A very very large corpus doesn't always yield reliable estimates

James R. Curran and **Miles Osborne**
Institute for Communicating and Collaborative Systems
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW
United Kingdom
{jamesc,osborne}@cogsci.ed.ac.uk

Abstract

Banko and Brill (2001) suggested that the development of very large training corpora may be more effective for progress in empirical Natural Language Processing than improving methods that use existing smaller training corpora.

This work tests their claim by exploring whether a very large corpus can eliminate the sparseness problems associated with estimating unigram probabilities. We do this by empirically investigating the convergence behaviour of unigram probability estimates on a one billion word corpus. When using one billion words, as expected, we do find that many of our estimates do converge to their eventual value. However, we also find that for some words, no such convergence occurs. This leads us to conclude that simply relying upon large corpora is not in itself sufficient: we must pay attention to the statistical modelling as well.

1 Introduction

The quantity and reliability of linguistic information is primarily determined by the size of the training corpus: with limited data available, extracting statistics for any given language phenomenon and its surrounding context is unreliable. Overcoming the sparse distribution of linguistic events is a key design problem in any statistical NLP system.

For some tasks, corpus size is no longer a limiting factor, since it has become feasible to acquire homogeneous document collections two or three orders of magnitude larger than existing resources.

Banko and Brill (2001) report on confusion set disambiguation experiments where they apply relatively simple learning methods to a one billion word training corpus. Their experiments show a logarithmic trend in performance as corpus size increases without performance reaching an upper bound. This leads them to believe that the development of large scale training material will yield superior results than further experimentation with machine learning methods on existing smaller scale training corpora.

Recent work has replicated the Banko and Brill (2001) results on the much more complex task of automatic thesaurus extraction, showing that contextual statistics, collected over a very large corpus, significantly improve system performance (Curran and Moens, 2002). Other research has shown that query statistics from a web search engine can be used as a substitute for counts collected from large corpora (Volk, 2001; Keller et al., 2002).

To further investigate the benefits of using very large corpora we empirically analyse the convergence behaviour of unigram probability estimates for a range of words with different relative frequencies. By dramatically increasing the size of the training corpus, we expect our confidence in the probability estimates for each word to increase. As theory predicts, unigram probability estimates for many words do converge as corpus size grows.

However, contrary to intuition, we found that for many commonplace words, for example `tightness`, there was no sign of convergence as corpus size approaches one billion words. This suggests that for at least some words, simply using a much larger corpus to reduce sparseness will not yield reliable estimates. This leads us to conclude that effective use of large corpora demands, rather than discourages, further research into sophisticated statistical language modelling methods. In our case, this means adding extra conditioning to the model. Only then could we reasonably predict how much training material would be required to ameliorate sparse statistics problems in NLP.

The next section briefly introduces the relevant limit theorems from statistics. Section 3 describes our experimental procedure and the collection of the billion word corpus. Section 4 gives examples of words with convergent and non-convergent behaviour covering a range of relative frequencies. We conclude with a discussion of the implications for language modelling and the use of very large corpora that our results present.

2 Theoretical Convergence Behaviour

Standard results in the theory of statistical inference govern the convergence behaviour and deviance from that behaviour of expectation statistics in the limit of sample size. The intuitive "Law of Averages" convergence of probabilities estimated from increasingly large samples is formalised by the *Law(s) of Large Numbers*. The definition[1] given in Theorem 1 is taken from Casella and Berger (1990):

Theorem 1 (Strong Law of Large Numbers) *Let* $X_1, X_2, X_3, \ldots$ *be i.i.d. random variables with* $\mathrm{E}X_i = \mu$ *and* $\mathrm{Var}\,X_i = \sigma^2 < \infty$, *and define the average* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. *Then, for every* $\varepsilon > 0$*:*

$$P\left(\lim_{n\to\infty} |\overline{X}_n - \mu| < \varepsilon\right) = 1 \tag{1}$$

The Law of the Iterated Logarithm relates the degree of deviance from convergent behaviour to the variance of the converging expectation estimates and the size of the sample. The definition in Theorem 2 is taken from Petrov (1995):

Theorem 2 (Law of the Iterated Logarithm) *Let* $X_1, X_2, X_3, \ldots$ *be i.i.d. random variables with* $\mathrm{E}X_i = \mu$, $\mu^2 < \infty$, *and* $\mathrm{Var}\,X_i = \sigma^2 < \infty$, *and define the average* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. *Then:*

$$P\left(\limsup_{n\to\infty} \frac{\overline{X}_n - \mu}{\sqrt{2\log\log n}} = \sigma\right) = 1 \tag{2}$$

Limit theorems codify behaviour as sample size n approaches infinity. Thus, they can only provide an approximate guide to the finite convergence behaviour of the expectation statistics, particularly for smaller samples. Also, the assumptions these limit theorems impose on the random variables may not be reasonable or even approximately so. It is therefore an open question whether a billion word corpus is sufficiently large to yield reliable estimates.

