Chinese and English Pairs

Linas Vepstas

July 8, 2017

Abstract

A short report on word-pair datasets for Chinese and English. Although these are strictly not comparable, because the pairs were formed between hanzi, and not words, for Chinese, the pair statistics are still .. interesting.

Word-Pair datasets

Some summary reports from various different word-pair datasets. These datasets hold word-pairs, parsed using the LG "ANY" link type: i.e. random parse trees.

Size	Pairs	Obs'ns	Obs/pr	Entropy	MI	Dataset
395K x 396K	8.88M	418M	47.0	19.28	-3.02	en_pairs_sim
134K x 135K	5.54M	174M	31.4	17.67	-1.94	en_pairs_rone_mst
185K x 188K	8.95M	321M	35.9	17.77	-1.79	en_pairs_rtwo_mst
428K x 434K	16.4M	639M	38.9	18.27	-1.90	en_pairs_rthree
158K x 159K	5.92M	729M	123	18.45	-2.02	zh_pairs_sone

The legend is as follows:

Size The dimensions of the array. This is the number of unique, distinct words observed occurring on the left-side of a word pair, times the number of words occurring on the right. We expect the dimensions to be approximately equal, as most words will typically occur on both the left and right side of a pair.

Pairs The total number of distinct pairs observed.

Obsn's The total number of observations of these pairs. Most pairs will be observed more than once. Distributions are typically Zipfian, as previous sections point out.

Obs/pr The average number of times each pair was observed.

Entropy The total entropy of these pairs in this dataset, as defined previously: for word-pairs (w_L, w_R) it is $H = -\sum_{w_L,w_R} p(w_L, w_R) \log_2 p(w_L, w_R)$.

MI The total mutual information for the pairs in this dataset, as defined previously:
$MI = -\sum_{w_L,w_R} p(w_L,w_R)\log_2[p(w_L,w_R)/p(w_L,*)p(*,w_R)]$

Recall that each dataset can be viewed as a sparse matrix, consisting of the number of observations $N(w_L,w_R)$ of a word-pair (w_L,w_R). The probabilities are defined as $p(w_L,w_R) = N(w_L,w_R)/N(*,*)$. The matrices are extremely sparse: of all possible word-pairs, only about one in twenty thousand are observed. The sparsity is the log-base-two of this ratio, and varies in the range of 12 to 15.

The datasets are as below.

en_pairs_sim This contains text parsed from Wikipedia, only. As noted previously, Wikipedia is painfully short of verbs and pronouns. Compared to the Gutenberg datasets below, it is also very rich in foreign words and proper names (product and brand names, geographical place names, biographical mentions and other named entities). Issue: missing connectors the LEFT-WALL.

en_pairs_rone_mst Text from Project Gutenberg “tranche one”, mostly all “famous authors”, popular, well-known 19th century books. Includes six modern sci-fi/fantasy novels from other sources, and some 20th century non-fiction, including a military appraisal of Vietnam.

en_pairs_rtwo_mst Tranche two - Everything from tranche one, plus fan-fiction from http://archiveofourown.org. Most of the selected texts were 10K words or longer. See the ’download.sh’ file for the precise texts. Issues: rone_mst and rtwo_mst are missing connectors the LEFT-WALL. Certain types of punctuation is mishandled.

en_pairs_rthree Tranche three - Everything in tranche two, plus several hundred of the most recently created Project Gutenberg texts, whatever they may be. See the ’download.sh’ file for the precise texts.

zh_pairs_sone A parse of Mandarin Wikipedia, with each individual character (hanzi) treated as a single item (so that, during pair-counting, pairs are formed between items). Non-Chinese characters are grouped into words in the normal way, by splitting according to white-space (and punctuation). Thus, the total dimensions of the dataset are given by the number of observed Chinese characters (hanzi) plus the number of observed non-Chinese words (and punctuation). Issue: missing connectors the LEFT-WALL.

Now, for some commentary, as to the summary stats. For English, as the number of pair observations increase, so do the number of unique, distinct words. The relation even seems to be linear: double the number of pair observations, and the number of different words also increases. This suggests something Zipfian at work. The explosion of words is hypothesized to be given names, although these datasets all fail to split hyphenated words, and so some may be due to that. The point is that the average observations per pair increases with difficulty, and the entropy and MI does not budge at all.

Comparing the English _sim dataset to the _rone, _rtwo and _rthree datasets does provide some contrast: The _sim dataset, built from Wikipedia, is distinctly different from the Gutenberg datasets. Certinly, the prose style in the two datasets is quite different, with Wikipedia consisting of statements of facts ("is", "has" relational statements) concerning a broad range of named entities, whereas the Gutenberg texts are primarily narrative adventures ("did", "went" activity statements) involving fictional personages.

