Measuring corpus homogeneity using a range of measures for inter-document distance

Gabriela Cavaglià

ITRI, University of Brighton
Lewes Road, Brighton
BN2 4GJ, United Kingdom
Gabriela.Cavaglia@itri.brighton.ac.uk

Abstract

With the ever more widespread use of corpora in language research, it is becoming increasingly important to be able to describe and compare corpora. The analysis of corpus homogeneity is preliminary to any quantitative approach to corpora comparison. We describe a method for text analysis based only on document-internal linguistic features, and a set of related homogeneity measures based on inter-document distance. We present a preliminary experiment to validate the hypothesis that in the presence of a homogeneous corpus the subcorpus that is necessary to train an NLP system is smaller than the one required if a heterogeneous corpus is used.

1. Introduction

Corpora are collections of documents, generally in electronic form, used mainly as a source of different kinds of linguistic information. In the last decade, the availability of many texts in machine-readable form and the development of powerful tools for exploiting them make corpora the basic resource for many areas of language research (Lexicography, Linguistics, Psycholinguistics, Natural Language Processing). When a study performed on a particular corpus obtains interesting results, the possibility of extending them to a larger population is very tempting. But only corpora built according to explicit design criteria, which constitute a representative sample of a defined language variety, can allow the result of the study to be extended without major bias errors. It follows that characterizations of existing corpora and the design of new ones are now receiving more attention: it becomes essential to be able to describe a corpus, compare it with others, and produce new corpora that are representative samples of particular language varieties.

The criteria that one might use to describe or design a corpus can be *external* or *internal*. External criteria are essentially non-linguistic, and therefore not present in the document itself; they cover the document's topic, genre and socio-cultural aspects (e.g., age and occupation of the author) and are standardly assigned by people. By contrast, internal criteria are based on linguistic features, which are more or less directly present inside a text (e.g., words are directly present in a text, while POS tags can be exploited only after further analysis).

Corpus descriptions and corpus design techniques are usually based on external criteria (e.g., the 'Wall Street Journal corpus'). The main problem with external features is that they are not always available and, when they are, not always reliable (e.g., you can not always use the title of a text to identify its topic). Moreover, corpora produced using external features can contain wide variations in internal features, which can cause problems when used by an NLP system. The decision to classify texts only on the basis of external criteria is motivated when the users are human beings, who can cope without any problem with differences in linguistic features. But when the user is a system, as in NLP, the performance of any task can be degraded by the presence of different linguistic features, as shown in Biber (1993; Sekine (1997; Roland and Jurafsky (1998; Folch et al. (2000).

2. Corpus profiling: homogeneity and similarity

The problem of describing and comparing corpora in relation to their internal features is becoming important. Work on corpora comparison started in the early '80s (Hofland and Johansson, 1982) with the study of the differences between British and American English and then extended to the opposition between spoken and written English (Biber, 1988) and later to differences in register (Biber, 1993; Kessler et al., 1997; Dewdney et al., 2001).

More recent developments focus on corpus homogeneity and similarity. Both homogeneity and similarity are complex and multi-dimensional issues: a corpus can be homogeneous, and two or more corpora can be similar, in relation to aspects such as lexis, syntax, semantics but also in relation to the structure of the texts or the presence of extra-textual information. Because we are mainly interested in textual information, we restrict our analysis to lexical, semantic and syntactic aspects, and we label the corpus profiling we are interested in as "linguistic". We call a corpus "homogeneous" when it does not contain major differences in internal features among its documents.

Kilgarriff (2001) defines corpus similarity as the "likehood that linguistic findings based on one corpus apply to another". He presents corpus homogeneity as the preliminary step to any quantitative study of corpus similarity: his claim is that without knowledge of corpus homogeneity it is not clear if it would be appropriate to measure similarity between, for example, a homogeneous corpus like the PILLs corpus of Patient Information Leaflets (Scott et al., 2001) and a balanced one like the Brown. He also states that ideally the measure used for corpus similarity can be used for corpus homogeneity, and presents an analysis based on word frequency lists. Illouz et al. (2000) present a methodology for text profiling that aims to produce measures for corpus homogeneity within the different parts of a corpus. Their supervised approach is similar to Biber's work on text classification, but they use a tagger/parser to analyze syntactic features.

We present a technique for corpus analysis strictly based on internal features and unsupervised learning techniques, together with a set of measures for corpus homogeneity and similarity.

