Measures for corpus similarity and homogeneity

Adam Kilgarriff[*]
ITRI, University of Brighton

Tony Rose
Canon Research Centre Europe

Abstract

How similar are two corpora? A measure of corpus similarity would be very useful for NLP for many purposes, such as estimating the work involved in porting a system from one domain to another. First, we discuss difficulties in identifying what we mean by 'corpus similarity': human similarity judgements are not fine-grained enough, corpus similarity is inherently multi-dimensional, and similarity can only be interpreted in the light of corpus homogeneity. We then present an operational definition of corpus similarity which addresses or circumvents the problems, using purpose-built sets of "known-similarity corpora". These KSC sets can be used to evaluate the measures. We evaluate the measures described in the literature, including three variants of the information theoretic measure 'perplexity'. A χ^2-based measure, using word frequencies, is shown to be the best of those tested.

The Problem

How similar are two corpora? The question arises on many occasions. In NLP, many useful results can be generated from corpora, but when can the results developed using one corpus be applied to another? How much will it cost to port an NLP application from one domain, with one corpus, to another, with another? For linguistics, does it matter whether language researchers use this corpora or that, or are they similar enough for it to make no difference? There are also questions of more general interest. Looking at British national newspapers: is the Independent more like the Guardian or the Telegraph?[1]

What are the constraints on a measure for corpus similarity? The first is simply that its findings correspond to unequivocal human judgements. It must match our intuition that, eg, a corpus of syntax papers is more like one of semantics papers than one of shopping lists. The constraint is key but is weak. Direct human intuitions on corpus similarity are not easy to come by, firstly, because large corpora, unlike coherent texts, are not the sorts of things people read, so people are not generally in a position to have any intuitions about them. Secondly, a human response to the question, "how similar are two objects", where those objects are complex and multi-dimensional, will themselves be multi-dimensional: things will be similar in some ways and dissimilar in others. To ask a human to reduce a set of perceptions about the similarities and differences between two complex objects to a single figure is an exercise of dubious value.

This serves to emphasise an underlying truth: corpus similarity is complex, and there is no absolute answer to "is Corpus 1 more like Corpus 2 than Corpus 3?". All there are, are possible measures which serve particular purposes more or less well. Given the task of costing the customisation of an NLP system, produced for one domain, to another, a corpus similarity measure is of interest insofar as it predicts how long the porting will take. It could be that a measure which predicts well for one NLP system, predicts badly for another. It can only be established whether a measure correctly predicts actual costs, by investigating actual costs.[2]

Having struck a note of caution, we now proceed on the hypothesis that there is a single measure which corresponds to pre-theoretical intuitions about 'similarity' and which is a good indicator of many properties of interest – customisation costs, the likelihood that linguistic findings based on one corpus apply to another, etc. We would expect the limitations of the hypothesis to show through at some point, when different measures are shown to be suited to different purposes, but in the current situation, where there has been almost no work

[*] Kilgarriff's part of the work was undertaken under EPSRC grant GR/K/18931

[1] The work presented here develops and extends that presented in Kilgarriff (1997).

[2] Cf. Ueberla (1997), who looks in detail at the appropriateness of perplexity as a measure of task difficulty for speech recognition, and finds it wanting.

Corpus 1	Corpus 2	Distance	Interpretation
equal	equal	equal	same language variety/ies
equal	equal	high	different language varieties
high	low	high	corpus 2 is homogeneous and falls within the range of 'general' corpus 1
high	low	higher	corpus 2 is homogeneous and falls outside the range of 'general' corpus 1
high	high	low	impossible
low	low	a bit lower	overlapping; share some varieties
high	high	a bit lower	similar varieties

Table 1: **Interactions between homogeneity and similarity:** a similarity measure can only be interpreted with respect to homogeneity.
High means a large distance between corpora, or large within-corpus distances, so the corpus is heterogeneous/corpora are dissimilar; **low**, that the distances are low, so the corpus is homogeneous/corpora are similar. **High, low** and **equal** are relative to the other columns in the same row, so, in row 2, 'equal' in the first two columns reads that the within-corpus distance (homogeneity) of Corpus 1 is roughly equal to the within-corpus distance of Corpus 2, and 'high' in the Distance column reads that the distance between the corpora is substantially higher than these within-corpus distances.

on the question, it is a good starting point.