[1] There are two different standard formulations: the *weak* and *strong* Law of Large Numbers. In the weak law, the probability is converging in the limit to one (called convergence *in probability*). In the strong law, the absolute difference is converging in the limit to less than epsilon with probability 1 (called *almost sure* convergence).

Corpus	# Words
NANC	434.4 million
NANC Supplement	517.4 million
RCV1	193.0 million

Table 1: Components of the billion word corpus

3 Experiments

We would like to answer the question: how much training material is required to estimate the unigram probability of a given word with arbitrary confidence. This is clearly dependent on the relative frequency of the word in question. Words which appear to have similar probability estimates on small corpora can exhibit quite different convergence behaviour as the sample size increases.

To demonstrate this we compiled a homogeneous corpus of 1.145 billion words of newspaper and newswire text from three existing corpora: the North American News Text Corpus, NANC (Graff, 1995), the NANC Supplement (MacIntyre, 1998) and the Reuters Corpus Volume 1, RCV1 (Rose et al., 2002). The number of words in each corpus is shown in Table 1.

These corpora were concatenated together in the order given in Table 1 without randomising the individual sentence order. This emulates the process of collecting a large quantity of text and then calculating statistics based counts from the entire collection. Random shuffling removes the discourse features and natural clustering of words which has such a significant influence on the probability estimates.

We investigate the large-sample convergence behaviour of words that appear at least once in a standard small training corpus, the Penn Treebank (PTB). The next section describes the convergence behaviour for words with frequency ranging from the most common down to hapax legomena.

From the entire 1.145 billion word corpus we calculated the gold-standard unigram probability estimate, that is, the relative frequency for each word. We also calculated the probability estimates for each word using increasing subsets of the full corpus. These subset corpora were sampled every 5 million words up to 1.145 billion.

To determine the rate of convergence to the gold-standard probability estimate as the training set increases, we plotted the ratio between the subset and gold-standard estimates. Note that the horizontal lines on all of the graphs are the same distance apart. The exception is Figure 5, where there are no lines

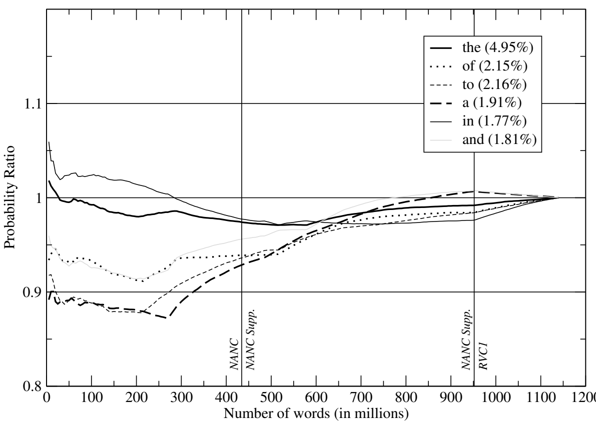

Figure 1: Estimate ratios for function words

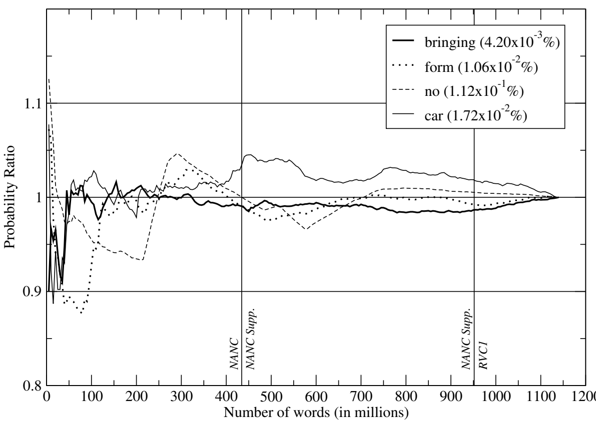

Figure 2: Ratios for accurate non-function words

because there would be too many to plot within the range of the graph. The legends list the selected words with the relative frequency (as a percentage) of each word in the full corpus. Vertical lines show the boundaries between the concatenated corpora.

