

The Effects of Corpus Size and Homogeneity on Language Model Quality

Tony G. Rose¹

tgr@cre.canon.co.uk

Canon Research Centre Europe Ltd.

Surrey Research Park, Guildford, Surrey GU2 5YF UK

Nicholas J. Haddock, Roger C.F. Tucker

{njh, rcft}@hplb.hpl.hp.com

Hewlett-Packard Laboratories

Stoke Gifford, Bristol BS12 6QZ UK

Abstract

Generic speech recognition systems typically use language models that are trained to cope with a broad variety of input. However, many recognition applications are more constrained, often to a specific task or domain. In cases such as these, a knowledge of the particular topic can be used to advantage. This report describes the development of a number of techniques for augmenting domain-specific language models with data from a more general source.

Two investigations are discussed. The first concerns the problem of acquiring a suitable sample of domain-specific language data from which to train the models. The issue here is essentially on *quality*, since it is shown that not all domain-specific corpora are equal. Moreover, they can display significantly different characteristics that affect the quality of any language models built therefrom. These characteristics are defined using a number of statistical measures, and their significance for language modelling is discussed.

The second investigation concerns the empirical development and evaluation of a set of language models for the task of email speech-to-text dictation. The issue here is essentially one of *quantity*, since it is shown that effective language models can be built from very modestly sized corpora, providing the training data matches the target application. Evaluations show that a language model trained on one million words can perform better than one trained on a corpus of over 100 times that size.

1. Introduction

The development of robust speech recognition technology offers great potential for the design of improved interfaces to a wide range of applications. The current project concerns the development of such an application: the speech-to-text dictation of email messages. The work makes use of the Abbot recogniser, which is a connectionist/HMM continuous speech recognition system developed by the Connectionist Speech Group at Cambridge University. It is designed to recognise British English and American English, clearly spoken in a quiet acoustic environment (Hochberg et al., 1994).

The Abbot system is available with a vocabulary of 20,000 words, which means that anything spoken outside this vocabulary cannot be recognised (and therefore will be recognised as another word or series of words). The vocabulary and grammar (LM) were optimised for the task of reading from a North American Business newspaper, in this case the Wall Street Journal. Some 227 million words of training text were used in building this LM and it is widely used throughout the speech community. However, despite the size of the original training corpus, this LM was clearly not designed for the specific task of email dictation, so its performance is likely to be sub-optimal. However, a new vocabulary and LM

¹ This work was completed at HP Labs during a previous appointment funded by the Royal Academy of Engineering.

easily be created and then substituted for the one supplied. The LMs described in this paper were 'back-off trigram' LMs (Katz, 1987), built using the CMU SLM toolkit (Rosenfeld, 1994).

2. Corpus Acquisition

2.1 The form of email messages

In order to build a LM for the task of email dictation, it is necessary to acquire a corpus of suitable email training data. However, behind this ostensibly simple objective lie several subtle challenges. "Email", as a general term, describes a great variety of types of communication. These types are perhaps best illustrated by considering the range of functions that email messages typically provide. For example, email can be used as a medium for:

- a formal face-to-face meeting;
- a casual face-to-face chat;
- a broadcast (e.g. "Tannoy") message;
- requesting information;
- replacing an office memo;
- replacing a phone call, etc.

Clearly, the purpose of each communication can be very different, and the language used will reflect this. Furthermore, apart from the issue of domain (i.e. subject matter), the type of language used will also vary according to the role of the participants. For example, when requesting advice from a mailing list one would tend to be more formal and polite than when requesting the same advice from a friend or colleague. Consequently, it would appear that email messages vary almost as much as spontaneous, spoken dialogue. If this is indeed so, then the prospects for building effective language models for email may appear somewhat limited. Clearly, in order to move forward, it is necessary to define some limits.

The first of these concerns the quantity. What is a reasonable size for an email corpus? There are few precedents for this so the question was answered empirically, by finding a compromise between the need to acquire sufficient training data and the need to complete the acquisition phase within a reasonable space of time. However, the time taken to reach a certain quantity depends very much on the source of the data, which forms the second limit: from where should the email be acquired? A range of possibilities exists, e.g. the Internet (i.e. bulletin boards, mailing lists, email archives) or specific individuals (i.e. previously saved messages, day-by-day output).

Evidently, email acquired from the Internet exhibits a wide range of authorship, function and subject matter. In addition, downloading large quantities of text from such sources without the authors' consent may involve certain copyright issues. Clearly, the limits of the source are more easily defined if the email is restricted to the output of a group of specific individuals. However, unless the group is very large acquiring just 1 million words from their day-by-day output would be too slow to enable the acquisition to be completed within a reasonable space of time. Therefore, individuals with a large collection of previously saved messages were identified as more suitable candidates. Furthermore, a restriction that all members of this group must be employees of HPLB (Hewlett-Packard Labs, Bristol) placed a further constraint on the source. To ensure controlled authorship, only the outgoing messages of these individuals were collected.

2.2 The content of email messages

The "content" of an email message is not an easy concept to define. Evidently, the body contains much important data, but what about the other elements, e.g. headers, signatures, quoted sections, etc. - wh

roles do they play? In the case of headers, a cursory analysis revealed that they could safely be discarded since few contained any useful information. However, other email components are not quite so easily categorised, e.g.

- quoted (included) messages: these are usually referred to by the message content, but are often product of a different author;
- email 'signatures': these are often quite verbose, but rarely contribute anything to the message content;
- samples of Postscript/Latex: these were problematic, since people would often quote verbatim long passages to illustrate a point that did indeed contribute to the content of the message. However, building models from data that included such heavily marked-up text is questionable.

