Good Bigrams

Christer Johansson
Dept. of Linguistics at Lund University
Helgonabacken 12
223 62 Lund, Sweden
email: Christer.Johansson@ling.lu.se

Abstract

A desired property of a measure of connective strength in bigrams is that the measure should be insensitive to corpus size. This paper investigates the stability of three different measures over text genres and expansion of the corpus. The measures are (1) the commonly used mutual information, (2) the difference in mutual information, and (3) raw occurrence. Mutual information is further compared to using knowledge about genres to remove overlap between genres. This last approach considers the difference between two products of the same process (human text-generation) constrained by different genres. The cancellation of overlap seems to provide the most specific word pairs for each genre.

1 Introduction

Statistical methods have been used to find cohesion between local items of language (such as phonemes, morphemes, or words). Early work (Stolz, 1965; Zellig, 1955) was inspired by the advances in information science (Shannon, 1951; Shannon & Weaver, 1963). The research benefited from the possibility to store huge amounts of information in computer systems, and the optimism could be overwhelming when the problems were simplified and thought mostly restricted by the size of the corpus. In this paper the stability of some bigram measures will be investigated. Bigrams are items (i.e. word forms) that occur frequently together in a specific order. The meanings of bigrams are not discussed since there is no meaning outside of a context. Co-occurrence is still interesting because bigrams occur non-randomly, sometimes to such an extent that we discern some structure beyond co-occurrence. The reason why it should be so is probably that part of the use of words is reflected by the company that words keep.

Researchers (Church & Hanks, 1990; Kita & al., 1994, inter al.) have noted that mutual information tends to be insensitive to high frequency patterns, and unstable for low frequency patterns. Johansson (1994) compared another measure, the difference in mutual information ($\Delta\mu$), of collocational strength with mutual information (μ). That measure ranked high frequency bigrams higher than other bigrams if the order was consistent, whereas mutual information tended to pick out combinations of low frequency items. Since low frequency items carry more specific information such bigrams give an illusion of semantic content. It is usually this semantic illusion that we are interested in, but what says that “of the” or ”in a” are worse bigrams than “wooden spades” or “various pretexts”. Johansson proposed the test of finding some of the characters in the children’s story “Alice in Wonderland”, and showed that a ‘new’ measure was to some degree "better" than mutual information. Unfortunately, some of that result was based on the fact that mutual information is very sensitive to low frequency items.

2 Definitions

2.1 Mutual information

In the following p(x) will denote the observed probability as defined by p(x)=F(x)/N where F(x) is the frequency of occurrence of x, and N is the number of observed cases. N is, in the calculations, equal to the corpus size in words. Given this, the mutual information ratio (Church & Hanks, 1990; Church & Mercer, 1993; Steier & Belew, 1991) is expressed by **Formula 1**. (Church & Hanks refer to this measure as the *association ratio* for technical reasons).

$$\mu = \log_2\left(\frac{p([w_1, w_2])}{p(w_1)p(w_2)}\right) \approx$$
$$\approx \log_2\left(\frac{N * Occ([w_1, w_2])}{Occ(w_1)Occ(w_2)}\right)$$

Formula 1: The mutual information ratio

The instability of statistical measures seems to be a problem in statistical bigrams. Especially low frequency counts cause instability. To avoid this use the rule of thumb that a bigram must occur more than four times (cf. Church & Hanks, 1990:p.24) to be considered as a candidate for an interesting bigram.

2.2 The difference in mutual information: temporal co-occurrence

A reasonable way of using the temporal ordering in word pairs is to consider the opposite ordering of the word pair as negative evidence against the present order. A reasonable measure would be to use the difference in mutual information between the two orderings, hereafter $\Delta\mu$. The size of the corpus cancels out and $\Delta\mu$ can be calculated by a ratio between frequencies. This is intuitively correct for a comparison between apples and pears, i.e. you can say that apples (w1 w2) occur twice as often as pears (w2 w1) in my fruit bowl (corpus). (p is the probability in the fixed corpus (f/N) which is different from the probability in the language. It is impossible to have a fixed corpus that equals the language since language does not have a fixed number of words or word patterns).