4 Empirical Convergence Behaviour

Figure 1 shows the convergence behaviour of some very frequent closed-class words selected from the PTB. This graph shows that for most of these extremely common words, the probability estimates are accurate to within approximately ±10% (a ratio of 1 ± 0.1) of their final value for a very small corpus of only 5 million words (the size of the first subset sample).

Some function words, for example, `the` and `in`, display much more stable probability estimates even amongst the function words, suggesting their usage is very uniform throughout the corpus. By chance, there are also some open-class words, such as `bringing`, `form` and `crucial`, that also have very stable probability estimates. Examples of these are shown in Figure 2. The main difference between the convergence behaviour of these words and the function words is the fine-grained smoothness of the convergence, because the open-class words are not as uniformly distributed across each sample.

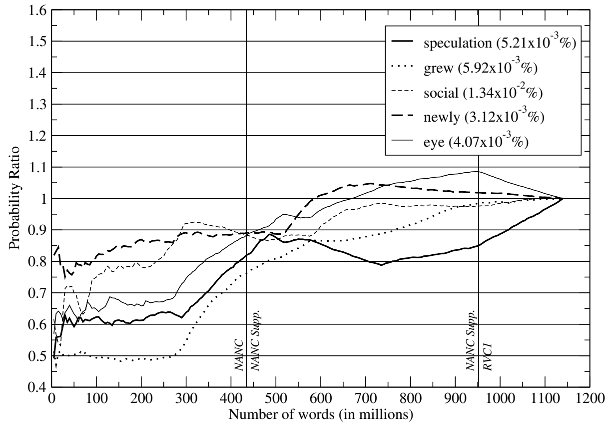

Figure 3: Ratios for commonplace words

Figure 3 shows the convergence behaviour of commonplace words that appear in the PTB between 30 and 100 times each. Their convergence behaviour is markedly different to the closed-class words. We can see that many of these words have very poor initial probability estimates, consistently low by up to a factor of almost 50%, five times worse than the closed-class words.

`speculation` is an example of convergence from a low initial estimate. After approximately 800 million words, many (but not all) of the estimates are correct to within about ±10%, which is the same error as high frequency words sampled from a 5 million words corpus. This is a result of the sparse distribution of these words and their stronger context dependence. Their relative frequency is two to three orders of magnitude smaller than the relative frequencies of the closed-class words in Figure 1.

What is most interesting is the convergence behaviour of rare but not necessarily unusual words, which is where using a large corpus should be most beneficial in terms of reducing sparseness. Figure 4 shows the very large corpus behaviour of selected hapax legomena from the PTB. Many of the words in this graph show similar behaviour to Figure 3, in that some words appear to converge relatively smoothly to an estimate within ±20% of the final value. This shows the improvement in stability of the estimates from using large corpora, although ±20% is a considerable deviation from the gold-standard estimate.

However, other words, for instance `tightness`, fail spectacularly to converge to their final estimate before the influence of the forced convergence of the ratio starts to take effect. `tightness` is an extreme example of the case where a word is seen very rarely, until it suddenly becomes very popular. A similar convergence behaviour can be seen for words with a very high initial estimate in Figure 5. The maximum decay ratio curve is the curve we would see if a word appeared at the very beginning of the corpus, but did not appear in the remainder of the corpus. A smooth decay with a similar gradient to the maximum decay ratio indicates that the word is extremely rare in the remainder of the corpus, after a high initial estimate. `rebelled`, `kilometers` and `coward` are examples of exceedingly high initial estimates, followed by very rare or no other occurrences. `extremists`, `shelling` and `cricket` are examples of words that were used more consistently for a period of time in the corpus, and then failed to appear later, with `cricket` having two periods of frequent usage.

Unfortunately, if we continue to assume that a unigram model is correct, these results imply that we cannot be at all confident about the probability estimates of some rare words even with over one billion words of material. We cannot dismiss this as an unreliable low frequency count because `tightness` occurs 2652 times in the full corpus. Thus we must look for an alternative explanation: and the most reasonable explanation is *burstiness*, the fact that word occurrence is not independent and identically distributed. So given that one billion words does not always yield reliable estimates for rare but not unusual words, it leaves us to ask if any finite number of words could accurately estimate the probability of pathologically bursty word occurrences.

5 Discussion

It is worth reflecting on why some words appear to have more bursty behaviour than others. As we would expect, function words are distributed most evenly throughout the corpus. There are also some content words that appear to be distributed evenly. On the other hand, some words appear often in the first 5 million word sample but are not seen again in the remainder of the corpus.