Since the above items were all rendered as ASCII strings, their surface form could fairly reliably be predicted and they were therefore removed from the corpus using suitably designed "filters". However, there is a further number of email attachments that are not composed of predictable ASCII strings. These include items such as shared or uuencoded files and word processor/DTP output. Evidently, such items need to be removed, but finding small fragments of such diverse data in a corpus of several million words is a non-trivial problem.

Alternatively, rather than trying to filter out "noise" from the email "signal", it is possible to adopt a converse approach, and try to identify those lines that constitute genuine English within the overall email data, which may then be retained as the true training data. It is possible to achieve this using various heuristics, e.g. "retain those lines that contain at least 90% English words". However, this approach assumes there exists some predefined vocabulary, which is somewhat tautological since the vocabulary is one of the things we seek to define in the first place.

2.3 The email collection process

A programme of email data collection took place over a period of 2-3 weeks, following the principles described above. This resulted in the acquisition of some 4 million words of email data. The donors were asked to provide both previously saved messages and intermittent day-by-day output. This was necessary since (by their very nature) saved messages tend to possess some sort of significant content and were therefore often of above average length. In contrast, much day-by-day email correspondence uses an informal dialogue that is heavily context dependent, and therefore may be no more than a single brief phrase or sentence. This was then filtered in the manner described above. The final output was a corpus of email data of 1,962,280 words (49% of the original size). This was then partitioned (95% : 5%) into training and test data.

3. Corpus adaptation

It is theoretically possible to build a LM using the tiniest of corpora. On balance, however, the 2 million words of email training data look somewhat inadequate compared to the 227 million words used for the WSJ LM. The problem is that the coverage of the n-grams is likely to be sparse and any LM so built will be degenerate since it does not reliably predict the characteristics of the source. To illustrate, consider the distribution of unigram frequencies: a mere 14,137 word types (19%) in the email corpus have frequencies of 6 or greater. Therefore, to acquire a vocabulary of just 20k words without unigrams of frequencies of 5 or less clearly requires a training corpus larger than 2 million words.

It would be highly desirable therefore if a method could be devised whereby information from a large corpus could be combined with a smaller sample of the domain-specific training data to create an optimal language model. One such approach involves augmenting a base model built from a larger, more general corpus with information from a small sample of the domain-specific language. There is evidence to suggest that this method can improve recognition performance (e.g. Rudnicky, 1995, Vergyri, 1995)

alternative approach is to use a suitable similarity metric to acquire further "email-like" training data from the larger corpus (henceforth referred to as the "background corpus"), and then build a new language model from the combined text. This approach offers interesting possibilities regarding the development of a general methodology for corpus adaptation, by attempting to "grow" a suitable corpus of training data for any domain using only a small sample as a "seed". A number of ways to implement this technique have been developed. Broadly speaking, they fall into two categories: "top-down" methods and "bottom-up" methods.

3.1 The top-down approach

At its simplest, this approach involves a combination of manual inspection and regular expression searching to identify those parts of the background corpus that contain suitable material. It relies on a good classification scheme and reliable organisation of the background corpus. The British National Corpus (BNC) is a suitable example, since it contains 100 million words of modern English, both spoken and written, sampled from the widest range of materials. It is annotated with part-of-speech codes, and SGML-encoded according to the Text Encoding Initiative's Guidelines (Burnard, 1995). It is therefore possible to use the SGML tags to identify suitable texts. For example, extracting 10 million words of text for a domain such as World Affairs is trivially easy, since domain information is encoded in the header of each individual file (of which there are over 4,000).

Since much HP email concerns the computing business, and the BNC classifies computing as a branch of Applied Science, it would appear that the 10 million words from the Applied Science section of the BNC may prove sufficiently similar. Likewise, the 10 million words classified as Commerce and Finance may also prove suitable. The effect of such an addition would be to increase the size of the training corpus from 10 million words to 22 million, which constitutes an increase of 1100%.

However, methods such as this cannot be justified by subjective judgement and anecdotal evidence. What is required is an objective measure that reliably identifies which of the domains in the BNC is most similar to HP email. There may be many standard statistical techniques for measuring the degree of similarity of two data sets, but not all are suitable for the task of comparing corpora (Church et al., 1991). For example, some assume a normal distribution, which is clearly inappropriate for textual data. What is needed therefore is a test that makes few assumptions about the distributions of the underlying data, but provides a directly usable measure of similarity. One such test is the rank correlation, using Spearman's ρ :

The assumptions behind rank correlation are few. It measures the degree of monotonic association between two rankable variables. The distribution of ρ as normal (mean 0, variance $1/(N-1)$, assuming independence) is asymptotic for large enough samples, and does not make any assumptions about normality. This test was therefore applied to the word frequency lists of each of the domains in the BNC and the email corpus, to identify which corpora were most similar. The correlation with the BNC as a whole was also measured. All calculations were based on Spearman's ρ , where D^2 denotes the sum of the squares of the differences between the ranks of each pair of word types, and N the number of rank pairs:

$$\rho = 1 - \frac{6D^2}{N^3 - N}$$

It is known that a sublanguage corpus can have very different characteristics to a general corpus (Biber, 1993), yet it is not obvious how the position on this scale of a given corpus can be assessed. Consequently, it is necessary to determine the *homogeneity* of a corpus prior to performing any similarity measures, since it is not clear what a measure of similarity would mean if a homogeneous corpus was being compared with a heterogeneous one (Kilgarriff, 1996). A homogeneity test was therefore performed on the corpus of each domain. This was calculated using the following algorithm:

1. For each domain corpus, do (times 10)
 - 2.1 Divide the corpus into two halves, by randomly placing 5k-word chunks in one of subcorpora;
 - 2.2 Produce a word frequency list (wfl) for each subcorpus;
 - 2.3 Calculate the rank correlation between the two subcorpora;
3. Calculate the mean and standard deviation of r .