Similarity and homogeneity

How homogeneous is a corpus? The question is both of interest in its own right, and is a preliminary to any quantitative approach to corpus similarity. In its own right, because a sublanguage corpus, or one containing only a specific language variety, has very different characteristics to a general corpus (Biber, 1993) yet it is not obvious how a corpus's position on this scale can be assessed. As a preliminary to measuring corpus similarity, because it is not clear what a measure of similarity would mean if a homogeneous corpus (of, eg, software manuals) was being compared with a heterogeneous one (eg. Brown). Ideally, the same measure can be used for similarity and homogeneity, as then, Corpus 1/Corpus 2 distances will be directly comparable with heterogeneity (or "within-corpus distances") for Corpus1 and Corpus2. This is the approach adopted here.

Not all combinations of homogeneity and similarity scores are logically possible. A corpus cannot be much more similar to something else than it is to itself. Some of the permutations, and their interpretations, are shown in Table 1.

The last two lines in the table point to the differences between general corpora and specific corpora. High within-corpus distance scores will be for general corpora, which embrace a number of language varieties. Corpus similarity between general corpora will be a matter of whether all the same language varieties are represented in each corpus, and in what proportions. Low within-corpus distance scores will typically relate to corpora of a single language variety, so here, scores may be interpreted as a measure of the distance between the two varieties.

Related Work

There is very little work which explicitly aims to measure similarity between corpora. Johansson and Hofland (1989) aim to find which genres, within the LOB corpus, most resemble each other. They take the 89 most common words in the corpus, find their rank within each genre, and calculate the Spearman rank correlation statistic ('spearman').

Rose, Haddock, and Tucker (1997) explore how performance of a speech recognition system varies with the size and specificity of the training data used to build the language model. They have a small corpus of the target text type, and experiment with 'growing' their seed corpus by adding more same-text-type material. They use spearman and log-likelihood (Dunning, 1993) as measures to identify same-text-type corpora. Spearman is evaluated below.

There is a large body of work aiming to find words which are particularly characteristic of one text, or corpus, in contrast to another, in various fields including linguistic variation studies (Rayson, Leech, and Hodges, 1997), author identification (Mosteller and Wallace, 1964) and information retrieval (Salton, 1989; Dunning, 1993). Biber (1988, 1995) explores and quantifies the differences between corpora from a sociolinguistic perspective. While all of this work touches on corpus-similarity, none looks at is as a topic of itself.

Sekine (1997) explores the domain dependence of parsing. He parses corpora of various text genres and counts the number of occurrences of each subtree of depth one. This gives him a subtree frequency list for each corpus, and he is then able to investigate which subtrees are markedly different in frequency between corpora. Such work is highly salient for customising parsers for particular domains. Subtree frequencies could readily replace word frequencies for the frequency-based measures below.

In information-theoretic approaches, perplexity is a widely-used measure. Given a language model and a corpus, perplexity "is, crudely speaking, a measure of the size of the set of words from which the next word is chosen given that we observe the history of ... words" (Roukos, 1996). Perplexity is most often used to assess how good a language modelling strategy is, so is used with the corpus held constant. Achieving low perplexity in the language model is critical for high-accuracy speech recognition, as it means there are fewer high-likelihood candidate words for the speech signal to be compared with.

Perplexity can be used to measure a property akin to homogeneity if the language modelling strategy is held constant and the corpora are varied. In this case, perplexity is taken to measure the intrinsic difficulty of the speech recognition task: the less constraint the domain corpus provides on what the next word might be, the harder the task. Thus Roukos (1996) presents a table in which different corpora are associated with different perplexities.

Perplexity measures are evaluated below.

