Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2

Kenneth W. Church
AT&T Labs-Research, 180 Park Ave., Florham Park, NJ., USA
kwc@research.att.com

Abstract

Repetition is very common. Adaptive language models, which allow probabilities to change or adapt after seeing just a few words of a text, were introduced in speech recognition to account for text cohesion. Suppose a document mentions *Noriega* once. What is the chance that he will be mentioned again? If the first instance has probability *p*, then under standard (bag-of-words) independence assumptions, two instances ought to have probability p^2, but we find the probability is actually closer to $p/2$. The first mention of a word obviously depends on frequency, but surprisingly, the second does not. Adaptation depends more on lexical content than frequency; there is more adaptation for content words (proper nouns, technical terminology and good keywords for information retrieval), and less adaptation for function words, cliches and ordinary first names.

1. Introduction

Adaptive language models were introduced in the Speech Recognition literature to model repetition. Jelinek (1997, p. 254) describes cache-based models which combine two estimates of word (ngram) probabilities, Pr_L, a local estimate based on a relatively small cache of recently seen words, and Pr_G, a global estimate based on a large training corpus.

1. Additive: $Pr_A(w) = \lambda Pr_L(w) + (1-\lambda) Pr_G(w)$
2. Case-based: $Pr_C(w) = \begin{cases} \lambda_1 Pr_L(w) & \text{if } w \in cache \\ \lambda_2 Pr_G(w) & \text{otherwise} \end{cases}$

Intuitively, if a word has been mentioned recently, then (a) the probability of that word (and related words) should go way up, and (b) many other words should go down a little. We will refer to (a) as *positive adaptation* and (b) as *negative adaptation*. Our empirical experiments confirm the intuition that positive adaptation, $Pr(+adapt)$, is typically much larger than negative adaptation, $Pr(-adapt)$. That is, $Pr(+adapt) >> Pr(prior) > Pr(-adapt)$. Two methods, $Pr(+adapt_1)$ and $Pr(+adapt_2)$, will be introduced for estimating positive adaptation.

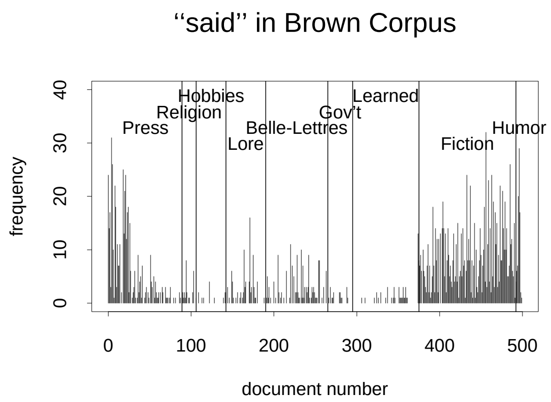
1. $Pr(+adapt_1) = Pr(w \in test | w \in history)$
2. $Pr(+adapt_2) = Pr(k \geq 2 | k \geq 1) \approx df_2 / df_1$

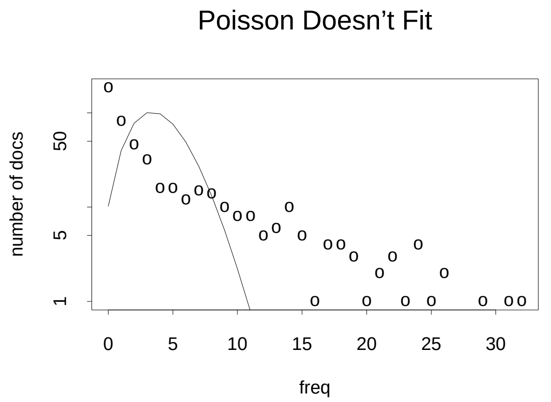
The two methods produce similar results, usually well within a factor of two of one another. The first method splits each document into two equal pieces, a *history* portion and a *test* portion. The adapted probabilities are modeled as the chance that a word will appear in the test portion, given that it appeared in the history. The second method, suggested by Church and Gale (1995), models adaptation as the chance of a second mention (probability that a word will appear two or more times, given that it appeared one or more times). $Pr(+adapt_2)$ is approximated by df_2/df_1, where df_k is the number of documents that contain the word/ngram k or more times. (df_k is a generalization of *document frequency*, *df*, a standard term in Information Retrieval.)

