Adam Kilgarriff, Information Technology Research Institute, University of Brighton
Raphael Salkie, Language Centre, University of Brighton

Corpus Similarity and Homogeneity via Word Frequency

Abstract

A measure of corpus similarity would be very useful for lexicography. Word frequency lists are cheap and easy to generate so a measure based on them can be used where a detailed comparison of the two corpora is not viable, for example, to judge how a new corpus relates to already-familiar ones. We show that corpus similarity can only be interpreted in the light of corpus homogeneity, and present a measure, based on the chi-square statistic, for measuring both corpus similarity and corpus homogeneity.

1. Introduction

How similar are two corpora? Does it matter whether lexicographers use this corpus or that, or are they similar enough for it to make no difference? How long will it take to adapt language-analysis software built with one corpus in mind, to work with another? In this paper we present a method of measuring corpus similarity.

Our approach uses word frequency lists. The full text is very rich in information, but that information is not available for automatic, objective manipulation. When a corpus is represented as a frequency list, much information is lost, but the tradeoff is an object that is susceptible to statistical processing. Word frequency lists are easy to generate, so measuring corpus similarity based on them will be viable in many circumstances where a more extensive analysis of the two corpora is not possible.

A judgement of similarity runs the risk of meaninglessness if a homogeneous corpus is compared with a heterogeneous one. We propose a method which can be used initially to measure corpus homogeneity, and subsequently to measure the similarity between two corpora. In brief, the method (for the homogeneity case) is as follows:

Divide the corpus into two halves by randomly placing texts in one of two subcorpora; produce a word frequency list for each subcorpus;

calculate the χ^2 statistic for the difference between the two lists; normalise; iterate (to give different random halves); interpret result by comparing values for different corpora.

For the corpus-similarity case, the only modifications are that one subcorpus is taken from the first corpus and the other from the second, and the similarity value is interpreted by reference to the homogeneity measure for each corpus.

Table 1 shows the possible outcomes for various permutations of the scores for homogeneity of corpus 1 (corp1), homogeneity of corpus 2 (corp2), and corpus dissimilarity (dis). High scores correspond to heterogeneous corpora and dissimilar corpora. The last two lines in the table point to the differences between general corpora and specific corpora. General corpora which embrace a number of language varieties will score highly for heterogeneity. Corpus similarity between such corpora will depend on whether all the same language varieties are represented in each corpus, and in what proportions. Low heterogeneity scores will typically apply to corpora of a single language variety, so here, similarity scores will be interpreted as a measure of the distance between the two language varieties.

2. The χ^2 test.

At a first pass, it would appear that the chi-square test will serve to indicate whether two corpora are drawn from the same population, or whether two or more phenomena are significantly different in their distributions between two corpora. For a contingency table of dimensions *m* x *n*, if the null hypothesis is true, the statistic

$$\sum \frac{(O-E)^2}{E}$$

(where O is the observed value, E is the expected value calculated on the basis of the joint corpus, and the sum is over the cells of the contingency table) will be χ^2 distributed with *(m-1)* x *(n-1)* degrees of freedom (1). If the figures are as in table 2, then the χ^2 statistic, with expected values based on probabilities in the joint corpus, is calculated as in table 3. The sum of the items in the last two columns of table 3 is 29.17, and four words were used for the comparison, so the χ^2 statistic is 29.17 on 4 degrees of freedom. (The "remainders" column is included in the

contingency table, giving a 5x2 table, so degrees of freedom = (5–1)x(2–1) = 4: the number of degrees of freedom equals the number of words used for the comparison). Looking at statistical tables for the distribution, we find that the critical value on 4 DF at the 99% significance level is 13.3. 29.17 is greater than 13.3, so we can conclude that corpus 1 and corpus 2 do not comprise words randomly drawn from the same population.

Hofland & Johansson (1989) use the χ^2 test to identify where words are significantly more frequent in the LOB corpus (of British English) than in the Brown corpus (of American English). In the table where they make the comparison, the χ^2 value for each word is given, with the value starred if it exceeds the critical value so one might infer that the LOB-Brown difference was non-random. Looking at the LOB-Brown comparison, we find that very many words, including most very common words, are starred. Much of the time, the null hypothesis is defeated. Does this show that all those words have systematically different patterns of usage in British and American English?