	<i>Ar</i>	<i>As</i>	<i>Bt</i>	<i>Cf</i>	<i>Im</i>	<i>Le</i>	<i>Np</i>	<i>Ss</i>	<i>Un</i>	<i>Wa</i>	<i>BNC</i>
Ar	0.789 0.001										
As	0.255 72127	0.758 0.001									
Bt	0.408 69932	0.238 73839	0.581 0.002								
Cf	0.340 70648	0.351 70452	0.291 72970	0.730 0.001							
Im	0.404 67604	0.159 73786	0.340 70807	0.215 72800	0.887 0.001						
Le	0.459 67013	0.284 71281	0.310 71768	0.337 70407	0.407 67615	0.824 0.001					
Np	0.122 75822	0.315 71596	0.150 76151	0.161 75339	0.061 76628	0.145 74952	0.662 0.002				
Ss	0.409 68263	0.342 70405	0.422 69756	0.470 68023	0.278 70475	0.325 69796	0.200 74053	0.812 0.001			
Un	0.273 74110	0.161 76754	0.174 77136	0.229 75423	0.244 74635	0.357 72373	0.032 79904	0.226 75161	0.486 0.002		
Wa	0.423 68005	0.280 71298	0.395 70189	0.432 68404	0.311 69838	0.395 68296	0.130 75399	0.469 66921	0.284 73981	0.865 0.001	
BNC	0.611 63522	0.497 66732	0.505 67920	0.541 66189	0.578 63755	0.605 63439	0.307 71615	0.609 63779	0.381 72033	0.653 62450	0.687 0.001
Email	0.012 81648	0.093 79745	-0.026 82897	0.055 80884	-0.062 83083	-0.003 81908	-0.032 82719	0.035 81143	-0.066 84338	-0.012 82015	0.073 80085

Table 1. Similarity and homogeneity of BNC domains and email

Table 1 shows both sets of results. The homogeneity values are across the diagonal, with mean standard deviation shown in each cell. The other cells show the rank correlation (r) and the value c . "BNC" refers to the complete corpus. The subdomains are labelled as follows:

- As: Applied Science
- Ar: Arts
- Be: Beliefs & thought
- Cf: Commerce & Finance

Im: Imaginative
Le: Leisure
Np: Natural & pure science
Ss: Social science
Un: Unclassified
Wa: World affairs.
BNC: the whole of the BNC
Email: the 2 million word email corpus

For large samples such as these the rank correlation coefficient has a normal distribution with mean 0 and variance $1/(n-1)$ where n is the number of common words. Although the significance of the correlation is not in doubt, the differences are highly significant too. The difference between two rank correlation coefficients will be normally distributed with mean 0. The maximum possible value for the standard deviation is $(1/\sqrt{(n_1-1)})+(1/\sqrt{(n_2-1)})$ where n_1, n_2 are the two common vocabulary sizes. Any difference greater than about 0.03 is therefore significant, and there are many pairs for which this is true. It is therefore possible to rank the rank correlations, and hence the BNC domains.

Evidently, the strongest correlation with the email corpus is from As (Applied Science). Interestingly, this figure is higher than that between email and the whole BNC. The second highest domain correlation is with Cf (Commerce & Finance). This agrees with intuitions based on a manual inspection of the content of the email corpus. The table also shows a polarity of the BNC - the "arts" domains at one pole attracting each other (e.g. Ar:Bt = 0.408) but repelling the sciences (e.g. Im:As = 0.159). Similarly, the sciences attract each other (e.g. Np:As = 0.315). In the middle are domains such as World Affairs, Social Sciences & Commerce & Finance that correlate with both poles to varying degrees. Moreover, the Email corpus really stands out on its own, having a very poor correlation with the others (in many cases it is negative). This suggests that even if the most strongly correlated domains are chosen, it is difficult to justify augmenting the email corpus with texts selected from the BNC using this method. Table 1 also shows the results of the homogeneity tests. Email is by far the most heterogeneous, more so even than the "Unclassified" section of the BNC (!) This brings into question the results of the similarity calculations in which the email corpus was involved, and mitigates further against the strategy of augmenting the email corpus with texts selected using the top-down method.

These results also provide insight into the relationship between homogeneity and language model quality. A common measure of LM quality is *perplexity* (PP), which can be thought of as a measure of the "branching factor" (i.e. the average size of the set of words between which a speech recogniser must choose) when transcribing a single word of the spoken text. PP thus measures the recognition difficulty of the text relative to the given LM, and is measured by applying the model to a sample of test data. Consequently, a LM derived from a heterogeneous corpus *should* have a higher perplexity than an equivalent one derived from a more homogenous corpus. However, homogeneity is defined here as a measure of unigram distributions, whereas perplexity is usually calculated using n -grams (where n is usually ≤ 3), so it is not clear to what extent the two measures would be related.