Both methods are non-parametric (unlike cache models). Parametric assumptions, when appropriate, can be very powerful (better estimates from less training data), but errors resulting from inappropriate assumptions can outweigh the benefits. In this empirical investigation of the magnitude and shape of adaptation we decided to use conservative non-parametric methods to hedge against the risk of inappropriate parametric assumptions.

The two plots (below) illustrate some of the reasons for being concerned about standard parametric assumptions. The first plot shows the number of times that the word ‘‘said’’ appears in each of the 500 documents in the Brown Corpus (Francis & Kucera, 1982). Note that there are quite a few documents with more than 15 instances of ‘‘said,’’ especially in *Press* and *Fiction*. There are also quite a few documents with hardly any instances of ‘‘said,’’ especially in the *Learned* genre. We have found a similar pattern in other collections; ‘‘said’’ is more common in newswire (Associated Press and Wall Street Journal) than technical writing (Department of Energy abstracts).

The second plot (below) compares these Brown Corpus observations to a Poisson. The circles indicate the number of documents that have *x* instances of ‘‘said.’’ As mentioned above, *Press* and *Fiction* documents can mention ‘‘said’’ 15 times or more, while documents in the *Learned* genre might not mention the word at all. The line shows what would be expected under a Poisson. Clearly the line does not fit the circles very well. The probability of ‘‘said’’ depends on many factors (e.g, genre, topic, style, author) that make the distributions broader than chance (Poisson). We find especially broad distributions for words that adapt a lot.





We will show that adaptation is huge. $Pr(+adapt)$ is often several orders of magnitude larger than $Pr(prior)$. In addition, we find that $Pr(+adapt)$ has a very different shape from $Pr(prior)$. By construction, $Pr(prior)$ varies over many orders of magnitude depending on the frequency of the word. Interestingly, though, we find that $Pr(+adapt)$ has almost no dependence on word frequency, although there is a strong lexical dependence. Some words adapt more than others. The result is quite robust. Words that adapt more in one corpus also tend to adapt more in another corpus of similar material. Both the magnitude and especially the shape (lack of dependence on frequency as well as dependence on content) are hard to capture in an additive-based cache model.

Later in the paper, we will study *neighbors*, words that do not appear in the history but do appear in documents near the history using an information retrieval notion of near. We find that neighbors adapt more than non-neighbors, but not as much as the history. The shape is in between as well. Neighbors have a modest dependency on frequency, more than the history, but not as much as the prior.

Neighbors are an extension of Florian & Yarowsky (1999), who used topic clustering to build a language model for contexts such as: ‘‘It is at least on the Serb side a real setback to the *x*.’’ Their work was motivated by speech recognition applications where it would be desirable for the language model to favor $x =$ ‘‘peace’’ over $x =$ ‘‘piece.’’ Obviously, acoustic evidence is not very helpful in this case. Trigrams are also not very helpful because the strongest clues (e.g., ‘‘Serb,’’ ‘‘side’’ and ‘‘setback’’) are beyond the window of three words. Florian & Yarowsky cluster documents into about 10^2 topics, and compute a separate trigram language model for each topic. Neighbors are similar in spirit, but support more topics.

2. Estimates of Adaptation: Method 1

Method 1 splits each document into two equal pieces. The first half of each document is referred to as the *history* portion of the document and the second half of each document is referred to as the *test* portion of the document. The task is to predict the *test* portion of the document given the *history*. We start by computing a contingency table for each word, as illustrated below:

Documents containing ‘‘hostages’’ in 1990 AP

	test	$\overline{test}$
history	$a = 638$	$b = 505$
$\overline{history}$	$c = 557$	$d = 76787$

This table indicates that there are (a) 638 documents with ‘‘hostages’’ in both the first half (history) and the second half (test), (b) 505 documents with ‘‘hostages’’ in just the first half, (c) 557 documents with ‘‘hostages’’ in just the second half, and (d) 76,787 documents with ‘‘hostages’’ in neither half. Positive and negative adaptation are defined in terms *a, b, c* and *d*.

$$Pr(+adapt_1) = Pr(w \in test \mid w \in history) \approx \frac{a}{a+b}$$

$$Pr(-adapt_1) = Pr(w \in test \mid \neg w \in history) \approx \frac{c}{c+d}$$

Adapted probabilities will be compared to:

$$Pr(prior) = Pr(w \in test) \approx (a+c)/D$$

where $D = a + b + c + d$.

