like 'canal', in that 'canal' has zero counts of (the-of+) while 'creation' has only one count of (the-). As nouns, they are actually quite different, even based on the thin evidence available.

The equivalence of 'in .. from' is much more interesting. The preposition 'in' has a total of 8338 different unique disjuncts – this is huge, compared to the nouns above. Likewise, 'from' has 2185 different disjuncts observed. The top-ranked disjuncts for each are shown in the tables below.

count	'in' disjunct
627	the+
75	a+
61	which+
49	the+ the+
36	used-
33	his+
32	order+

count	'from' disjunct
155	the+
20	:+ :+ :+ .. 0+ 0+ 0+
18	a+
9	derived-
9	until+
9	degree-
8	his+

The similarity of these two are driven primarily by the word-pairs "in the" and "from the", followed by "in a" and "from a". Notable also is the common appearance of the pronoun "his", with "in his" and "from his". The second disjunct on 'from' appears to be some sort of garbage, presumably due to the parse of a list of numeric values of some kind.

Just as the case with nouns, the evidence for the similarity for these two prepositions seems quite thin. The surprise is that, despite this, the identification does seem correct.

At any rate, the dataset seems a bit thin, but we already knew this; its based on a fairly small sample.

Pair Cosine Similarity

To correctly gauge the disjunct-based cosine similarity, it should be contrasted against the simpler pair-based cosine similarity. This is given by

$$\text{sim}_{pair}(w_1, w_2) = \frac{\sum_w N_{pair}(w_1, w) N_{pair}(w, w_2)}{\text{len}(w_1)\text{len}(w_2)}$$

where $\text{len}(w)$ is the root-mean-square length (Euclidean length) of the pair vector:

$$\text{len}(w) = \sqrt{\sum_v N_{pair}(w, v) N_{pair}(v, w)}$$

and $N_{pair}(w, v)$ is the count of having observed the ordered word-pair (w, v). Equivalently, writing the normalized frequency of observing a word pair as $p(w, v) = N(w, v)/N(*, *)$, this similarity can be written in the form

$$\text{sim}_{pair}(w_1, w_2) = \frac{\sum_w p(w_1, w) p(w, w_2)}{\sqrt{\sum_v p(w_1, v) p(v, w_1)} \sqrt{\sum_v p(w_2, v) p(v, w_2)}}$$

Note that this similarity measure is NOT symmetric: $\text{sim}(w_1, w_2) \neq \text{sim}(w_2, w_1)$. This is because it's built out of a manifestly non-symmetric count: $p(w, v) \neq p(v, w)$ and should really be written as $p(w, v) = p(R; w, v)$ with the relation R encompassing all of the constraints of pair-wise word relationships (including, for example, that the pair might have been extracted from a random planar tree parse). Of course, one could construct a symmetrized similarity measure.

The point here is that this measure treats words as similar with they link, pair-wise, to the same kinds of words, with the same kinds of frequencies. This is not unlike the similarity that the disjunct-cosine is measuring, except that the disjunct carries additional grammatical information with it: it captures more complex relationships between the words in a sentence.

Cosine Information

The cosine similarity was defined as

$$\text{sim}(w_1, w_2) = \frac{\sum_d N(w_1, d) N(w_2, d)}{\sqrt{\sum_d N^2(w_1, d)} \sqrt{\sum_d N^2(w_2, d)}}$$

which, after dividing by $N(*, *)$ so that $p(w, d) = N(w, d)/N(*, *)$, gives the equivalent expression

$$\text{sim}(w_1, w_2) = \frac{\sum_d p(w_1, d) p(w_2, d)}{\sqrt{\sum_d p^2(w_1, d)} \sqrt{\sum_d p^2(w_2, d)}}$$

Comparing this to the expression for mutual information suggests that using the vector support, instead of the vector length, could be interesting. In particular, these might be interesting:

$$\text{com}(w_1, w_2) = -\log_2 \frac{\sum_d p(w_1, d) p(w_2, d)}{p(w_1, *) p(w_2, *)}$$

and

$$\text{mim}(w,d) = -\log_2 \frac{p(w,d)}{\sum_{w,d} p^2(w,d)}$$

I don't know what to call these; the first seems to be some kind of "cosine information", the second, some sort of "mutual length" device.

Quality Cosine

More possibilities exist. The motivation for using cosine similarity is to find pairs of words that act in a grammatically similar fashion: they are used in the same way, with the same kinds of disjuncts. However, observational counts are subject to the vicissitudes of the input text: perhaps, instead of using vectors where the components are given by the frequency, once could instead use components based on, say, the "quality" of the disjunct itself. A disjunct could be judged as being "high quality" if $mi(w,d) = MI_{pair}(w,d)$ as defined in equation 1 on page 14 is high. This motivates a "quality cosine":

$$\text{qim}(w_1,w_2) = \frac{\sum_d mi(w_1,d)mi(w_2,d)}{\sqrt{\sum_d mi^2(w_1,d)}\sqrt{\sum_d mi^2(w_2,d)}}$$

But what if the high quality is based on a pathetically small number of observations? Perhaps there should be some observational weighting. One can contemplate defining $pi(w,d) = p(w,d)mi(w,d)$ and so the cosine

$$\text{pim}(w_1,w_2) = \frac{\sum_d pi(w_1,d)pi(w_2,d)}{\sqrt{\sum_d pi^2(w_1,d)}\sqrt{\sum_d pi^2(w_2,d)}}$$

The suitability of these different means of judging similarity is not clear.

The End