3.2 The bottom-up approach

The top-down approach assumes that the BNC classification system is perfect, in that each text classified as belonging to a certain domain really belongs in that domain. However, this is ultimately a subjective judgement, and frequently more than one classification is possible or even preferable (Lewis, 1992). Moreover, it is often the case that texts from the same medium are more similar to each other than texts from the same domain (e.g. a journal paper on computing may be more similar to a journal paper on geology than an item from a popular computing magazine, because the "content" features are lost among the much more salient "genre" features). Besides, no classification system is 100% reliable, so techniques

that are based on them will inherit this uncertainty. Furthermore, domains such as Applied Science are very coarse-grained: they contain many more types of material than just those of computing. Even if corpora are subdivided to a further level of classification they still suffer the same problem, albeit at a finer level of detail.

An alternative strategy is to work in a "bottom-up" direction. In this approach, a similarity metric is used to find and extract related material from the background corpus, regardless of the top-down classification. This method may not be as structured as the previous approach, but it is more robust in that it involves minimal manual intervention and does not rely on correct organisation or SGML tagging of the background corpus. Moreover, it will not "miss" material that is classified under an unexpected domain or method but is otherwise suitable. The success of this approach depends on the use of a reliable similarity metric (even more so than the top-down approach, since it is now being applied to each of the 4,000+ files in the BNC rather than the 10 domain-based collections). Using this statistic to find texts that are similar to email in the BNC could be achieved using the following algorithm:

1. Create a wfl for the email corpus.
2. For each individual text in the BNC, do:
 - 2.1 Create the wfl for the BNC text
 - 2.2 Create a contingency table from the 2 wfls (ignoring function words)
 - 2.3 Calculate the number of common words N and the rank correlation r
3. Store the filename, title, N and r in RESULTS
4. Output the RESULTS sorted on the value for r .

Although the rank correlation may be applied to the wfls regardless of their content, it was found empirically that performance improved if function words were excluded from the contingency table. A stop list of 241 function words was therefore applied in Step 2.2 of the above algorithm. The algorithm was run on the entire BNC (i.e. each of its 4,000+ files). The output was a list of the files sorted according to the value of r . The top and bottom 10 texts on this list are as follows:

/BNC/1.0/H/HA/HAC	ARTICLES FROM PRACTICAL PC NOV 95	7558	0.606
/BNC/1.0/J/J0/JOV	ELECTRONIC INFORMATION RESOURCES AND THE HI	5393	0.592
/BNC/1.0/C/CT/CTX	WHAT PERSONAL COMPUTER	4337	0.581
/BNC/1.0/F/FT/FTB	WHAT PERSONAL COMPUTER: THE ULTIMATE GUIDE	4513	0.577
/BNC/1.0/G/G0/G00	MISCELLANEOUS ARTICLES ABOUT DESK-TOP PUBLI	5228	0.572
/BNC/1.0/C/CB/CBU	ACCOUNTANCY	6053	0.567
/BNC/1.0/C/CB/CBX	ACCOUNTANCY	5902	0.557
/BNC/1.0/K/KR/KRG	IDEAS IN ACTION PROGRAMMES (03) -- AN ELECT	3231	0.556
/BNC/1.0/H/HR/HRD	MULTIMEDIA IN THE 1990S	3914	0.554
/BNC/1.0/E/EE/EEB	PEOPLE IN ORGANISATIONS	3199	0.553
.....			
/BNC/1.0/F/FU/FUS	RESULTS OF PRSTATECTOMY SURVEY -- AN ELECTR	239	0.096
/BNC/1.0/J/J5/J5B	ECOVER BIO-DEGRADABLE HOUSEHOLD CLEANING PR	248	0.088
/BNC/1.0/H/H4/H4P	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRAN	137	0.087
/BNC/1.0/G/GY/GY5	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRAN	90	0.078
/BNC/1.0/K/KP/KPS	SPOKEN MATERIAL FROM RESPONDENT PAMELA2 --	21	0.063
/BNC/1.0/G/G5/G54	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRAN	17	0.052
/BNC/1.0/H/HK/HKM	ROCKWELL/THE BETTER ALTERNATIVE TO THE FLAT	215	0.050
/BNC/1.0/F/FD/FDE	THE WEEKLY LAW REPORTS 1992 VOLUME 3	196	0.045
/BNC/1.0/G/G5/G5A	AUCTION ROOMS -- AN ELECTRONIC TRANSCRIPTIO	186	0.000
/BNC/1.0/J/JJ/JJ2	BRISTOL UNIVERSITY -- AN ELECTRONIC TRANSCR	18	-0.259

Each line shows the filename, the title of the text, the number of common words and the value for r . At first glance, the results appear to be intuitively satisfying. Of the top ten texts, six have titles that are clearly related to computing, including all of the top five. The remaining four could arguably be classified as Commerce & Finance (which was identified as the second most similar domain to email). However, a suitable title is no guarantee of suitable contents. As far as can reasonably be expected, the list should constitute a fair and accurate reflection of the contents of each text. Of course, the whole point of

approach is to develop techniques that do not rely on ambiguous manual annotations such as title domain, so the presence of suitable titles is merely an initial indication of success.

One way of evaluating this result is to go through the list and calculate the mean rank of the "Computergram International" texts, which are typical of the sort of texts this technique should identify being similar to the email corpus. If the technique is working perfectly, the mean rank should be 31. If it is completely random, the mean rank would be 2062. It transpires that the mean rank is 959.85 (std dev 524.44). Clearly, this result is better than chance, but far from significant. One of the main reasons for this was a tendency to sometimes give high scores to texts that were actually too short to constitute reliable samples (the BNC attempts to maintain a standard sample size but this is not always possible). A logical modification was therefore to ignore those texts for which the number of common words was below a certain threshold. A number of thresholds were investigated, and the optimum value (determined empirically) was around 1,370 words. However, even with this modification, the mean rank remained as high as 818.41 (std dev = 407.81). It is possible to reduce this value still further, but only by compromising the overall recall value (i.e. genuine texts are eliminated along with the "noise").