To test this, we took two corpora which were indisputably of the same language type: each was a random subset of the British National Corpus (BNC). The sampling was as follows: all texts shorter than 20,000 words were excluded and all others were truncated at 20,000 words. The truncated texts were randomly assigned to either corpus 1 or corpus 2, and frequency lists for each corpus were generated. As in the LOB-Brown comparison, for very many words (2), including most common words, the null hypothesis was defeated.

We conclude that the British-American differences were not the reason so many words were starred in the LOB-Brown corpus. Rather, any two corpora covering a range of registers (and comprising, say, less than 1000 samples of over 1000 words each) will show such differences. While it might seem plausible that oddities would balance out to give a population that was indistinguishable from one where the individual words (as opposed to the individual texts) had been randomly selected, this turns out not to be the case.

Let us look more closely at why this occurs. A key word in the last paragraph is "indistinguishable". In hypothesis testing, the objective is generally to see if the population can be distinguished from one that has been randomly generated – or, in our case, to see if the two populations are distinguishable from two populations which have been randomly generated on the basis of the frequencies in the joint corpus. Since speakers and writers do not choose words at random, we know that our corpora are not randomly generated. The only question, then, is whether there is enough evidence to say that they are not, with confidence. In

general, where a word is more common, there is more evidence. This is why a higher proportion of common words than of rare ones defeat the nullhypothesis. On the null hypothesis, the expected value for the $(O–E)^2/E$ term would be 0.5 and would not vary with word frequency. Table 4 shows that this term tends to be substantially higher than 0.5, and tends to increase with word frequency.

We cannot, then, use the χ^2 statistic for testing the null hypothesis, but nonetheless it does come close to meeting our requirements. The $(O–E)^2/E$ term gives a measure of the difference in a word's frequency between two corpora, and, while the measure tends to increase with word frequency, it does not increase by orders of magnitude. The strategy we adopt is therefore to calculate χ^2 for (sub)corpus pairs, but then, rather than comparing it with a critical value based on the null hypothesis, we compare it with the χ^2 value for other (sub)corpus pairs (3).

3. Normalisation

At a first pass, three desiderata for a measure of corpus homogeneity or similarity are that a similarity measure based on data for more words should be directly comparable with one based on fewer words; one based on data for higher-frequency words should be directly comparable with one based on lower-frequency words; and corpora of different sizes should be directly comparable.

None of these hold for the χ^2 statistic as it stands. In relation to different numbers of words (which correspond to different numbers of degrees of freedom): for χ^2 statistics drawn from two independent samples, with n and m degrees of freedom, the F-test can be used to test the null hypothesis. Each χ^2 value is divided by its degrees of freedom, and then their ratio is taken. If the ratio is close to one, the null hypothesis that the variances of the samples are drawn from the same population is not defeated. The F-distribution on m and n degrees of freedom provides critical values for "close to one" to a specified confidence level (and is tabulated in statistics textbooks). In our case, although the assumptions of normality which underlie the F distribution do not hold, the empirical evidence shows that the basic method holds good: if χ^2 is divided by the degrees of freedom (to give a statistic we shall call CBDF, "Chi By Degrees of Freedom"), then statistics based on different numbers of degrees of freedom become directly comparable. CBDF figures tend to be in the range 1–50.

How do we compare evidence from high-frequency and low-frequency words? As the discussion of the χ^2 test shows in theory, and Table

4 shows in practice, common words tend to have substantially higher $(O–E)^2/E$ values than less common ones. If we compare CBDF figures for the 500 most common words with ones for the top 5,000 words, the former will tend to be higher because the high scores of the very common words are "diluted" in the latter. We are currently investigating the issue further, particularly since there is the possibility of comparing corpora both in terms of form (looking at the very common words which are predominantly closed-class, 'grammatical', form words) and in terms of content (excluding the most common words, thus looking at predominantly open-class, 'lexical', content words). In the meantime, we sidestep the issue by simply taking the most common N words in the two corpora, for some convenient value of N.