3. The methodology

We propose a stochastic method to describe and compare corpora, which is based only on their internal features. This method can be computed for any corpus, and is independent of any particular theory of language: it uses all the linguistic features of the documents, and not just a special sub-set as for example in Biber's work. The method has four steps:

1. choose the aspect you want to study and the type of feature you want to use;

2. collect data for each document in the corpus;

3. calculate the similarity between each pair of documents;

4. quantify the characteristics of the corpus: we produce both a description of the corpus and measures of its homogeneity and its similarity in relation to other corpora.

3.1. Deciding aspect and feature types

Corpus profiling can be studied from different perspectives. As has been said, we restrict our interest to linguistic analysis and in particular to the lexical, syntactic and semantic *aspects*. Each aspect can be studied using different *feature types* (i.e. words or POS tags). At the moment just the lexical and syntactic aspects have been investigated.

Lexical analysis is performed to detect possible restrictions in the vocabulary. As feature types for lexical analysis, either all-words or content words or lemmas are used. To identify restrictions at the syntactic level, either function words or POS tags or POS bi-grams are used. To detect function words a list of function words is needed, while to produce POS tag and POS bi-gram frequency lists a POS tagger is required.

3.2. Collecting the data

The objects employed to study corpus profiling are the texts that make up the corpus. Each text is represented by a vector of features (attributes) or a *frequency list*. A frequency list is a list of pairs $< x, f(x) >$ in which x is a feature instance, e.g., the function word "with" or the lemma "to cut", and $f(x)$ is the frequency of the feature x in the document (the number of occurrences of "with" or "to cut" in the document). Instead of using the sample frequency $f(x)$ direcly, we compute the estimate of the probability $p(x)$.

This step yields a matrix which represents a corpus by the frequency lists of its documents.

3.3. Computing similarity

Probability lists representing texts in the corpus can also be seen as distributions. Two documents are considered similar if their probability distributions are similar. We explored the use of three different text-similarity measures.

Relative entropy, also know as Kullback-Leibler divergence, is a well-known measure for calculating how similar two probability distributions are (over the same event space). If $p(i)$ and $q(i)$ are the distributions which represent two documents, the relative entropy, $D(p||q)$, is calculated as follow:

$$D(p||q) = \sum_{i=1}^{n} p(i) log \frac{p(i)}{q(i)}$$

Because it is not defined for $q(i) = 0$, for which $D(p||q) = \infty$, we compute the *centroid,* the average probability distribution of the corpus, and then add it to each distribution before calculating the similarity. The formula for relative entropy becomes:

$$D'(p||q) = \sum_{i=1}^{n} (p(i) + c(i)) log \frac{p(i) + c(i)}{q(i) + c(i)}$$

with $c(i)$ the centroid of the entire corpus.

We also tested two other similarity measures based on the divergence from the null hypothesis that the two documents are random samples from the same distribution: Chi Squared and Log-likelihood.

Chi Square measure (χ^2): for each feature in the frequency list, we calculate the number of occurrences in each document that would be expected. Suppose the sizes of documents *A* and *B* are respectively N_A and N_B and feature *w* has observed frequency $o_{w,A}$ in *A* and $o_{w,B}$ in *B*, then the expected value $e_{w,A}$ for *A* is:

$$e_{w,A} = \frac{N_A(o_{w,A} + o_{w,B})}{N_A + N_B}$$

and likewise for $e_{w,B}$ for document *B*. Then the χ^2 value for the document pair, A and B, is computed as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

with the sum over all the features.

Log-likelihood (G^2): Dunning (1993) showed that G^2 is a much better approximation of the binomial distribution than χ^2 especially for events with frequencies smaller than 5. It is a measure that works quite well with both large and small documents and allows the comparison of the significance of both common and rare events. A **Contingency table**, as presented in table 1, helps us to understand the formula for G^2.

$$\begin{aligned} G_w^2 = & \\ & 2(a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & -(a+b)\log(a+b) - (a+c)\log(a+c) \\ & -(b+d)\log(b+d) - (c+d)\log(c+d) \\ & +(a+b+c+d)\log(a+b+c+d) \end{aligned}$$

	Doc. A	Doc. B
w	a	b
¬ w	c	d

Table 1: Contingency table

$$G^2 = \sum_{i=1}^{n} G_i^2$$

This step yields a similarity matrix: to each pair of documents a distance is associated. Relative entropy, χ^2 and G^2 are all distance measures, so in the matrix the more similar text couples appear with a small value assigned.