However, there is a more fundamental limitation to the above methodology. The rank correlation statistic compares differences in rank, ignoring absolute value (which can be significant). To illustrate, consider the case where the word "of" is ranked 3 in one corpus and 6 in another. This is a very important difference. Conversely, if "banana" is ranked 10,000 in one corpus and 100,000 in another, this is a very insignificant difference. But the difference of ranks for "of" = 3, for "banana" = 90,000. Clearly this technique is missing something important. Consequently, it was decided to investigate an alternative measure: the Loglikelihood Ratio Statistic.

The Loglikelihood Ratio, G^2 , is a mathematically well-grounded and accurate method for calculating how "surprising" an event is (Dunning, 1993). This is true even when the event has only occurred once (as is often the case with linguistic phenomena). It is an effective measure for the determination of domain specific terms (e.g. Daille, 1995) and can be also used as a measure of corpus similarity. In the case where two corpora are being compared, it is possible to calculate the G^2 statistic either for single words (using a 2x2 contingency table) or for a vocabulary of N words (an Nx2 table). The analysis of the 4,000 BNC files was therefore repeated using the Loglikelihood (instead of rank correlation) as the similarity measure. This produced the following top and bottom 10 texts:

/BNC/1.0/H/H4/H4L	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION	23226	108.897
/BNC/1.0/G/G5/G54	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION	23226	244.251
/BNC/1.0/H/H5/H58	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION	23228	318.442
/BNC/1.0/F/F7/F78	STAFF MEETING -- AN ELECTRONIC TRANSCRIPTION	23226	358.087
/BNC/1.0/H/H5/H5B	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION	23230	385.826
/BNC/1.0/G/G4/G4Y	MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION	23231	419.826
/BNC/1.0/K/KN/KNU	SPOKEN MATERIAL FROM RESPONDENT 716 -- AN ELECTRONIC	23227	425.806
/BNC/1.0/J/JJ/JJJ	BRISTOL UNIVERSITY -- AN ELECTRONIC TRANSCRIPTION	23231	526.359
/BNC/1.0/A/A9/A9B	GUARDIAN, ELECTRONIC EDITION OF 19891210; APPSCI MAT	23232	532.956
/BNC/1.0/H/HK/HKC	FREEMANS	23234	547.160

/BNC/1.0/C/CB/CBG	TODAY	32658	351001.583
/BNC/1.0/K/K5/K5D	&LSQB;UNCATALOGUED TEXT SAWLDA&RSQB;	34572	352223.071
/BNC/1.0/H/HU/HU4	GUTS1\$1\$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY	29557	367084.662
/BNC/1.0/H/HW/HWS	GUTS1\$1\$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY	29531	367801.892
/BNC/1.0/H/HU/HU2	GUTS1\$1\$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY	29356	379109.473
/BNC/1.0/H/HU/HU3	GUTS1\$1\$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY	29500	382579.770
/BNC/1.0/H/HH/HHX	HANSARD PROCEEDINGS 19951&NDASH;92 SESSION	30854	383471.711
/BNC/1.0/K/K9/K97	LIVERPOOL ECHO \$1\$21DAILY POST\$1\$2NOVEMBER 1992 --	38865	402754.606
/BNC/1.0/C/CR/CRM	NATURE	36867	463853.336
/BNC/1.0/H/HH/HHV	SELECTION FROM HANSARD 19951&NDASH;1992	30389	469043.781

Each line shows the filename, the title of the text, the length of the contingency table and the value for G^2 . These are sorted in ascending order since comparing two identical documents would produce a G^2 of zero.

A brief inspection of the titles of the documents at the top of the list would indicate that the metric has produced an improvement. Moreover, it transpires that the mean rank of the CI texts is now 1171.98, std dev = 178.54. However, as before, the number of common words is very small for some of the 1. Therefore, the filter was applied to ignore cases where there were fewer than 1,370 words in common. This produced a mean rank of 125.15 (std. dev. = 75.62), which is significantly lower than that produced by the rank correlation (mean rank = 818.41, std. dev. = 75.621).

So despite the absence of apparently suitable candidates in the top 10, the overall accuracy of the technique (measured by the mean rank of the 61 CI texts) is higher. The G^2 statistic appears to be suitable for this type of data since it uses the actual frequency values for the words in the wfls, rather than just their ranks. Other independent sources indicate that the G^2 produces results that appear to correspond reasonably well with human judgement (Daille, 1995).