Positive adaptation tends to be much larger than the prior, which is just a little larger than negative adaptation, as illustrated in the table below for the word ''hostages'' in four years of the Associated Press (AP) newswire. We find remarkably consistent results when we compare one year of the AP news to another (though topics do come and go over time). Generally, the differences of interest are huge (orders of magnitude) compared to the differences among various control conditions (at most factors of two or three). Note that values are more similar within columns than across columns.

Pr(+adapt) >> Pr(prior) > Pr(–adapt)

prior	+adapt	–adapt	source	w
0.014	0.56	0.0069	AP87	hostages
0.015	0.56	0.0072	AP90	
0.013	0.59	0.0057	AP91	
0.0044	0.39	0.0030	AP93	

3. Adaptation is Lexical

We find that some words adapt more than others, and that words that adapt more in one year of the AP also tend to adapt more in another year of the AP. In general, words that adapt a lot tend to have more content (e.g., good keywords for information retrieval (IR)) and words that adapt less have less content (e.g., function words).

It is often assumed that word frequency is a good (inverse) correlate of content. In the psycholinguistic literature, the term ''high frequency'' is often used synonymously with ''function words,'' and ''low frequency'' with ''content words.'' In IR, inverse document frequency (IDF) is commonly used for weighting keywords. The table below is interesting because it questions this very basic assumption. We compare two words, ''Kennedy'' and ''except,'' that are about equally frequent (similar priors). Intuitively, ''Kennedy'' is a content word and ''except'' is not. This intuition is supported by the adaptation statistics: the adaptation ratio, $Pr(+adapt)/Pr(prior)$, is much larger for ''Kennedy'' than for ''except.'' A similar pattern holds for negative adaptation, but in the reverse direction. That is, $Pr(-adapt)/Pr(prior)$ is much smaller for ''Kennedy'' than for ''except.''

Kennedy adapts more than *except*

prior	+adapt	–adapt	source	w
0.012	0.27	0.0091	AP90	Kennedy
0.015	0.40	0.0084	AP91	
0.014	0.32	0.0094	AP93	
0.016	0.049	0.016	AP90	except
0.014	0.047	0.014	AP91	
0.012	0.048	0.012	AP93	

In general, we expect more adaptation for better keywords (e.g., ''Kennedy'') and less adaptation for less good keywords (e.g., function words such as ''except''). This observation runs counter to the standard practice of weighting keywords solely on the basis of frequency, without considering adaptation. In a related paper, Umemura and Church (submitted), we describe a term weighting method that makes use of adaptation (sometimes referred to as burstiness).

Distinctive surnames adapt more than ordinary first names

prior	+adapt	–adapt	source	w
0.0079	0.71	0.0026	AP90	Noriega
0.0038	0.80	0.0009	AP91	
0.0006	0.90	0.0002	AP90	Aristide
0.0035	0.77	0.0009	AP91	
0.0011	0.47	0.0006	AP90	Escobar
0.0014	0.74	0.0006	AP91	
0.068	0.18	0.059	AP90	John
0.066	0.16	0.057	AP91	
0.025	0.11	0.022	AP90	George
0.025	0.13	0.022	AP91	
0.029	0.15	0.025	AP90	Paul
0.028	0.13	0.025	AP91	

The table above compares surnames with first names. These surnames are excellent keywords unlike the first names, which are nearly as useless for IR as function words. The adaptation ratio, $Pr(+adapt)/Pr(prior)$, is much larger for the surnames than for the first names.

What is the probability of seeing two *Noriega*s in a document? The chance of the first one is $p \approx 0.006$. According to the table above, the chance of two is about $0.75p$, closer to $p/2$ than p^2. Finding a rare word like *Noriega* in a document is like lightning. We might not expect lightning to strike twice, but it happens all the time, especially for good keywords.