2.2.1 Handling zero negative evidence

In the case that the reversed ordering of a word pair has not been observed in the corpus, the measure becomes undefined. To relieve this the frequency[1] is multiplied by a constant (10), and the frequency of the reversed ordering is set to 1. Subtracting 9 from that value does not add anything to the measure for a single occurrence (log(10-9)=0).

Other ways of handling zero-frequencies are evaluated in (Gale & Church, 1994), e.g. the Good-Turing method. Relative frequencies of non-observed word pairs are hard to estimate. For example, the frequencies of frequencies (X) and frequency (Y) used in the Good-Turing method are linearly dependent in a log-log scale, i.e., there is an infinite frequency of non-observed items (which is another way of saying that we cannot expect the unexpected).

$$\Delta\mu = \begin{cases} \log_2\left(\dfrac{Occ([w_1, w_2])}{Occ([w_2, w_1])}\right) \\ \quad :if \quad Occ([w_2, w_1]) > 0 \\ \log_2(10 * Occ([w_1, w_2]) - 9) \\ \quad :if \quad Occ([w_2, w_1]) = 0 \end{cases}$$

Formula 2: Handling zero frequencies

3 Illustration

The difference between the two measures are perhaps best illustrated with some concrete examples. In a previous paper (Johansson, 1994) "*Alice's adventures in Wonderland*" (AIW) was used as an experimental corpus to compare phrase finding for μ, and a new measure — Δμ. A critique against that corpus is that the corpus is very small. "*Through the Looking Glass*" and "*The Hunting of the Snark*" extend that corpus to about 63 000 words of which 26 831 occurred more than 4 times. With the criterion that an interesting bigram occurs more than 4 times 1970 bigram candidates were found in this larger corpus.

Effect of delta	Effect of mu	bigram
215	1883	cheshire cat
7	34	humpty dumpty
48	202	looking glass
33	136	march hare
7	28	mock turtle
204	1400	red king
7	-9	red queen
29	-519	the dormouse
174	931	white king
160	932	white knight
29	5	white queen
47	190	white rabbit

In the previous table the effect is measured by the number of steps a bigram is moved up compared to a sorted frequency list. The effect of mutual information under these conditions is higher than the proposed measure for finding most characters in AIW, except for some names defined by definite article + noun, and common adjective + noun.

[1] I will use 'frequency' as equivalent to 'occurrence' in the sample corpus.

4 Material

In the rest of this paper, the corpus is the SUSANNE corpus (Sampson, 1994). This corpus consists of an extensively tagged and annotated subset from the Brown Corpus of American English. The corpus is fairly small, but provides information on grammatical roles on the word and phrase level. This makes the SUSANNE corpus suitable for further research.

The SUSANNE corpus is divided into 4 (approximately equally large) genre subcategories:

"A: press reportage
G: belles lettres, biography, memoirs
J: learned (mainly scientific and technical) writing
N: adventure and Western fiction" (Sampson, 1994:p.174)

Each genre has approximately 20,000 unique word pairs[2]. The four genres will be used as one factor in the comparison between different measures. The question is whether the genre interacts with the ability of the different measures to discover bigrams. In category A 439 unique bigrams (occurring more than 4 times) were found, in G 486, in J 598, N 620, and 2573 for the used corpus[3].

5 Method

The highest ranking bigrams according to the measure are sampled at 5 different levels: the 10, 50, 100, 200 and 400 top collocations. Samples are sorted and compared for overlap by the UNIX command 'comm -12 SAMPLE1 SAMPLE2 | wc -l', and the percentage of overlap was calculated from the size of the sample.
Stability of bigrams was tested by three different overlaps. 1) The overlap between samples from genres, and samples for the entire corpus for the same measure. 2) The overlap between different measures at the five different levels for the different genres and the entire corpus. 3) The overlap between different genres.

6 Results

6.1 Mutual Information

The average overlap between genres and the corpus showed that the J sample was much more stabile than the other genres[4]. The J genre would be the genre that information retrieval applications would be most interested in. The ranking of the genres according to the stability of the overlap is: JANG. The highest collocations are most stabile for J, where the other genres show less specificity (i.e. equal or growing percentages as the overlap grows).

10	50	100	200	400	mean	
20	22	30	27.5	21.5	24.2	A
0	6	10	14.5	16.7	9.4	G
60	62	48	36.5	31	47.5	J
10	6	7	15	22	12.0	N

6.2 Delta Mutual Information

Delta mutual information shows little effect of genre, and sample size. Growing sample size predicts less overlap. The ranking of genres is: GANJ. Delta mutual information seems to rank the less specific genres high.