3.4. Quantify homogeneity

The similarity values calculated in the previous step for each pair of documents in a corpus are now employed to produce information to help in describing the corpus and in quantifying its homogeneity and its similarity in relation to other corpora. The information we provide for the corpus is:

- a homogeneity measure which quantifies the variability of the features inside the corpus. The homogeneity measure corresponds to the maximum distance among its documents;
- the corpus prototypical element, to give the user an idea of what kind of text he/she can find in the corpus. In clustering, such an object is called "the medoid", the cluster element which is the nearest to the centroid;
- a similarity measure which describes the relative position of the corpus in relation to the others. The similarity of corpus A and B is the distance between their centroids.

The usefulness of a prototypical element and the validity of a similarity measure depend directly on the homogeneity of the corpora analyzed. The more a corpus is homogeneous the better its prototypical element can describe the corpus documents, because there is a smaller variance between it and the other documents of the corpus. The interpretation of a similarity measure computed between a homogeneous corpus and a heterogeneous one, or between two heterogeneous corpora, is not clear, and needs further analysis. In this paper we focus only on the evaluation of the homogeneity measures.

4. Evaluation

The aim of our first experiment is to understand which text-similarity measure is most reliable, among the three currently used ($D(p||q)$, χ^2 and G^2).

To evaluate the homogeneity of a corpus is difficult owing to the lack of gold-standard judgments with which the measures can be compared. The hypothesis at the base of homogeneity is that a NLP system can reach better results when it uses an homogeneous corpus rather than an heterogeneous one. In the experiment we run an NLP system using homogeneous and heterogeneous corpora. Then we compare the accuracy that the system achieved on each corpus, with the degree of homogeneity that the corpus scores. We expect to find that the accuracy for the homogeneous corpora is higher.

The NLP system we use for the evaluation is Rainbow (McCallum, 1996), which performs text classification. We choose Rainbow because it is freely available, fast, and does not require any particular annotation or linguistic resource other than the corpus itself. Because Rainbow performs text analysis (builds its model) using *all words* or *content words*, we have to restrict the evaluation to just these two internal features in this experiment. We collect a set of corpora for which we have a reliable classification, and compute the homogeneity measure for each corpus. For each corpus we measure homogeneity using the three inter-document similarity measures. Then, for each similarity measure, we rank the corpora according to their homogeneity value in increasing order, so that homogeneous corpora appear at the beginning of the list. For each of the two features, three ranked lists of homogeneity values are produced.

We then use Rainbow to produce similar ranked lists; using both all-words and content words, to use as a gold standard. All the corpora for which we measure the homogeneity are merged to form a single big corpus. We then use Rainbow to classify the new big corpus using different sizes of training corpus. The task for Rainbow is to rebuild from the merged corpus all the corpora it was made of. According to our hypothesis, in order to achieve the same level of accuracy, homogeneous corpora need to be trained on a smaller subcorpus than heterogeneous corpora. The accuracy of the classification of each class is computed. Classes are then ranked in a descending order, so that the homogeneous ones appear at the beginning of the list. For each of the two features, a rank list of Rainbow accuracy values is produced. Finally, the Spearman's rho test is employed to identify the correlation between the homogeneity values and Rainbow accuracy values.

5. Experiment

The corpus we used for this first experiment is the British National Corpus (BNC). The BNC is a 100 million-word collection of samples of written and spoken language, from a wide range of sources, designed to represent a wide cross-section of current British English (monolingual synchronic corpus). Moreover, it is a general corpus which includes many different language varieties, and is not limited to any particular subject field, genre or register. There has been a lot of work on the classification of BNC documents. The BNC Index (Lee, 2001) is an attempt to combine and consolidate some of these suggestions. The result is a resource which provides an accurate classification of the documents in the BNC, according to many different kinds of external criteria such as *medium*, *domain* and *genre*. According to the medium, BNC documents can be classified into six different classes: spoken, written-to-be-spoken, book, periodical, published miscellanea, and unpublished miscellanea. For domain, spoken English can be classified into five classes (e.g., transcription of business recordings, spontaneous natural conversations), and written English into nine (e.g., applied science, arts, belief and thought). There are 24 genres for spoken and 46 genres for written English (among the genres for written English there are for example personal letters, university essays, tabloid newspaper, bibliographies and instructional texts).

To avoid comparing classes whose size is too dissimilar:

- each BNC document is divided into chunks of a fixed size. For the first experiment, chunks of 20,000 words were produced. If a document is too small it is discharged. If it is big enough to contain more than one chunk, multiple chunks are produced, and considered as individual documents in the analysis;

- for each BNC classes from medium, domain and genre, a corpus is created with the same number of chunks. For the experiment we produced corpora of 20 chunks each. If a class does not contains enough documents it is discharged; otherwise 20 chunks are chosen randomly.