4. Smoothing (for low frequency words)

Thus far, we have seen that adaptation can be large, but to demonstrate the shape property (lack of dependence on frequency), the counts in the contingency table need to be smoothed. The problem is that the estimates of *a*, *b*, *c*, *d*, and especially estimates of the ratios of these quantities, become unstable when the counts are small. The standard methods of smoothing in the speech recognition literature are Good-Turing (GT) and Held-Out (HO), described in sections 15.3 & 15.4 of Jelinek (1997). In both cases, we let *r* be an observed count of an object (e.g., the frequency of a word and/or ngram), and $r*$ be our best estimate of *r* in another corpus of the same size (all other things being equal).

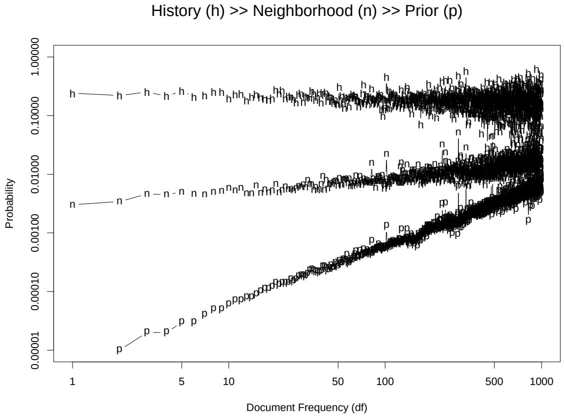
4.1 Standard Held-Out (HO)

HO splits the training corpus into two halves. The first half is used to count *r* for all objects of interest (e.g., the frequency of all words in vocabulary). These counts are then used to group objects into bins. The r^{th} bin contains all (and only) the words with count *r*. For each bin, we compute N_r, the number of words in the r^{th} bin. The second half of the training corpus is then used to compute C_r, the aggregate frequency of all the words in the r^{th} bin. The final result is simply: $r* = C_r / N_r$ If the two halves of the training corpora or the test corpora have different sizes, then $r*$ should be scaled appropriately.

We chose HO in this work because it makes few assumptions. There is no parametric model. All that is assumed is that the two halves of the training corpus are similar, and that both are similar to the testing corpus. Even this assumption is a matter of some concern, since major stories come and go over time.

4.2 Application of HO to Contingency Tables

As above, the training corpus is split into two halves. We used two different years of AP news. The first half is used to count document frequency *df*. (Document frequency will be used instead of standard (term) frequency.) Words are binned by *df* and by their cell in the contingency table. The first half of the corpus is used to compute the number of words in each bin: $N_{df,a}$, $N_{df,b}$, $N_{df,c}$ and $N_{df,d}$; the second half of the corpus is used to compute the aggregate document frequency for the words in each bin: $C_{df,a}$, $C_{df,b}$, $C_{df,c}$ and $C_{df,d}$. The final result is simply: $a^*_{df} = C_{df,a} / N_{df,a}$, $b^*_{df} = C_{df,b} / N_{df,b}$, $c^*_{df} = C_{df,c} / N_{df,c}$ and $d^*_{df} = C_{df,d} / N_{df,d}$. We compute the probabilities as before, but replace *a*, *b*, *c*, *d* with $a*$, $b*$, $c*$, $d*$, respectively.



With these smoothed estimates, we are able to show that $Pr(+adapt)$, labeled *h* in the plot above, is larger and less dependent on frequency than $Pr(prior)$, labeled *p*. The plot shows a third group, labeled *n* for neighbors, which will be described later. Note that the *n*s fall between the *p*s and the *h*s.

Thus far, we have seen that adaptation can be huge: $Pr(+adapt) >> Pr(prior)$, often by two or three orders of magnitude. Perhaps even more surprisingly, although the first mention depends strongly on frequency (*df*), the second does not. Some words adapt more (e.g., *Noriega, Aristide, Escobar*) and some words adapt less (e.g., *John, George, Paul*). The results are robust. Words that adapt more in one year of AP news tend to adapt more in another year, and vice versa.

5. Method 2: $Pr(+adapt_2)$

So far, we have limited our attention to the relatively simple case where the history and the test are the same size. In practice, this won't be the case. We were concerned that the observations above might be artifacts somehow caused by this limitation.