"Known-Similarity Corpora"

A "Known-Similarity Corpora" (KSC) set is built as follows: two reasonably distinct text types, A and B, are taken. Corpus 1 comprises 100% A; Corpus 2, 90% A and 10% B; Corpus 3, 80% A and 20% B; and so on. We now have at our disposal a set of fine-grained statements of corpus similarity: Corpus 1 is more like Corpus 2 than Corpus 1 is like Corpus 3. Corpus 2 is more like Corpus 3 than Corpus 1 is like Corpus 4, etc. Alternative measures can now be evaluated, by determining how many of these 'gold standard judgements' they get right. For a set of n Known-Similarity Corpora there are

$$\sum_{i=1}^{n}(n-i)\left(\frac{i(i+1)}{2}-1\right)$$

gold standard judgements (see Appendix for proof) and the ideal measure would get all of them right. Measures can be compared by seeing what percentage of gold standard judgements they get right.

Two limitations on the validity of the method are, first, there are different ways in which corpora can be different. They can be different because each represents one language variety, and these varieties are different, or because they contain different mixes, with some of the same varieties. The method only directly addresses the latter model.

Second, if the corpora are small and the difference in proportions between the corpora is also small, it is not clear that all the 'gold standard' assertions are in fact true. There may be a finance supplement in one of the copies of the Guardian in the corpus, and one of the copies of Accountancy may be full of political stories: perhaps, then, Corpus 3 *is* more like Corpus 5 than Corpus 4. This was addressed by selecting the two text types with care so they were similar enough so the measures were not 100% correct yet dissimilar enough to make it likely that all gold-standard judgements were true, and by ensuring there was enough data and enough KSC-sets so that oddities of individual corpora did not obscure the picture of the best overall measure.

Measures

All the measures use spelt forms of words. None make use of linguistic theories. Comments on an earlier version of the paper included the suggestion that lemmas, or word senses, or syntactic constituents, were more appropriate objects to count and perform computations on than spelt forms. This would in many ways be desirable. However there are costs to be considered. To count, for example, syntactic constituents requires, firstly, a theory of what the syntactic constituents are; secondly, an account of how they can be recognised in running text; and thirdly, a program which performs the recognition. Shortcomings or bugs in any of the three will tend to degrade performance, and it will not be straightforward to allocate blame. Different theories and implementations are likely to have been developed with different varieties of text in focus, so the degradation may well effect different text types differentially. Moreover, practical users of a corpus-similarity measure cannot be expected to invest energy in particular linguistic modules and associated theory. To be of general utility, a measure should be as theory-neutral as possible.

While we are planning to explore counts of lemmas and part-of-speech categories, in these experiments we consider only raw word-counts.

Word Frequency measures

Two word frequency measures were considered. For each, the statistic did not dictate which words should be compared across the two corpora. In a preliminary investigation we had experimented with taking the most frequent 10, 20, 40 ... 640, 1280, 2560, 5120 words in the union of the two corpora as data points, and had achieved the best results with 320 or 640. For the experiments below, we used the most frequent 500 words.

Both word-frequency measures can be directly applied to pairs of corpora, but only indirectly to measure homogeneity. To measure homogeneity:

1. divide the corpus into 'slices';
2. create two subcorpora by randomly allocating half the slices to each;
3. measure the similarity between the subcorpora;
4. iterate with different random allocations of slices;
5. calculate mean and standard deviation over all iterations.

Wherever similarity and homogeneity figures were to be compared, the same method was adopting for calculating corpus similarity, with one subcorpus comprising a random half of Corpus 1, the other, a random half of Corpus 2.

Spearman Rank Correlation Co-efficient

Ranked wordlists are produced for Corpus 1 and Corpus 2. For each of the n most common words, the difference in rank order between the two corpora is taken. The statistic is then the normalised sum of the squares of these differences,

$$1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Comment Spearman is easy to compute and is independent of corpus size: one can directly compare ranked lists for large and small corpora. However there was an *a priori* objection to the statistic. For very frequent words, a difference of rank order is highly significant: if *the* is the most common word in corpus 1 but only 3rd in corpus 2, this indicates a high degree of difference between the genres. At the other end of the scale, if *bread* is in 400th position in the one corpus and 500th in the other, this is of no significance, yet Spearman counts the latter as far more significant than the former.