In relation to corpora of different sizes, there is a theoretical problem: it is not clear what it means to say a bigger corpus is as homogeneous as a smaller one. If corpus 1 is twice as big as corpus 2, should it contain twice as many language varieties, or should it contain the same range of language varieties but twice as much of each? We are currently invesigating possible approaches.

4. Experiments

The experiments used subcorpora of the BNC each comprising 200 5,000-word samples, of five language varieties: "spoken" (Sp), "written-imaginative" (Im), and three written, non-fiction varieties, "low status" (Lo), "mid status" (Mid), and "high status" (Hi). The meaning of "status" is not documented in the BNC Users Reference Guide but would appear to relate to breadth of readership. National newspapers, bestselling books and TV scripts have "high status", the bulk of books have "mid status", and specialist books and periodicals and most unpublished material have "low status". The benefit of selecting material in this way is that one can make an independent judgement of how similar the various corpora are, so there is a "gold standard" to compare the outcomes of the experiments with. Given the defining criteria for the five corpora, one would expect the three non-fiction corpora to be most similar to each other, with Hi and Lo being less similar to each other than either is to Mid. One would expect Im and Sp both to be quite different to all others.

The algorithm (for corpus homogeneity) was:

- For each variety, identify all BNC files in that variety. Reject those with less than 5,000 words and randomly select 200 of the remainder.

- For each of these 200 files, truncate the text at 5,000 running words and produce a frequency list. Put the frequencies in a table, with a row for each word and a column for each file. Identify the N highest-frequency words in the entire subcorpus (Nkeywords).

- Repeat the following R times: randomly assign half the columns of the table to each of two half-corpora; calculate CBDF for the difference between the half-corpora on the basis of the Nkeywords.

- Calculate mean and standard deviation for CBDF over R iterations.

Various values of R and N were experimented with. R=10 and N=5,000 were used for the homogeneity results shown in table 5.

To measure corpus similarity, the algorithm was very similar but one half-corpus was taken from each of the corpora being compared. No value for N was set: the number of degrees of freedom was determined by the number of words for which there was sufficient data in the two half-corpora for the "expected" value in each half-corpus to exceed 5. In practise, this gave between 3456 and 6085 degrees of freedom – numbers of the same order of magnitude as 5,000.

Table 5 gives the scores for all corpus pairs (so the pairs where the two corpora are the same are homogeneity measures; others are similarity measures.) Main figures are average values for CBDF. The first number in brackets on the line below is the standard deviation and the second, for corpus-similarity cases, is the average number of degrees of freedom. The results are promising. They conform with the "gold standard". Hi is less similar to Lo, than either is to Mid. Im and Sp, while not notably more or less homogeneous than the non-fiction corpora, are of resoundingly different varieties. All the similarity scores are over 20 for Im, and over 30 for Sp. As we might expect, Spoken is still further removed from non-fiction than Imaginative. Also, Hi is the least homogeneous corpus, possibly suggesting that its two major components – bestsellers and newspapers – are quite dissimilar. Mid is slightly more homogeneous than Lo, and it would appear that the variation within Mid mostly falls within the variation of Lo, since the similarity measure for the two is very close to the homogeneity measure for Lo, but significantly above that for Mid.

5. Conclusion

A measure of corpus similarity has been presented. It uses word frequency information for the two corpora, and the χ^2 statistic. The measure can also be used to quantify the homogeneity of a corpus. The relation between corpus homogeneity and corpus similarity was considered: a corpus similarity score must be interpreted relative to the homogeneity scores of the two corpora. Homogeneity and similarity scores were calculated for various corpora where an independent judgement of their similarity could be made, and there was a good fit between the independent judgement and the (interpreted) similarity scores. The measure is potentially of value for lexicography and language engineering.