Proper names and topic-specific nouns and verbs

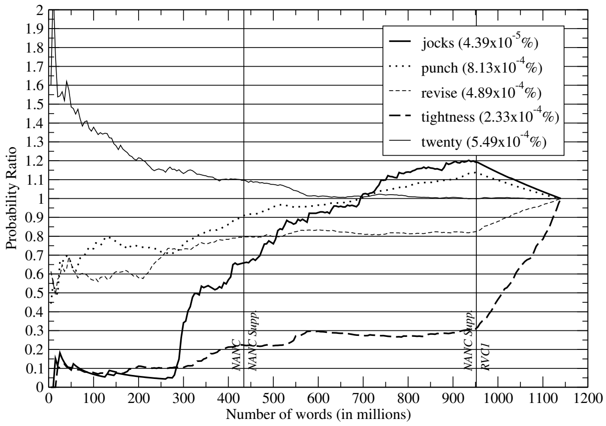

Figure 4: Example ratios for hapax legomena

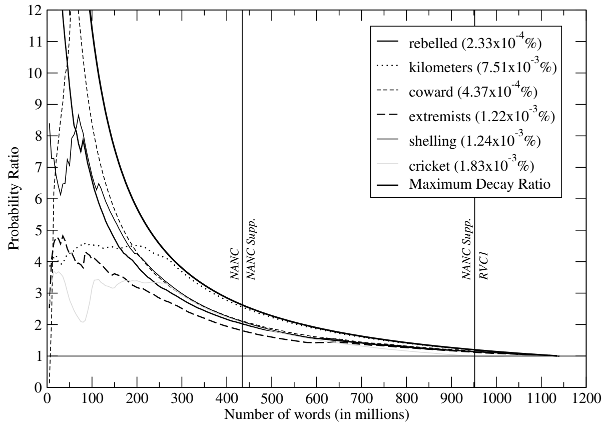

Figure 5: Example ratios for decaying initial words

exhibit the most bursty behaviour, since the newspaper articles are naturally clustered together according to the chronologically grouped events. The most obvious and expected conditioning of the random variables is the topic of the text in question.

However, it is hard to envisage seemingly topic-neutral words, such as `tightness` and `newly`, being conditioned strongly on topic. Other factors that apply to many different types of words include the stylistic and idiomatic expressions favoured by particular genres, authors, editors and even the in-house style guides.

These large corpus experiments demonstrate the failure of simple Poisson models to account for the burstiness of words. The fact that words are not distributed by a simple Poisson process becomes even more apparent as corpus size increases, particularly as the effect of noise and sparseness on the language model is reduced, giving a clearer picture of how badly the current language models fail. With a very large corpus it is obvious that the usual independence assumptions are not always appropriate.

Using very large corpora for simple probability estimation demonstrates the need for more sophisticated statistical models of language. Without better models, all that training upon large corpora can achieve is better estimates of words which are approximately i.i.d.

To fully leverage the information in very large corpora, we need to introduce more dependencies into the models to capture the non-stationary nature of language data. This means that to gain a significant advantage from large corpora, we must develop more sophisticated statistical language models.

We should also briefly mention the other main benefit of increasing corpus size: the acquisition of occurrences of otherwise unseen words. Previously unseen linguistic events are frequently presented to NLP systems. To handle these unseen events the statistical models used by the system must be smoothed. Smoothing typically adds considerable computational complexity to the system since multiple models need to be estimated and applied together, and it is often considered a black art (Chen and Goodman, 1996). Having access to very large corpora ought to reduce the need for smoothing, and so ought to allow us to design simpler systems.

6 Conclusion

The difficulty of obtaining reliable probability estimates is central to many NLP tasks. Can we improve the performance of these systems by simply using a lot more data? As might be expected, for many words, estimating probabilities on a very large corpus can be valuable, improving system performance significantly. This is due to the improved estimates of sparse statistics, made possible by the relatively uniform distribution of these words.

However, there is a large class of commonplace words which fail to display convergent behaviour even on very large corpora. What is striking about these words is that proficient language users would not recognise them as particularly unusual or specialised in their usage, which means that broad-coverage NLP systems should also be expected to handle them competently.

The non-convergence of these words is an indication of their non-stationary distributions, which a simple Poisson model is unable to capture. Since it is no longer a problem of sparseness, even exceptionally large corpora cannot be expected to produce reliable probability estimates. Instead we must relax the independence assumptions underlying the existing language models and incorporate conditional information into the language models.