χ^2

For each of the n most common words, we calculate the number of occurrences in each corpus that would be expected if both corpora were random samples from the same population. If the size of corpora 1 and 2 are N_1, N_2 and word w has observed frequencies $o_{w,1}, o_{w,2}$, then expected value $e_{w,1} = \frac{N_1 \times (o_{w,1}+o_{w,2})}{N_1+N_2}$ and likewise for $e_{w,2}$; then

$$\chi^2 = \Sigma \frac{(o-e)^2}{e}$$

Comment The inspiration for the statistic comes from the χ^2-test for statistical independence. As Kilgarriff (1996) shows, the statistic is not in general appropriate for hypothesis-testing in corpus linguistics: a corpus is never a random sample of words, so the null hypothesis is of no interest. But once divested of the hypothesis-testing link, χ^2 is suitable. 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, in contrast to the raw frequencies it does not increase by orders of magnitude.

The measure does not directly permit comparison between corpora of different sizes.

Perplexity and Cross-entropy

From an information-theoretic point of view, *prima facie*, entropy is a well-defined term capturing the informal notion of homogeneity, and the cross-entropy between two corpora captures their similarity. Entropy is not a quantity that can be directly measured. The standard problem for statistical language modelling is to aim to find the model for which the cross-entropy of the model for the corpus is as low as possible. For a perfect language model, the cross-entropy would be the entropy of the corpus (Church and Mercer, 1993; Charniak, 1993).

With language modelling strategy held constant, the cross-entropy of a language model (LM) trained on Corpus 1, as applied to Corpus 2, is a similarity measure. The cross-entropy of the LM based on nine tenths of Corpus 1, as applied to the other 'held-out' tenth, is a measure of homogeneity. We standardised on the 'tenfold cross-validation' method for measures of both similarity and homogeneity: that is, for each corpus, we divided the corpus into ten parts[3] and produced ten LMs, using nine tenths and leaving out a different tenth each time. (Perplexity is the log of the cross-entropy of a corpus with itself: measuring homogeneity as self-similarity is standard practice in information theoretic approaches.)

To measure homogeneity, we calculated the cross-entropy of each of these LMs as applied to the left-out tenth, and took the mean of the ten values. To measure similarity, we calculated the cross-entropy of each of the Corpus 1 LMs as applied to a tenth of Corpus 2 (using a different tenth each time). We then repeated the procedure with the roles of Corpus 1 and Corpus 2 reversed, and took the mean of the 20 values.

[3] For the KSC corpora, we ensured that each tenth had an appropriate mix of text types, so that, eg, each tenth of a corpus comprising 70% Guardian, 30% BMJ, also comprised 70% Guardian, 30% BMJ.

All LMs were trigram models. All LMs were produced and calculations performed using the CMU/Cambridge toolkit (Rosenfeld, 1995).

The treatment of words in the test material but not in the training material was critical to our procedure. It is typical in the language modelling community to represent such words with the symbol UNK, and to calculate the probability for the occurrence of UNK in the test corpus using one of three main strategies.

Closed vocabulary The vocabulary is defined to include all items in training and test data. Probabilities for those items that occur in training but not test data, the 'zerotons', are estimated by sharing out the probability mass initially assigned to the singletons and doubletons to include the zerotons.

Open, type 1 The vocabulary is chosen independently of the training and test data, so the probability of UNK may be estimated by counting the occurrence of unknown words in the training data and dividing by N (the total number of words).

Open, type 2 The vocabulary is defined to include all and only the training data, so the probability of UNK cannot be estimated directly from the training data. It is estimated instead using the discount mass created by the normalisation procedure.

All three strategies were evaluated.

Data

All KSC sets were subsets of the British National Corpus (BNC)[4]. A number of sets were prepared as follows.

For those newspapers or periodicals for which the BNC contained over 300,000 running words of text, word frequency lists were generated and similarity and homogeneity were calculated (using χ^2). We then selected pairs of text types which were moderately distinct, but not too distinct, to use to generate KSC sets. (In initial experiments, more highly distinct text types had been used, but then both Spearman and χ^2 had scored 100%, so 'harder' tests involving more similar text types were selected.)