This gives 51 corpora with 20 documents of 20,000 words each. Three random corpora made of 20 chunks chosen randomly from the BNC are also created. We expected random corpora to be less homogeneous than all the other corpora.

The three homogeneity measures (using $D(p||q)$, χ^2 and G^2) for each corpus are calculated, and the corpora are ranked according to their homogeneity score: corpora with a lower score are considered more homogeneous than ones with a higher score. Then the 54 corpora are merged to form one big corpus, and Rainbow is used to see how accurately it can recover each of the 54 corpora, using training sets of different sizes. The sizes used were 1, 5, 10, 15; e.g., when the training set size was 5, the task for Rainbow was to recover the other 15 same-class documents out of the pot of 1080 documents. For each corpus, we compute the accuracy, the proportion of correctly classified documents, and the standard deviation calculated on 50 trials.

6. Results

Tables 2 and 3 list the homogeneity measures for the five corpora at the beginning and end of the lists ranked by homogeneity and Rainbow accuracy values, using all words as features. Using Rainbow, the three random corpora appear to be less homogeneous than all the other corpora, as expected. The homogeneity measures based on inter-document distance instead partially failed; in fact, although they all appear somewhere at the bottom of the rank list, just one of them turns up after all the non-random corpora.

We use Spearman's rho test (Owen and Jones, 1977) to compare the ranks obtained using the homogeneity measure and Rainbow. Spearman's correlation is 1 when the two ranks are exacly the same, and 0 when no correlation is found between the two ranked lists. The results, presented in table 4 for all words and in table 5 for content words, are always positive and usually within the significance level of 5%.

Corpus	$D(p\|\|q)$	χ^2	G^2
g-W_news_script	0.0489	0.0304	0.0663
g-W_newsp_tabloid	0.0948	0.0640	0.1503
g-W_newsp_other_report	0.1208	0.0817	0.1951
g-W_newsp_other_sports	0.1335	0.0750	0.1756
g-W_hansard	0.1459	0.0950	0.2283
d-W_app_science	0.6165	0.2973	0.7737
m-m_unpub	0.6502	0.3711	0.9649
g-W_misc	0.6572	0.2818	0.7040
g-W_advert	0.7201	0.2949	0.7519
random2	0.94344	0.4200	1.0906

Table 2: Homogeneity scores computed using the 500 most frequent words in each corpus

Rainbow	Homogeneity	Spearman's correlation
1 doc per class	$D(p\|\|q)$	0.526
1 doc per class	χ^2	0.527
1 doc per class	G^2	0.530
5 doc per class	$D(p\|\|q)$	0.447
5 doc per class	χ^2	0.473
5 doc per class	G^2	0.474
10 doc per class	$D(p\|\|q)$	0.432
10 doc per class	χ^2	0.451
10 doc per class	G^2	0.451
15 doc per class	$D(p\|\|q)$	0.387
15 doc per class	χ^2	0.413
15 doc per class	G^2	0.415

Table 4: Spearman's correlation between Rainbow accuracy values and Homogeneity values using words

7. Conclusion and future work

The Spearman correlation values show that the original distinction between homogeneous and heterogeneous corpora is maintained in Rainbow: corpora with a low homogeneity score need a small training set to achieve a high accuracy in the classification task. By contrast, heterogeneous and random corpora need a bigger training set to achieve an accuracy which, however, is smaller than the one obtained by the homogeneous corpora.

None of the three text-similarity measures used to compute homogeneity produces a rank which follows exactly the same order identified with Rainbow, even if the constant high values of the standard deviation suggest that the rank order identify by Rainbow is not fixed. Among the three measures, G^2 provides the closest rank, expecially when all words are used.

Various reasons may be responsible for the unclarity of the results:

- lack of data: chunks, in which we divide the documents, and the number of chunks, we set for each corpus, are not big enough. For this experiment we produce chunks of 20,000 words and we use corpora of 20 chunks each. We also try to use chunks of 50,000 words and corpora made of 50 chunks each, but the

Class	1	5	10	15
g-W_hansard	73.05 (32.61)	97.2 (6.74)	97.6 (4.31)	94.8 (8.86)
g-W_newsp_tabloid	70.52 (25.62)	89.46 (5.05)	83.4 (7.98)	85.6 (16.18)
g-W_ac_medicine	58.63 (26.75)	85.06 (9.67)	85.4 (8.85)	80.4 (16.89)
g-W_newsp_other_sports	56.31 (21.24)	88.26 (13.57)	92.6 (8.99)	96.8 (7.40)
g-W_news_script	53.57 (30.64)	63.73 (24.32)	71.2 (20.06)	66.4 (19.56)
m-periodical	3.26 (4.36)	0 (0)	0 (0)	0 (0)
m-book	2.52 (5.12)	1.86 (3.31)	0.6 (2.39)	0.8 (3.95)
random3	2.42 (4.28)	0.13 (0.94)	0 (0)	0 (0)
random2	2.21 (3.84)	0.26 (1.32)	0 (0)	0 (0)
random1	1.05 (2.12)	0.13 (0.94)	0.6 (2.39)	0 (0)