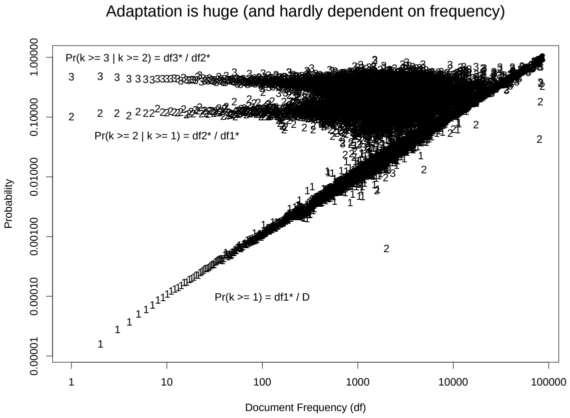
We experimented with two approaches for understanding the effect of this limitation and found that the size of the history doesn't change $Pr(+adapt)$ very much. The first approach split the history and the test at various points ranging from 5% to 95%. Generally, $Pr(+adapt_1)$ increases as the size of the test portion grows relative to the size of the history, but the effect is relatively small (more like a factor of two than an order of magnitude).

We were even more convinced by the second approach, which uses $Pr(+adapt_2)$, a completely different argument for estimating adaptation and doesn't depend on the relative size of the history and the test. The two methods produce remarkably similar results, usually well within a factor of two of one another (even when adapted probabilities are orders of magnitude larger than the prior).

$Pr(+adapt_2)$ makes use of $df_j(w)$, a generalization of document frequency. $df_j(w)$ is the number of documents with j or more instances of w; (df_1 is the standard notion of df).

$$Pr(+adapt_2) = Pr(k \geq 2 \mid k \geq 1) = df_2 / df_1$$

Method 2 has some advantages and some disadvantages in comparison with method 1. On the positive side, method 2 can be generalized to compute the chance of a third instance: $Pr(k \geq 3 \mid k \geq 2)$. But unfortunately, we do not know how to use method 2 to estimate negative adaptation; we leave that as an open question.



The plot (above) is similar to the plot in section 4.2 which showed that adapted probabilities (labeled h) are larger and less dependent on frequency than the prior (labeled p). So too, the plot (above) shows that the second and third mentions of a word (labeled 2 and 3, respectively) are larger and less dependent on frequency than the first mention (labeled 1). The plot in section 4.2 used method 1 whereas the plot (above) uses method 2. Both plots use the HO smoothing, so there is only one point per bin (df value), rather than one per word.

6. Neighborhoods (Near)

Florian and Yarowsky's example, ''It is at least on the Serb side a real setback to the *x*,'' provides a nice motivation for neighborhoods. Suppose the context (history) mentions a number of words related to a peace process, but doesn't mention the word ''peace.'' Intuitively, there should still be some adaptation. That is, the probability of ''peace'' should go up quite a bit (positive adaptation), and the probability of many other words such as ''piece'' should go down a little (negative adaptation).

We start by partitioning the vocabulary into three exhaustive and mutually exclusive sets: *hist*, *near* and *other* (abbreviations for *history*, *neighborhood* and *otherwise*, respectively). The first set, *hist*, contains the words that appear in the first half of the document, as before. *Other* is a catchall for the words that are in neither of the first two sets.

The interesting set is *near*. It is generated by query expansion. The history is treated as a query in an information retrieval document-ranking engine. (We implemented our own ranking engine using simple IDF weighting.) The neighborhood is the set of words that appear in the $k \approx 10$ or $k \approx 100$ top documents returned by the retrieval engine. To ensure that the three sets partition the vocabulary, we exclude the history from the neighborhood:

near = words in query expansion of hist – hist

The adaptation probabilities are estimated using a contingency table like before, but we now have a three-way partition (*hist*, *near* and *other*) of the vocabulary instead of the two-way partition, as illustrated below.

Documents containing ''peace'' in 1991 AP

	test	
history	a =2125	b =2160
	c =1963	d =74573

	test	
hist	a =2125	b =2160
near	e =1479	f =22516
other	g =484	h =52057

In estimating adaptation probabilities, we continue to use a, b, c and d as before, but four new variables are introduced: e, f, g and h, where $c = e + g$ and $d = f + h$.