However, both the rank correlation and Loglikelihood Ratio both make use only of unigram information. Clearly, much of the information that humans use to measure textual similarity is found not (solely) in individual word frequencies (unigrams), but rather in the way they combine (n-grams). The logical step is therefore to compare word bigrams (or trigrams) instead of just unigram data. A variation on this would be to compare texts using the Loglikelihood applied to bigrams that are not necessarily adjacent, i.e. counting occurrences of word1 and word2 within a limiting distance of each other. Indeed, such methods have been previously used for actually building the LMs themselves, and have been successfully applied to both speech (Rose & Lee, 1994) and handwriting data (Rose & Evett, 1995). Counting word occurrences within a limited window would be smoother than using strict bigrams and consequently less affected by the problems caused by sparse data (which are inevitable when small, individual text files are compared). Another interesting possibility is to use the LM itself as the similarity metric. From an information theoretic point of view, entropy is a measure of a corpus's homogeneity, and the cross-entropy between two corpora is a measure of their similarity (Charniak, 1993). After all, when a LM is applied to a test text to produce a perplexity score, this value is a measure of the cross-entropy which reflects how well the LM predicts the words in the text. So if a LM is trained on text that is very similar to the test text, it should predict the test data well and the perplexity should be low. Conversely, if the test text is different from the training text, then the perplexity will be high. The perplexity score can therefore be used to measure textual similarity. Moreover, it has the advantage of doing so by considering (typically) unigram, bigram and trigram data. Indeed, this method has already been successfully used within the development of a similarity-based Internet search agent, and preliminary findings indicate that perplexity is indeed an effective corpus similarity measure (Rose & Wyard, 1997).

However, the use of such an approach is not entirely beyond question. Firstly, the LM is being used as a representation of a training text against which similarity is to be judged, and yet it is, by definition, untrained and therefore degenerate. Secondly, the method by which similarity is measured should ideally be independent to the method by which success is evaluated. To use perplexity both as a similarity measure and an evaluation metric implies a certain amount of circular reasoning. However, the use of iterative techniques is not totally without precedent within the LM community. Several researchers have reported the successful improvement of LMs using techniques that iteratively tune the parameters using new samples of training data (e.g. Jelinek, 1990). So, this approach may transpire to be sufficiently well principled to merit further investigation.

4. Language model quality

A LM is built by collecting trigram, bigram & unigram data from a training corpus. However, it is not always desirable to store all of this data. Thresholds can be set such that some of the lower frequency n-grams are discarded. For example, a trigram cut-off of 5 implies that all the trigrams with frequencies of 5 or fewer in the training data are not used in building the model. Setting lower thresholds allows the n

to focus on more frequent events, and produces a proportionately smaller model. The LMs described in this paper were built using the CMU SLM toolkit (Rosenfeld, 1994) which facilitated the construction of a variety of LMs representing a range of different settings for each of the pertinent parameters.

The first of these was the Email LM. This was constructed using a vocabulary of 20,000 words that were derived directly from the email training data. The bigram and trigram cut-offs were both set to zero. The second LM was built from the whole of the BNC, using the same vocabulary as the Email LM (in order to ensure consistency). So although their n-grams had been based on general English rather than Email, the vocabulary was derived from the Email data. For comparison therefore, a third BNC LM was built, using a vocabulary derived directly from the BNC (rather than email). This allowed the comparative evaluation of the contribution of vocabulary vs. n-grams to the LM effectiveness (measured using both perplexity and word error rate). Due to memory constraints it was not possible to build the BNC models with cut-offs lower than 2-2. The fourth LM investigated was the 20k WSJ LM that is available from the Abbot site at Cambridge University.

The standard measure by which LMs are assessed is by calculating their perplexity using a sample of test data. This process is usually performed off-line, i.e. independently of the speech recogniser for which the models are intended. For the models described above, testing was performed using the CMU toolkit, by applying each LM to a sample of 10,000 words from the transcriptions of a database of video mail messages, developed by Cambridge University as part of their "Video Mail Retrieval using Voice project (Jones et al., 1994). Evidently, this data is not actually spoken email, but its domain and genre are nevertheless closely related to email. Unfortunately, it was not possible to calculate the PP of the WSJ LM due to the absence of a readily available version in the correct format.

A second evaluation method is to integrate the LM with the speech recogniser and test the combined system using recorded speech data. The models can be interchanged between trials, allowing comparative evaluation by measuring the word error rate (WER) produced by each model. More precisely, the error rates are measured using two standard metrics, percentage correct and accuracy:

$$1. \quad \%Correct = \frac{H}{N} \times 100\%$$

$$2. \quad Accuracy = \frac{(H - I)}{N} \times 100\%$$

where: H is the number of correct transcriptions (words in the utterance that are found in the transcription), D is the number of deletions (words in the utterance that are missing from the transcription), S is the number of substitutions (words in the utterance that are replaced by an incorrect word in the transcription), and I is the number of insertions (extra words in the transcription). Accuracy is more critical than %correct in that it directly penalises insertions. Deletions & substitutions reduce the value of H, since $H = N - (D+S)$.

As mentioned above, the VMR database is a collection of speech data with transcriptions (of which the latter were used in the above evaluation). The speech part contains audio files for 15 speakers, of which 10 were used in the current investigation. The Abbot recogniser was run using each combination of the speakers' data files (as input) and each of the four LMs: email, BNC with email vocabulary, BNC and the WSJ LM. The output transcriptions were assessed for %correct and accuracy using the HResults program which is part of HTK - the Hidden Markov Model Toolkit (Young & Woodland, 1993).

Table 2 shows the results of this investigation. The results for %correct and accuracy show the combined effect of the recogniser and LM. The contribution of the LM depends on its vocabulary and perplexity.

the LM changes, it produces different behaviour in the combined system and therefore different types of errors (e.g. insertions, deletions & substitutions). The net effect is that the email LM produces the highest %correct and also the highest accuracy. It is around 5% better (on both measures) than the WSJ LM. This difference is significant, considering the tiny corpus from which it was derived (2 million vs. 227 million in the case of WSJ). In between these two extremes are the two BNC LMs - the one with the email vocabulary performs slightly better (~0.5%) than the one with the BNC vocabulary.