Table 3: The accuracy obtained by Rainbow analyzing all words for homogeneous and heterogeneous subcorpora using training set of different size: 1, 5, 10 and 15 document per class respectively

Rainbow	Homogeneity	Spearman's correlation
1 doc per class	$D(p\|\|q)$	0.445
1 doc per class	χ^2	0.383
1 doc per class	G^2	0.389
5 doc per class	$D(p\|\|q)$	0.291
5 doc per class	χ^2	0.277
5 doc per class	G^2	0.286
10 doc per class	$D(p\|\|q)$	0.273
10 doc per class	χ^2	0.269
10 doc per class	G^2	0.281
15 doc per class	$D(p\|\|q)$	0.240
15 doc per class	χ^2	0.232
15 doc per class	G^2	0.245

Table 5: Spearman's correlation between Rainbow accuracy values and Homogeneity values using content words

number of BNC classes which contain these amounts of data are few and appear to be all quite heterogeneous;

- presence of noise in the data: at the moment we use the N most frequent internal features present in each corpus, and for this experiment we set N equal to 500. We would like to consider ways of identifying the features that seem more likely to show differences among the documents, and of filtering out those which instead can only create noise;

- the use of Rainbow as a gold-standard judgment for homogeneity: to classify texts any system uses a mix of homogeneity and similarity, so the attempt to use its classification to evaluate a homogeneity measure can be misleading ;

- the three text similarity measures used may not be the best for studying corpus homogeneity and similarity.

The results obtained from this first attempt to evaluate the homogeneity measure confirm the hypothesis that homogeneous corpora need a smaller training set than heterogeneous corpora to achieve a certain degree of accuracy. But the methodology we have used is still too unrefined to produce clear results.

As far as the methodology is concerned, the aspect requiring further attention is the use of some kind of feature selection in order to analyze just the more distinctive features. At the moment we are considering different types of feature selection. We also want to use a fourth text-similarity measure - perplexity.

As far as evaluation is concerned, other experiments to study the validity and reliability of the measures proposed to quantify homogeneity and similarity are needed. Because the main applications of the two measures are in NLP, they should still be tested in relation to a NLP task. We would like to consider a different system from text classification and also possible ways of combining the two measures.

8. References

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

Douglas Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–41.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: Classification of genres in text. In *Workshop on human language technology and knowledge management*, Toulouse, France. ACL 2001.

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

Helka Folch, Serge Heiden, Benoit Habert, Serge Fleury, Gabriel Illouz, Pierre Lafon, Julien Nioche, and Sophie Prevost. 2000. Typtex: Inductive typological text classification by multivariate statical analysis for nlp systems tuning/evaluation. In *Second International Conference on Language Resources and Evaluation*, pages 141–148, Athens, Greece. Lrec 2000.

k. Hofland and S. Johansson. 1982. *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities.

Gabriel Illouz, Benoit Habert, Helka Folch, Serge Heiden, Serge Fleury, Pierre Lafon, and Sophie Prevost. 2000.

Typtex: Generic features for text profiler. In *Content-Based Multimedia Information Access*, Paris, France. RIAO'2000.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *Procceedings of the 37th Annual Meeting of the Association for Computational Linguistic and the 8th Conference of the European Chapter of the Association for Computational Linguistic*, pages 32–38, Madrid, Spain. ACL'97.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

David Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72. Special Issue on: "Using Corpora in Language Teaching and Learning.

Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling,text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Frank Owen and Ronald Jones. 1977. *Statistics*. Polytech Publishers, Stockport, UK.

Douglas Roland and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *36th Annual Meeting of the Association for Computational Linguistics*, pages 1122–1128, Montreal, Canada.

D. Scott, N. Bouayad-Agha, R. Power, S. Schulz, R. Beck, D. Murphy, and R. Lockwood. 2001. Pills: A multilingual authoring system for patient information. In *Vision of the Future and Lessons from the past. Proceeding of the 2001 AMIA Annual Symposium*, Washington DC.

Satoshi Sekine. 1997. The domain dependence of parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington D.C., USA. ACL.