$Pr(w \in test) \approx (a+c)/D$	prior
$Pr(w \in test \mid w \in hist) \approx a/(a+b)$	hist
$Pr(w \in test \mid w \in near) \approx e/(e+f)$	near
$Pr(w \in \text{test} \mid w \in other) \approx g/(g+h)$	other

The table below shows that ‘‘Kennedy’’ adapts more than ‘‘except’’ and that ‘‘peace’’ adapts more than ‘‘piece.’’ That is, ‘‘Kennedy’’ has a larger spread than ‘‘except’’ between the history and the otherwise case.

prior	hist	near	other	src	w
0.026	0.40	0.022	0.0050	AP91	Kennedy
0.020	0.32	0.025	0.0038	AP93	
0.026	0.05	0.018	0.0122	AP91	except
0.019	0.05	0.014	0.0081	AP93	
0.077	0.50	0.062	0.0092	AP91	peace
0.074	0.49	0.066	0.0069	AP93	
0.015	0.10	0.014	0.0066	AP91	piece
0.013	0.08	0.015	0.0046	AP93	

When *df* is small ($df < 100$), HO smoothing is used to group words into bins by *df*. Adaptation probabilities are computed for each bin, rather than for each word. Since these probabilities are implicitly conditional on *df*, they have already been weighted by *df* in some sense, and therefore, it is unnecessary to introduce an additional explicit weighting scheme based on *df* or a simple transform thereof such as IDF.

The experiments below split the neighborhood into four classes, ranging from *better neighbors* to *worse neighbors*, depending on expansion frequency, *ef*. $ef(t)$ is a number between 1 and *k*, indicating how many of the *k* top scoring documents contain *t*. (Better neighbors appear in more of the top scoring documents, and worse neighbors appear in fewer.) All the neighborhood classes fall between *hist* and *other*, with better neighbors adapting more than worse neighbors.

7. Experimental Results

Recall that the task is to predict the *test* portion (the second half) of a document given the *history* (the first half). The following table shows a selection of words (sorted by the third column) from the test portion of one of the test documents. The table is separated into thirds by horizontal lines. The words in the top third receive much higher scores by the proposed method (S) than by a baseline (B). These words are such good keywords that one can fairly confidently guess what the story is about. Most of these words receive a high score because they were mentioned in the history portion of the document, but ‘‘laid-off’’ receives a high score by the neighborhood mechanism. Although ‘‘laid-off’’ is not mentioned explicitly in the history, it is obviously closely related to a number of words that were, especially ‘‘layoffs,’’ but also ‘‘notices’’ and ‘‘cuts.’’ It is reassuring to see the neighborhood mechanism doing what it was designed to do.

The middle third shows words whose scores are about the same as the baseline. These words tend to be function words and other low content words that give us little sense of what the document is about. The bottom third contains words whose scores are much lower than the baseline. These words tend to be high in content, but misleading. The word ‘‘arms,’’ for example, might suggest that story is about a military conflict.

S	B	log2(S/B)	Set	Term
0.19	0.00	11.06	hist	Binder
0.22	0.00	7.45	hist	layoff
0.06	0.00	5.71	hist	notices
0.36	0.01	5.66	hist	Boeing
0.02	0.00	5.11	near3	laid-off
0.25	0.02	3.79	hist	cuts
0.01	0.01	0.18	near3	projects
0.89	0.81	0.15	hist	said
0.06	0.05	0.11	near4	announced
0.06	0.06	0.09	near4	As
0.00	0.00	0.09	near1	employed
0.00	0.00	–0.61	other	714
0.00	0.01	–0.77	other	managed
0.01	0.02	–1.05	near2	additional
0.00	0.01	–1.56	other	wave
0.00	0.03	–3.41	other	arms

The proposed score, *S*, shown in column 1, is:

$$Pr_S(w) = \begin{cases} Pr(w \mid hist) & \text{if } w \in hist \\ Pr(w \mid near_1) & \text{if } w \in near_1 \\ Pr(w \mid near_2) & \text{if } w \in near_2 \\ Pr(w \mid near_3) & \text{if } w \in near_3 \\ Pr(w \mid near_4) & \text{if } w \in near_4 \\ Pr(w \mid other) & \text{otherwise} \end{cases}$$

where $near_1$ through $near_4$ are four neighborhoods ($k = 100$). Words in $near_4$ are the best neighbors ($ef \geq 10$) and words in $near_1$ are the worst neighbors ($ef = 1$). The baseline, *B*, shown in column 2, is: $Pr_B(w) = df/D$. Column 3 compares the first two columns.