	%Correct	Accuracy		%Correct	Accuracy
Speaker 1			Speaker 6		
email	42.32	33.23	email	52.17	42.88
BNC	41.47	32.02	BNC/email	49.53	39.76
BNC/email	41.42	31.06	BNC	48.97	39.11
WSJ	37.84	28.50	WSJ	46.36	36.65
Speaker 2			Speaker 7		
BNC/email	37.14	24.58	email	65.05	54.23
email	36.77	25.03	BNC/email	64.49	53.93
BNC	36.19	24.40	BNC	64.23	53.83
WSJ	32.90	22.11	WSJ	60.97	50.15
Speaker 3			Speaker 8		
email	44.42	39.14	email	54.70	43.44
BNC/email	44.42	37.86	BNC/email	51.40	38.74
BNC	43.34	36.74	BNC	50.99	37.64
WSJ	39.72	33.90	WSJ	47.93	35.55
Speaker 4			Speaker 9		
email	59.82	50.04	email	70.08	62.75
BNC/email	59.28	49.50	BNC/email	69.86	62.03
BNC	58.37	48.28	BNC	68.85	61.14
WSJ	56.99	46.68	WSJ	66.67	58.37
Speaker 5			Speaker 10		
BNC	76.52	71.99	email	65.91	56.14
BNC/email	75.94	70.84	BNC/email	65.83	55.38
WSJ	73.15	68.43	BNC	65.12	54.67
email	71.90	66.31	WSJ	61.58	51.13
OVERALL					
	%Correct	Accuracy	Perplexity		
email	54.92	45.85	261.58		
BNC/email	54.04	44.24	241.70		
BNC	53.40	43.71	227.54		
WSJ	50.42	40.93	N/A		

Table 2. %Correct, accuracy and perplexity of the language models

The result for the PP testing is highly revealing. As described earlier, a corpus of low homogeneity should produce a LM of higher PP than a corpus of high homogeneity. This is indeed shown to be the case, as the PP for email is 261.58 (homogeneity = 0.362), whereas the PP for the BNC is 227.54 (homogeneity = 0.687). These PP values are calculated using the 10K test data sample from the transcriptions of the V project. The higher PP value for email would tend to indicate that this is the poorer LM. However, it is clear that when used on the real spoken data, the email LM provides the lowest error rates. In other words, explanations for this centred on the vocabulary, since a higher incidence of out-of-vocabulary (OOV) words can produce a lower PP but a higher WER. However, the email LM performs better (by 0.8

correct) than the BNC/email LM even though both share the same vocabulary. Two explanations for this are possible. Firstly, there may be n-grams in the email corpus that are simply not found in the BNC (even though the BNC is 50 times larger). Secondly, the email LM may be better because it “wastes” its probability mass on n-grams that never actually occur in the test data. This implies that quality, as well as quantity, is a major factor in training effective LMs. Further PP testing, possibly using the complete transcriptions of the VMR data is necessary to clarify this issue.

Evidently, the choice of vocabulary also makes an important contribution. The BNC LM with the email vocabulary performs better (by 0.64% correct) than the BNC LM with the BNC vocabulary, so clearly the email vocabulary provides better coverage of the test data. In fact, it is possible to directly compare the OOV rates with the performances shown above: the BNC LM with the email vocabulary has an OOV rate of 1.16% on the VMR data, and a %correct of 54.04. By contrast, the BNC LM with the BNC vocabulary has an OOV rate of 1.69% and a %correct of 53.40. These figures suggest that an increase in OOV rate of 0.56% leads to a reduction in %correct of 0.64%, or, in other words, a 1% increase in OOV rate produces a reduction in %correct of around 1.14%. Interestingly, this figure correlates extremely well with the results of a similar experiment performed by Rosenfeld (1995), who found that a 1% increase in the OOV rate can lead to a 1.2% increase in the word error rate.

5. Conclusions

The analysis of the corpora has provided several revealing insights. Firstly, it is necessary to determine the homogeneity of a corpus prior to performing any similarity measures, since it is not clear what a measure of similarity would mean if a homogeneous corpus was being compared with a heterogeneous one. A methodology for calculating homogeneity has been described and the accuracy and usefulness of this is further described in Kilgarriff (1997).

Clearly, the email corpus is highly heterogeneous. This means it is particularly prone to “burstiness” and unpredictability, which affects all levels of n-grams (including unigrams). This may be due in part to the particular training corpus used, but it is more likely to be inherent to the medium, since email can fulfil many communicative functions. It therefore exhibits a level of diversity surpassed perhaps only by spontaneous speech. Investigation of the spoken part of the BNC is therefore suggested as an area for further work.

To a certain extent, the apparent heterogeneity of the email undermines the results of any similarity measures applied to this corpus. Nevertheless, the extent to which the email is unlike all the other BNC domains is quite apparent and therefore mitigates any unprincipled approaches to corpus augmentation using crude, top-down techniques that involve complete domains taken from the BNC. Consequently, the best way to acquire more email data appears to be either: (a) to instigate a further collection initiative, (b) to use more sophisticated bottom-up methods, or (c) to use self-organising adaptation techniques (see Clarkson and Robinson, 1997). The similarity metric used in (b) must be chosen carefully. Although both Loglikelihood and rank correlation metrics both produce results that can look intuitively plausible, the former merely underlines the need for an objective, thorough evaluation method. Loglikelihood appears to be the more principled of the two measures, and it is suggested that this offers the greater potential.