Notes

1. Provided all expected values are over a threshold of 5. Where there is just one degree of freedom, Yates' correction is applied.
2. Strictly, word-form-and-part-of-speech lists; the BNC is part-of-speech tagged, and part-of-speech distinctions were retained in the lists. No lemmatisation was carried out. All experiments reported in this paper were performed on such <wordform, POS> lists.
3. The log-likelihood statistic (Dunning, 1993) would have the same advantages, and is, mathematically, a more appropriate test. We shall consider using it for future experiments. It has not been used in the current trials because it is more complex to compute and, where expected values for word frequencies are over 5 and the probability of the next word being the word of interest is less than 1 in 50, the difference between chi-squared and log-likelihood is very small. These two conditions hold for all the data (except the data for *the, of, and* and *a*) that we are using. For a survey of statistical approaches, see Kilgarriff (1996).

Tables

corp1	corp2	dis	Comment
equal	equal	equal	same language variety/ies.
equal	equal	much higher	different language varieties.
high	low	high	corp2 is homogeneous and falls within the range of "general" corp1.
high	low	higher	corp2 is homogeneous and falls outside the range of "general" corp1.
high	high	low	impossible
high	high	a bit higher	overlapping; share some varieties
low	low	a bit higher	similar varieties

Table1: Interactions between homogeneity and similarity.

	Corpus 1	Corpus 2
Totals	1234567	1876543
the	80123	121045
of	36356	56101
and	25143	37731
a	19976	29164

Table 2: Word frequencies in two corpora.

	o1	o2	e1	e2	$\frac{(o1-e1)^2}{e1}$	$\frac{(o2-e2)^2}{e2}$
the	81023	121045	79828.5	121339.5	1.09	0.71
of	36356	56101	36689.3	55767.7	3.03	1.99
and	25143	37731	24850.0	37924.0	1.49	0.98
a	19976	29164	19500.0	29640.0	11.62	7.64
Remainders	1072969	1632502	1073599.2	1631871.8	0.37	0.24

Table 3: The χ^2 statistic for two corpora.

Class (Words in freq. order)	*First item in class* Word	POS	Mean error term for items in class
First 10 items	the	DET	18.76
Next 10 items	for	PREP	17.45
Next 20 items	not	NOT	14.39
Next 40 items	have	V-BASE	10.71
Next 80 items	also	ADV	7.03
Next 160 items	know	V-INF	6.40
Next 320 items	six	CARD	5.30
Next 640 items	finally	ADV	6.71
Next 1280 items	plants	N-PL	6.05
Next 2560 items	pocket	N-SING	5.82
Next 5120 items	represent	V-BASE	4.53
Next 10240 items	peking	PROPER	3.07
Next 20480 items	fondly	ADV	1.87

Table 4: Variation of $(O-E)^2/E$ term with word frequency for same-variety corpora, for high-frequency and low-frequency word-POS pairs. Part-of-speech codes are from the CLAWS tagset as used in the BNC (modified/lengthened for easier reading).

	Lo	Mid	Hi	Im	Sp
Lo	5.1 (.3)				
Mid	5.4 (.3; 6085)	4.5 (0.2)			
Hi	9.3 (.5; 5450)	6.7 (0.4; 5729)	6.0 (.2)		
Im	26.3 (1.7; 4460)	28.7 (2.2; 4407)	42.0 (1.6; 3290)	4.6 (.2)	
Sp	35.4 (2.0; 4126)	36.3 (0.8; 4144)	47.6 (1.7; 3820)	35.4 (1.2; 3456)	4.8 (0.3)

Table 5: Corpus homogeneity and corpus similarity results. Unbracketed: Mean homogeneity/similarity figures. Bracketed: standard deviation, and number of degrees of freedom. Number of degrees of freedom for all homogeneity figures (on leading diagonal) is 5,000.

References

Dunning, T. 1993. " Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics.* 19(1). Pp 61–74.

Hofland, K. and Johanssen, S. 1989. *Frequency analysis of English vocabulary and grammar, based on the LOB corpus.* Oxford: Clarendon.

Kilgarriff, A. 1996. " Which words are particularly characteristic of a text? A survey of statistical approaches." *Proceedings, ALLC-ACH '96.* Bergen, Norway.