To fully harness the extra information in a very large corpus we must spend more, and not less, time and effort developing sophisticated language models and machine learning systems.

7 Further Work

We are particularly interested in trying to characterise the burstiness tendencies of individual words and word classes, and the resulting convergence behaviour of their probability estimates. An example of this is calculating the area between unity and the ratio curves. Some example words with different convergence behaviour selected using this area measure are given in Table 2 in the Appendix. We are also interested in applying the exponential models of lexical attraction and repulsion described by Beeferman et al. (1997) to the very large corpus.

We would like to investigate the overall error in the probability mass distribution by comparing the whole distributions at each sample with the final distribution. To estimate the error properly will require smoothing methods to be taken into consideration.

Acknowledgements

We would like to thank Marc Moens, Steve Finch, Tara Murphy, Yuval Krymolowski and the many anonymous reviewers for their insightful comments that have contributed significantly to this paper. This research is partly supported by a Commonwealth scholarship and a Sydney University Travelling scholarship.

References

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France, 9–11 July.

Doug Beeferman, Adam Berger, and John Lafferty. 1997. A model of lexical attraction and repulsion. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 373–380, Madrid, Spain, 7–11 July.

George Casella and Roger L. Berger. 1990. *Statistical Inference*. Duxbury Press, Belmont, CA USA.

Stanley F. Chen and Joshua T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, CA USA, 23–28 June.

James R. Curran and Marc Moens. 2002. Scaling context space. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, Philadelphia, PA USA, 7–12 July.

David Graff. 1995. *North American News Text Corpus*. Linguistic Data Consortium. LDC95T21.

Frank Keller, Maria Lapata, and Olga Ourioupina. 2002. Using the web to overcome data sparseness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA USA, 6–7 July.

Robert MacIntyre. 1998. *North American News Text Supplement*. Linguistic Data Consortium. LDC98T30.

Valentin V. Petrov. 1995. *Limit theorems of probability theory: Sequences of independent random variables*, volume 4 of *Oxford Studies in Probability*. Clarendon Press, Oxford, UK.

T.G. Rose, M. Stevenson, and M. Whitehead. 2002. The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain, 29–31 May.

Martin Volk. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of the Corpus Linguistics 2001 Conference*, pages 601–606, Lancaster, UK, 29 March–2 April.

Appendix

It is possible to get some sense of the convergence behaviour of individual words by calculating the area between the ratio curve and unity. Table 2 lists words with largest and smallest areas, and words that fell in between large and small areas. A large area (MAX $\sum_{i=1}^{n} \frac{\overline{X}_i}{\mu}$) indicates either non-convergent behaviour or convergence from poor initial estimates, and so many of the words are highly conditioned (primarily on topics such as war). These words behave like the words shown in Figure 4 and Figure 5. A small area (MIN $\sum_{i=1}^{n} \frac{\overline{X}_i}{\mu}$) indicates strongly convergent behaviour with accurate initial estimates, and so includes a number of function words. These words behave like the words shown in Figure 1 and Figure 2.

MAX $\sum_{i=1}^{n} \frac{\overline{X}_i}{\mu}$	MID $\sum_{i=1}^{n} \frac{\overline{X}_i}{\mu}$	MIN $\sum_{i=1}^{n} \frac{\overline{X}_i}{\mu}$
convoys	unending	bringing
rebelled	buildings	has
coward	instrument	string
hick	poisoning	the
routing	awesome	been
shelling	livelihood	give
secede	sharpness	form
truce	likewise	remains
convoy	phantom	received
kilometers	acquitted	before
artillery	comfortable	quit
kilometer	complement	wants
shelled	entities	crucial
atolls	generous	allowing
quake	island	seek
showers	advancements	considered
gunners	demonstrates	no
centimeters	linden	in
kilograms	politicking	chosen
shells	spur	involved
armored	veer	nearest
hideouts	scoop	hands
seahorse	drill	with
expedited	skill	car
meters	arrows	respect
airlift	bats	day
skirmished	rewrite	dominate
clays	toughness	avoid
civilians	expands	stay
stronghold	negligence	joins
centimeter	swaying	covered
neighboring	mellowed	removing
downed	rendering	established
besieged	wording	asked
hostilities	disaffected	being
cessation	tempt	preparation
detaining	discourages	houses
meson	jumpy	reeling
rebel	landlords	into
disarm	geared	food
thunderstorm	planet	face

Table 2: Convergence detection using curve area