For each pair a and b, all the text in the BNC for each of a and b was divided into 10,000-word tranches. These tranches were randomly shuffled and allocated as follows:

first 10 of a	into	b0a
next 9 of a, first 1 of b	into	b1a
next 8 of a, next 2 of b	into	b2a
next 7 of a, next 3 of b	into	b3a
...		

until either the tranches of a or b ran out, or a complete 11-corpus KSC-set was formed. A sample of KSC sets are available on the web.[5] There were 21 sets containing between 5 and 11 corpora. The method ensured that the same piece of text never occurred in more than one of the corpora in a KSC set.

The text types used were:
Accountancy (acc); The Art Newspaper (art); British Medical Journal (bmj); Environment Digest (env); The Guardian (gua); The Scotsman (sco); and Today ('low-brow' daily newspaper, tod).

To the extent that some text types differ in content, whereas others differ in style, both sources of variation are captured here. Accountancy and The Art Newspaper are both trade journals, though in very different domains, while The Guardian and Today are both general national newspapers, of different styles.

Results

For each KSC-set, for each gold-standard judgement, the 'correct answer' was known, eg., "the similarity 1,2 is greater than the similarity 0,3". A given measure either agreed with this gold-standard statement, or disagreed. The percentage of times it agreed is a measure of the quality of the measure. Results for the cases where all four measures were investigated are presented in Table 2.

	spear	χ^2	closed	type 1	type 2
KSC-set					
acc_gua	93.33	91.33	82.22	81.11	80.44
art_gua	95.60	93.03	84.00	83.77	84.00
bmj_gua	95.57	97.27	88.77	89.11	88.77
env_gua	99.65	99.31	87.07	84.35	86.73

Table 2: Comparison of four measures

The word frequency measures outperformed the perplexity ones. It is also salient that the perplexity measures required far more computation: ca. 12 hours on a Sun, as opposed to around a minute.

Spearman and χ^2 were tested on all 21 KSC-sets, and χ^2 performed better for 13 of them, as shown in Table 3.

	spear	χ^2	tie	total
Highest score	5	13	3	21

Table 3: Spearman/χ^2 comparison on all KSCs

[4] http://info.ox.ac.uk/bnc

[5] http://www.itri.bton.ac.uk/~Adam.Kilgarriff/KSC/

The difference was significant (related t-test: t=4.47, 20DF, significant at 99.9% level). χ^2 was the best of the measures compared.

Conclusions and further work

We have argued that computational linguistics is in urgent need of measures for corpus similarity and homogeneity. Without one, it is very difficult to talk accurately about the relevance of findings based on one corpus, to another, or to predict the costs of porting an application to a new domain. We note that corpus similarity is complex and multifaceted, and that different measures might be required for different purposes. However, given the paucity of other work in the field, at this stage it is enough to seek a single measure which performs reasonably.

The Known-Similarity Corpora method for evaluating corpus-similarity measures was presented, and measures discussed in the literature were compared using it. For the corpus-size used and this approach to evaluation, χ^2 and Spearman both performed better than any of three cross-entropy measures. These measures have the advantage that they are cheap and straightforward to compute. χ^2 outperformed Spearman.

Further work is to include:

- developing a scale-independent χ^2-based statistic
- investigating a 2-dimensional measure for similarity, with one dimension for closed-class words and another for open-class words, to see whether differences in style and in domain can be distinguished
- evaluation of a log-likelihood-based measure, and of different vocabulary-sizes for open models. Then it will be possible to compare the 500-word measure for spearman and χ^2 more directly with the perplexity measures
- gathering data on the actual costs of porting systems, for correlation with results given by similarity measures
- comparing the method with Biber's feature-set and analysis.

References

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge University Press.

Biber, Douglas. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–242.

Biber, Douglas. 1995. *Dimensions in Register Variation.* Cambridge University Press.

Charniak, Eugene. 1993. *Statistical Language Learning.* MIT Press, Cambridge, Mass.