10	50	100	200	400	mean	
70	64	53	47.5	44.2	55.7	A
60	58	54	58.5	51.5	56.4	G
60	54	48	43	39.2	48.8	J
50	52	49	51	45.5	49.5	N

A factorial ANOVA on measure and genre shows that there is a significant effect ($p<0.001$) of measure ($\Delta\mu$ or μ), genre and interaction between measures. F(measure, 1df)=136.2, F(genre, 3df)=9.8, F(measure, genre, 1, 3)=15.4, $p<0.001$. These two measures are significantly different.

6.3 Occurrence

The results for the samples are similar to $\Delta\mu$. The overlap is generally higher for occurrence than $\Delta\mu$, but the ranking of genres is the same: GANJ. An ANOVA on measure ($\Delta\mu$ and occurrence) and genre show less significant effect on measure, and no significant effect of genre, or interaction (these measures behave in the same direction).

[2] (A 21198 unique / 29969 total / 5332 unique words; G 22248 / 31006 / 6048; J 19039 / 29484 / 4676; N 20902 / 31959 / 4876; all 74126 / 122421 / 13458)

[3] The last small part of each genre was excluded from the start for future purposes.

[4] In preliminary investigations the J genre was the least stabile genre for mutual information. This was 'corrected' by the demand that candidate bigrams should occur more than 4 times.

10	50	100	200	400	mean	
60	70	65	60.5	51	61.3	A
60	70	69	65.5	61	65.1	G
70	62	53	48.5	43.5	55.4	J
70	64	57	54.5	54.2	59.9	N

F(measure, 1df) = 11.1 p<0.02, F(genre, 3df) = 2.7, p>0.05, F(measure, genre, 1, 3) = 0.218, p>0.8. Occurrence is significantly more stabile than the other measure, but there is only a small difference of genres (occurrence and Δμ react in a similar way to genre — i.e. on high occurrence).

6.4 Comparison between measures

The overlap between measures is calculated for all combinations of measures. At the higher levels a high overlap can be expected since there is little possibility to fall out (e.g. in A 400 out of 439 is 91% of the sample). The results from this test indicate that the overlap between D (Δμ) and F (occurrence) is significantly and consistently higher than between the other combinations (especially for the entire corpus).

10	50	100	200	400	Genre	Test	mean overlap
0	6	22	44.5	93.2	A(439)	M=D	33.1
0	6	16	37.5	91.0	A	M=F	30.1
90	*64*	*74*	*78.0*	*91.2*	*A*	*D=F*	*79.4*
0	18	23	45.5	86.0	G(486)	M=D	34.5
0	14	20	43.0	82.0	G	M=F	31.8
80	*76*	*78*	*77.5*	*84.0*	*G*	*D=F*	*79.1*
0	8	13	34.0	72.2	J(598)	M=D	25.4
0	4	11	28.5	64.0	J	M=F	21.5
60	*84*	*78*	*72.5*	*75.5*	*J*	*D=F*	*74.0*
0	8	22	33.5	70.5	N(620)	M=D	26.8
0	6	20	28.0	63.7	N	M=F	23.5
40	*68*	*71*	*67.0*	*72.5*	*N*	*D=F*	*63.7*
0	0	1	7.0	15.7	all(2573)	M=D	4.7
0	0	1	4.0	13.0	all	M=F	3.6
40	*54*	*58*	*58.0*	*59.5*	*all*	*D=F*	*53.9*

6.5 Overlap between genres

To estimate the overlap of the genres the number of common bigrams between two genres were found and compared to the size of the smallest genre. The results indicate an average overlap between the genres of 10%.

Overlap of genres (% of smallest genre)				
	A	G	J	N
A	-			
G	11.0	-		
J	9.4	11.0	-	
N	10.0	12.0	7.5	-

6.6 Reduction of the bigrams

The bigrams that are rated high by the measures (especially mutual information) are mixed between two different types of bigrams: **(1)** bigrams with high internal cohesion between low frequency items that may be associated with a specific interpretation (e.g. "carbon tetrachloride" or "cheshire cat"), **(2)** bigrams with high internal cohesion with usually high frequency of both items that may be associated with a "syntactical" interpretation (e.g. "in the").