Comparing English to Chinese is very interesting. The Chinese dataset has three times, almost four times more observations per pair; equivalently about 3-4 fewer "words". This is partly due to the fixed number of ideograms in the language. Remarkably, the entropy and MI are untouched. This suggests that the entropy and MI are capturing something about the human nature of language use, as opposed to something descriptive of the language itself. However, a lot more data would be needed to see if this is really true.

There's something else interesting going on, shown in the table below.

Size		Support		Count		Length		Dataset
L	R	L	R	L	R	L	R	Name
158K	159K	6819	6411	548	487	41.7	37.7	zh_pairs_sone
428K	434K	45.6K	45.1K	208	187	22.9	19.4	en_pairs_rthree
185K	188K	24.7K	23.8K	199	173	21.5	17.8	en_pairs_rtwo_mst
134K	135K	17.4K	17.4K	143	129	16.6	14.0	en_pairs_rone_mst

The columns are as follows:

Size The left and right dimensions, as before. Viz, the number of unique, distinctly different words observed on the left and the right side of a pair. Viewed as a matrix, this is the number of columns and rows in the matrix.

Support The support is the average number of word-pairs that a word participates in (on the left, or on the right). Viewed as a matrix, this is the average number of non-zero entries in each row or column. Viewed as (row or column) vectors, this is the "support" of a (row or column) vector. Mathematically, this is the l_0 norm of each vector: $|(w_L,*)| = \sum_{w_R}(0 < N(w_L,w_R))$ and likewise $|(*,w_R)| = \sum_{w_L}(0 < N(w_L,w_R))$.

Count The count is the average number of observations that a word-pair was observed, for a given word. Viewed as a matrix, this is the average value of each non-zero entry (averaged over rows, or columns). Viewed as vectors, this is the l_1 norm divided by the l_0 norm. The l_1 norm is just the wild-card counts $N(w_L,*)$ and $N(*,w_R)$, where as always, the wild-card counts are defined as $N(w_L,*) = \sum_{w_R} N(w_L,w_R)$. The count shown in the table is then the average count: $N(w_L,*)/|(w_L,*)|$ for the rows, and likewise for the columns.

Length The length is the average length of the row and column vectors. This is the l_2 norm divided by the l_0 norm. The l_2 norm is just the standard concept of the length of a vector in Euclidean space. Here, $L(w_L,*) = \sqrt{\sum_{w_R} N^2(w_L,w_R)}$, and likewise $L(*,w_R) = \sqrt{\sum_{w_L} N^2(w_L,w_R)}$. The length is interesting, because it "penalizes" word-pairs with only a small number of counts. The act of squaring the count has the effect of giving much higher "confidence" to large observation counts: a word-pair observed twice as often is given four times the credit. The length shown in this table is the "average" length: it is $L(w_L,*)/\left|(w_L,*)\right|$ for the rows, and likewise for the columns.

So here's what is so interesting in this table: the support, for Chinese, is outrageously different than it is for English. For a given item (hanzi, for Chinese, word, for English), the Chinese hanzi participates in three to four fewer item-pairs! Since pairs are formed on a sentence-by-sentence basis, this means that the variety of different hanzi that can occur in a single sentence is much more constrained, much more strongly correlated. Now, perhaps this comparison is not quite valid: because we are not comparing words to words, but rather English words to Chinese "morphemes" (in the sense that Chinese words are typically composed of 1, 2 or 3 hanzi). Still, its interesting and surprising. This has knock-on effects: the observational counts are much higher, as are the average lengths. It would be interesting to repeat the previously given analysis of the various distributions, and see how they differ.

Disjuncts

Next, datasets that hold disjuncts.

Size	Csets	Obs'ns	Ob/cs	Entropy	H_{left}	H_{right}	MI	Notes
37K x 291K	446K	661K	1.48	18.30	16.00	10.28	-7.98	en_pairs_sim
131K x 2.24M	4.28M	9.33M	2.18	20.72	14.93	9.94	-4.15	en_pairs_rone_mst
179K x 3.98M	7.50M	17.0M	2.27	21.15	15.11	9.95	-3.91	en_pairs_rtwo_mst
								en_pairs_rthree_mst

An updated legend for the columns:

Size The dimensions of the array. The left dimension counts words, the right dimension counts the number of unique, distinct pseudo-disjuncts.

Left-Right The left and right entropies, as defined previously. Note that $MI = H - H_{left} - H_{right}$ holds, by definition. Not given for the word-pairs table, because these two are nearly equal, and are half the difference between the entropy and the MI.

The End.