The results of the language modelling exercise provide clear evidence that it is possible to build effective LMs from small corpora. The email LM outperformed the other LMs on real spoken data (albeit taken from a technical, “email-like” domain) for eight of the ten speakers. This is significant, considering that the other LMs were trained on corpora that were several times larger. This effect can be mainly attributed to the source of the n-grams and the extent to which the larger LMs “waste” probability mass on n-grams that never actually occur in the test data. Other researchers also have investigated methods for adapting large, general LMs using data from a small domain corpus and have found merit in simply building

smaller LM directly from the domain corpus. For example, Ueberla (1997) observed that improvements gained by using adaptation techniques compared to simply "starting from scratch" on domain data become quite small when several tens of thousands of words of domain data are available (Since the email corpus is almost 2 million words it clearly meets this criterion.) It is interesting also that this threshold is seen to vary according to the level of similarity between the adaptation (domain specific) corpus and the background (general) corpus.

It is also possible to adapt LMs *dynamically*, using cache-based methods (e.g. Kuhn & de Mori 1990) evidence suggests that this may prove the more effective approach (Matsunaga et al., 1992). It is clear that email is highly heterogeneous and therefore inherently unpredictable. Attempting to model this by static means can thus produce only limited success. By contrast, a dynamic LM would adapt to the current input and update its probabilities accordingly. However, dynamic LMs still need a set of static, baseline probabilities, so the email LM may present the best starting point for this.

6. References

- Biber, D. (1993) "Using register-diversified corpora for general language studies", *Computational Linguistics*, 19, No. 2.
- Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., P. (1989) "A statistical approach to machine translation", Technical Report, IBM Research Division.
- Burnard, L. (1995) *Users Reference Guide for the British National Corpus*, Oxford University Computing Services.
- Charniak, E. (1993) "Statistical Language Learning", MIT Press, Cambridge, Mass.
- Church, K., Hanks, P., Hindle, D. and Gale, W. (1991) "Using statistics in lexical analysis", in Zernil (Ed.) "Lexical Acquisition: Using On-Line Resources to Build a Lexicon", LEA, NJ.
- Clarkson, P.R. & Robinson, A.J. (1997) "Language model adaptation using mixtures and an exponent decaying cache", Proceedings of ICASSP, Munich, Germany.
- Daille, B. (1995) "Combined approach for terminology extraction", Technical Report 5, UCF Lancaster University.
- Dunning, E. (1993) "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, 19, No. 1.
- Hochberg, M., Robinson T. & Renals S. (1994) "Large vocabulary continuous speech recognition using a hybrid connectionist HMM system", Proc. of ICSLP, pp. 1499-1502.
- Jelinek, F. (1990) "Self-organized language modeling for speech recognition", in Waibel and Lee (Eds) *Readings in Speech Recognition*, Morgan Kaufmann, San Mateo, CA.
- Jones, G., Foote, J., Sparck Jones, K. & Young, S. (1994) "Video mail retrieval using voice", Technical Report 335, Cambridge University Computer Laboratory.
- Katz, S. (1987) "Estimation of probabilities from sparse data", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-35.

Kilgarriff, A. (1996) "Which words are particularly characteristic of a text?" in Rose & Evett (Eds) *Language Engineering for Document Analysis & Recognition*, AISB Workshop proceedings.

Kilgarriff, A (1997) "Using word frequency lists to measure corpus homogeneity and similarity between corpora", *Proceedings of the Fifth Workshop on Very Large Corpora*, Hong Kong.

Kuhn, R. & De Mori, R. (1990) "A cache-based natural language model for speech recognition" *IEEE Trans. on PAMI*, 12(6), pp. 570-583.

Lewis, D. (1992) "Text representation for intelligent text retrieval: a classification-oriented view", in Jacobs "Text-based Intelligent Systems", LEA Publishers, Hillsdale, NJ.

Matsunaga, S., Yamada, T. & Shikano, K. (1992) "Task adaptation in stochastic language models for continuous speech recognition", *Proc. ICASSP Vol. 1*, pp. 165-168.

Rose, T.G. & Evett, L. (1995) "The use of context in cursive script recognition", *Machine Vision and Applications*, Springer International.

Rose, T.G. & Lee, M (1994) "Language modelling for large vocabulary speech recognition", *Proc. IOE Meeting on LVSR*, Cambridge, England.

Rose, T.G. & Wyard, P.J. (1997) "A similarity-based agent for Internet searching", *Proceedings of RIAO'97 - Computer-assisted Searching on the Internet*, Montreal, Canada.

Rosenfeld, R. (1994) "The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARP/CSR Evaluation", *Proceedings of the Spoken Language Technology Workshop 1995*, Austin (TX).

Rosenfeld, R. (1995) "Optimizing lexical and n-gram coverage via judicious use of linguistic data", *Proc Eurospeech 95*.

Rudnicky, A. (1995) "Language modeling with limited domain data", *Proceedings of the ARP/Workshop on Spoken Language Technology*, Morgan Kaufmann, San Mateo, pp. 66-69.

Ueberla, J.P. (1997) "Domain adaptation with clustered language models", *Proceedings of ICASSP Munich*, Germany.

Vergyri, D. (1995) Unpublished web page <http://www.clsp.jhu.edu/~dvergy/bleaching.html>

Young S. & Woodland P. (1993) "HTK: Hidden Markov Model Toolkit V1.5 User Manual", Cambridge University Engineering Dept. and Entropic Research Labs Ltd.