To separate type 1 from type 2 some information about the overlap of genres might be used. The type 2 bigrams are typically found in most genres, whereas type 1 bigrams are specific to a text. The results above indicate that we can use the genres with least overlap to filter out common bigrams (i.e. A use J, G use J, J use N, N use J).

In the following table the effect of the genre (column 2) is shown by the number of 'surviving' bigrams from the candidate bigrams (column 1). The third column shows the effect of removing the bigrams that occur (more than 4 times) in both directions after common bigrams have been removed (first parenthesis shows actual removed, second shows those that would have been removed (i.e. those bigrams with both orderings in the candidate set). The fourth column shows the effect of removing bigrams that contains words that occur more than 4 times in the rest of the corpus (i.e. in A G N for J) after the bigrams have been formed. The reason for filtering after forming bigrams is that words that are filtered out later work as place holders, and prevent some bigrams to form. The reduction is most notable for removing bigrams that contain common words between genres: genre G and N contain few good candidates of collocations type 1.

Cand.	Genre	Word order filter	Freq. words	
439	216	179 (-63) (-80)	12	A
486	159	119 (-40) (-127)	1	G
598	355	277 (-78) (-131)	37	J
620	395	291 (-104)(-159)	0	N

The following bigrams survived the harshest condition of removing bigrams containing words of other genres. (Genre J, later ordered by mutual information). Some good candidates were (of course) removed, e.g. "black body", "per cent", "united states".

12.2	poynting robertson
11.8	indirect coombs
11.6	burning arcs
11.4	anionic binding
11.1	binding capacity
11.0	starting buffer
10.7	antenna beam
10.6	wave lengths
10.3	wave length
10.1	multiple secant
10.0	carbon tetrachloride
9.9	bronchial arteries
9.9	heat transfer
9.9	ideal gas
9.8	agglutinin activity
9.5	hydrogen atoms
9.4	multiple secants
9.3	antibody activity
9.1	particle size
9.1	pulmonary vein
8.9	active agent
8.9	bronchial artery
8.9	liquid phase
8.8	pulmonary artery
8.6	anode holder
8.3	solar radiation
8.2	reaction tubes
8.0	quadric surface
7.8	brightness temperature
7.8	mass flow
7.7	gas phase
7.7	surface cleaning
7.1	reaction cells
7.1	surface active
6.7	artery pulmonary
5.0	anode surface
4.7	surface temperature

In the A genre (News) the following 12 bigrams survived:

12.5	anne arundel
12.0	rhode island
10.0	grand jury
9.9	rule charter
9.2	austin texas
8.9	sunday sales
8.9	sales tax
8.9	payroll tax
8.2	fulton county
8.0	football league
7.5	kennedy administration
7.3	tax bill

Genres G and N contain few candidates for collocations (among the 'best' ones in N were "gray eyes", "picked up", "help me" and "stared at" which are quite telling about the prototypical western story: "The gray eyes stared at the villain who picked up his knife, while the girl cried "help me"."

7 Other approaches

The temporal dependencies of an ordered collocation [word1, word2] has been seen as a problem since the theory of mutual information assumes the frequencies of word pairs to be symmetric (i.e., f([w1, w2]) and f([w2, w1]) to be equal). Delta mutual information relies on this difference in temporal ordering.

> "[...] f(x, y) encodes linear precedence. [...] Although we could fix this problem by redefining f(x, y) to be symmetric (by averaging the matrix with its transpose), we have decided not to do so, since order information appears to be very interesting." (Church & Hanks, 1990:p.24)

Merkel, Nilsson, & Ahrenberg (1994) have constructed a system that uses frequency of recurrent segments to determine long phrases. In their approach they have to chunk the text into contiguous segments. Significant frequency counts are achieved through the use of a very large corpus, and/or a corpus specialised for a specific task. They report that it was possible for them to divide a large corpus into smaller sub-sections with little loss.

Smadja (1993) finds significant bigrams using an estimate of z-score (deviation from an expected mean). Smadja's method seems to require very large corpora, since the method needs to estimate a reliable measure of the variance of the frequencies with which words co-occur. This makes the method dependent on the corpus size. Smadja reports the use of a corpus of size 10 million words.