We applied this procedure to a year of the AP news and found a sizable gain in information on average: 0.75 bits per word type per document. In addition, there were many more big winners (20% of the documents gained 1 bit/type) than big losers (0% lost 1 bit/type). The largest winners include lists of major cities and their temperatures, lists of major currencies and their prices, and lists of commodities and their prices. Neighborhoods are quite successful in guessing the second half of such lists.

On the other hand, there were a few big losers, e.g., articles that summarize the major stories of the day, week and year. The second half of a summary article is almost never about the same subject as the first half. There were also a few end-of-document delimiters that were garbled in transmission causing two different documents to be treated as if they were one. These garbled documents tended to cause trouble for the proposed method; in such cases, the history comes from one document and the test comes from another.

In general, the proposed adaptation method performed well when the history is helpful for predicting the test portion of the document, and it performed poorly when the history is misleading. This suggests that we ought to measure topic shifts using methods suggested by Hearst (1994) and Florian & Yarowsky (1999). We should not use the history when we believe that there has been a major topic shift.

8. Conclusions

Adaptive language models were introduced to account for repetition. It is well known that the second instance of a word (or ngram) is much more likely than the first. But what we find surprising is just how large the effect is. The chance of two Noriegas is closer to $p/2$ than p^2.

In addition to the magnitude of adaptation, we were also surprised by the shape: while the first instance of a word depends very strongly on frequency, the second does not. Adaptation depends more on content than frequency; adaptation is stronger for content words such as proper nouns, technical terminology and good keywords for information retrieval, and weaker for function words, cliches and first names.

The shape and magnitude of adaptation has implications for psycholinguistics, information retrieval and language modeling. Psycholinguistics has tended to equate word frequency with content, but our results suggest that two words with similar frequency (e.g., ‘‘Kennedy’’ and ‘‘except’’) can be distinguished on the basis of their adaptation. Information retrieval has tended to use frequency in a similar way, weighting terms by IDF (inverse document frequency), with little attention paid to adaptation. We propose a term weighting method that makes use of adaptation (burstiness) and expansion frequency in a related paper (Umemura and Church, submitted).

Two estimation methods were introduced to demonstrate the magnitude and shape of adaptation. Both methods produce similar results.

- $Pr(+adapt_1) = Pr(test \mid hist)$
- $Pr(+adapt_2) = Pr(k \geq 2 \mid k \geq 1)$

Neighborhoods were then introduced for words such as ‘‘laid-off’’ that were not in the history but were close (‘‘laid-off’’ is related to ‘‘layoff,’’ which was in the history). Neighborhoods were defined in terms of query expansion. The history is treated as a query in an information retrieval document-ranking system. Words in the k top-ranking documents (but not in the history) are called *neighbors*. Neighbors adapt more than other terms, but not as much as words that actually appeared in the history. Better neighbors (larger *ef*) adapt more than worse neighbors (smaller *ef*).

References

Church, K. and Gale, W. (1995) ‘‘Poisson Mixtures,’’ *Journal of Natural Language Engineering*, 1:2, pp. 163-190.

Florian, R. and Yarowsky, D. (1999) ‘‘Dynamic Nonlocal Language Modeling via Hierarchical Topic-Based Adaptation,’’ ACL, pp. 167-174.

Francis, W., and Kucera, H. (1982) *Frequency Analysis of English Usage*, Houghton Mifflin Company, Boston, MA.

Hearst, M. (1994) *Context and Structure in Automated Full-Text Information Access*, PhD Thesis, Berkeley, available via www.sims.berkeley.edu/~hearst.

Jelinek, F. (1997) *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, USA.

Umemura, K. and Church, K. (submitted) ‘‘Empirical Term Weighting: A Framework for Studying Limits, Stop Lists, Burstiness and Query Expansion.