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

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

Johansson, Stig and Knut Hofland, editors. 1989. *Frequency Analysis of English vocabulary and grammar, based on the LOB corpus.* Clarendon, Oxford.

Kilgarriff, Adam. 1996. Which words are particularly characteristic of a text? a survey of statistical approaches. In *Language Engineering for Document Analysis and Recognition*, pages 33–40, Brighton, England, April. AISB Workshop Series.

Kilgarriff, Adam. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings, ACL SIGDAT workshop on very large corpora*, pages 231–245, Beijing and Hong Kong, August.

Mosteller, Frederick and David L. Wallace. 1964. *Applied Bayesian and Classical Inference - The Case of The Federalist Papers.* Springer Series in Satistics, Springer-Verlag.

Rayson, Paul, Geoffrey Leech, and Mary Hodges. 1997. Social differentiation in the use of English vocabulary: some analysis of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1):133–152.

Rose, Tony, Nicholas Haddock, and Roger Tucker. 1997. The effects of corpus size and homogeneity on language model quality. In *Proceedings, ACL SIGDAT workshop on very large corpora*, pages 178–191, Beijing and Hong Kong, August.

Rosenfeld, Ronald. 1995. The CMU Statistical Language Modelling Toolkit and its use in the 1994 ARPA CSR Evaluation. In *Proc. Spoken Language Technology Workshop*, Austin, Texas.

Roukos, Salim, 1996. *Language Representation*, chapter 1.6. National Science Foundation and European Commission, `www.cse.ogi/CSLU/HLTsurvey.html`.

Salton, Gerard. 1989. *Automatic Text Processing.* Addison-Wesley.

Sekine, Satshi. 1997. The domain dependence of parsing. In *Proc. Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington DC, April. ACL.

Ueberla, Joerg. 1997. Towards an improved performance measure for language models. Technical Report DERA/CIS/CIS5/TR97426, DERA. cmp-lg/9711009.

Appendix

The proof is based on the fact that the number of similarity judgements is the triangle number of the number of corpora in the set (less one), and that each new similarity judgement introduces a triangle number of gold standard judgements (once an ordering which rules out duplicates is imposed on gold standard judgements).

- A KSC set is ordered according to the proportion of text of type 1. Call the corpora in the set 1...n.
- A similarity judgement ('sim') between a and b (a,b) compares two corpora. To avoid duplication, we stipulate that a<b. Each sim is associated with a number of steps of difference between the corpora: dif(a,b)=b-a.
- A gold standard judgement ('gold') compares two sims; there is only a gold between a,b and c,d if a<b and c<d (as stipulated above) and also if a<=c, b>=d, and not (a=c and b=d). Each four-way comparison can only give rise to zero or one gold, as enforced by the ordering constraints. Each gold has a difference of difs ('difdif') of (b-a)-(d-c) (so, if we compare 3,5 with 3,4, difdif=1, but where we compare 2,7 with 3,4, difdif = 4). difdif(X,Y) = dif(X)-dif(Y).
- Adding an nth corpus to a KSC set introduces n-1 sims. Their difs vary from 1 (for (n-1),n) to n-1 (for 1,n).
- The number of golds with a sim of dif m as first term is a triangle number less one, $\sum_{i=2}^{m} i$ or $\frac{m(m+1)}{2} - 1$ For example, for 2,6 (dif=4) there are 2 golds of difdif 1 (eg with 2,5 and 3,6), 3 of difdif 2 (with 2,4, 3,5, 4,6), and 4 of difdif 3 (with 2,3, 3,4, 4,5, 5,6).
- With the addition of the nth corpus, we introduce n-1 sims with difs from 1 to n-1, so we add $\sum_{i=1}^{n-1} \frac{i(i+1)}{2} - 1$ golds. For the whole set, there are $\sum_{i=1}^{n} \sum_{j=1}^{i-1} \frac{j(j+1)}{2} - 1$ and collecting up repeated terms gives $\sum_{i=1}^{n} (n-i)(\frac{i(i+1)}{2} - 1)$