> "More precisely, the statistical methods we use do not seem to be effective on low frequency words (fewer than 100 occurrences)." (Smadja, 1993:p.168)

Kita & al. (1994) proposed another measure of collocational strength that was based on the notion of a reduction in 'processing cost' if a frequent chunk of text can be processed as one chunk. Cost reduction tended to extract conventional 'predicate phrase patterns', e.g., "*is that so*" and "*thank you very much*".

Steier & Belew (1991) discuss the 'exporting' of phrases into a general vocabulary, where a word pair with high mutual information within a topic tends to have lower mutual information within the collection, and vice versa. They relate a higher mutual information within a topic than in the collection to a lower value of discrimination.

Church & Gale (1995) have found it useful to compare the distribution of terms across documents. They showed that a distribution different from what could be expected by a (random) Poisson process indicates interesting terms. This approach is similar to the use of one genre to find interesting items in another. However, removal of the overlap needs some knowledge about the genres — apart from checking explicitly for a genre with least overlap. Cancelling overlap has the advantage that it can cancel out similar underlying causes, while it exaggerates the underlying causes that differ between genres. Some questions remain: at which level should overlap be formed? overlap in words or in bigrams; how many repetitions does it take for a word or bigram to 'belong' to a genre?

8 Conclusion

The question is "what is gained by using a measure?". Mutual information tends to find combinations of words that are highly co-ordinated with each other, but these bigrams show both interesting bigrams (e.g. "cheshire cat") and conventional (and uninteresting for keywords) bigrams (e.g. "in a"). The stability of interesting bigrams is improved by demanding candidate bigrams to occur more than a fixed number of times.

In this paper it has been shown that genre matters, and can be used to extract items that differ between genres. Instead of balancing one big corpus, the analysis of one corpus might benefit from finding out how it is different from another corpus. The bigrams that were formed by using different genres as filters showed interesting characteristics.

However, if we are to deal with larger amounts of data it might be unrealistic to compare differences directly between two large genres without the exclusion of terms that occur by chance.

The method that could be recommended from the results presented in this study is to triangulate a sample by the difference to other genres that we have some meta-knowledge about (i.e. we know that Western Fiction and Scientific Writing, at least on the surface, have little vocabulary in common).

References

Church, K., & Gale, W. (1995). Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. D. Yarowsky & K. Church (Eds.), *Third Workshop on Very Large Corpora* (pp. 121-130), MIT, Cambridge, Mass.

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics, 16*(1), 22-29.

Church, K., & Mercer, R. (1993). Introduction to the special issue on Computational Linguistics using Large Corpora. *Computational Linguistics, 19*(1), 1-24.

Gale, W., & Church, K. (1994). What is wrong with adding one? In N. Oostdijk & P. de Haan (Eds.), *Corpus based research into language* (pp. 189-198). Amsterdam - Atlanta: Rodopi.

Johansson, C. (1994). Catching the Cheshire Cat, *Coling* (pp. 1021-1025), Kyoto, Japan.

Kita, K., Omoto, T., Yano, Y., & Kato, Y. (1994). Application of Corpora in Second Language Learning — The Problem of Collocational Knowledge Acquisition —, *second annual workshop on very large corpora* (pp. 43-56), Kyoto, Japan.

Merkel, M., Nilsson, B., & Ahrenberg, L. (1994). A phrase-retrieval system based on recurrence, *second annual workshop on very large corpora* (pp. 99-108), Kyoto, Japan.

Sampson, G. (1994). SUSANNE: A Domesday Book of English Grammar. In N. Oostdijk & P. de Haan (Eds.), *Corpus-based research into language* (pp. 169-187). Amsterdam - Atlanta.

Shannon, C.E. (1951). Prediction and Entropy of printed English. *Bell Systems Technical Journal, 30*(1), 50-65.

Shannon, C.E., & Weaver, W. (1963). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics, 19*(1), 143-177.

Steier, A.M., & Belew, R.K. (1991). Exporting phrases: A statistical analysis of topical language. R. Casey & B. Croft (Eds.), *2nd Symposium on Document Analysis and Information Retrieval.*

Stolz, W. (1965). A probabilistic procedure for grouping words into phrases. *Language and Speech, 8*, 219-235.

Zellig, H. (1955). From phoneme to morpheme. *Language, 31*, 190-222.