A Bayesian mixture model for term re-occurrence and burstiness

Avik Sarkar[1], **Paul H Garthwaite**[2], **Anne De Roeck**[1]
[1] Department of Computing, [2] Department of Statistics
The Open University
Milton Keynes, MK7 6AA, UK
{a.sarkar, p.h.garthwaite, a.deroeck}@open.ac.uk

Abstract

This paper proposes a model for term re-occurrence in a text collection based on the gaps between successive occurrences of a term. These gaps are modeled using a mixture of exponential distributions. Parameter estimation is based on a Bayesian framework that allows us to fit a flexible model. The model provides measures of a term's re-occurrence rate and *within-document burstiness*. The model works for all kinds of terms, be it rare content word, medium frequency term or frequent function word. A measure is proposed to account for the term's importance based on its distribution pattern in the corpus.

1 Introduction

Traditionally, Information Retrieval (IR) and Statistical Natural Language Processing (NLP) applications have been based on the "bag of words" model. This model assumes term independence and homogeneity of the text and document under consideration, i.e. the terms in a document are all assumed to be distributed homogeneously. This immediately leads to the Vector Space representation of text. The immense popularity of this model is due to the ease with which mathematical and statistical techniques can be applied to it.

The model assumes that once a term occurs in a document, its overall frequency in the entire document is the only useful measure that associates a term with a document. It does not take into consideration whether the term occurred in the beginning, middle or end of the document. Neither does it consider whether the term occurs many times in close succession or whether it occurs uniformly throughout the document. It also assumes that additional positional information does not provide any extra leverage to the performance of the NLP and IR applications based on it. This assumption has been shown to be wrong in certain applications (Franz, 1997).

Existing models for term distribution are based on the above assumption, so they can merely estimate the term's frequency in a document or a term's topical behavior for a content term. The occurrence of a content word is classified as *topical* or *non-topical* based on whether it occurs once or many times in the document (Katz, 1996). We are not aware of any existing model that makes less stringent assumptions and models the distribution of occurrences of a term.

In this paper we describe a model for term re-occurrence in text based on the gaps between successive occurrences of the term and the position of its first occurrence in a document. The gaps are modeled by a mixture of exponential distributions. Non-occurrence of a term in a document is modeled by the statistical concept of *censoring*, which states that the event of observing a certain term is censored at the end of the document, i.e. the document length. The modeling is done in a Bayesian framework.

The organization of the paper is as follows. In section 2 we discuss existing term distribution models, the issue of burstiness and some other work that demonstrates the failure of the "bag of words" as-

Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL),
pages 48–55, Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

sumption. In section 3 we describe our mixture model, the issue of censoring and the Bayesian formulation of the model. Section 4 describes the Bayesian estimation theory and methodology. In section 5 we talk about ways of drawing inferences from our model, present parameter estimates on some chosen terms and present case studies for a few selected terms. We discuss our conclusions and suggest directions for future work in section 6.

2 Existing Work

2.1 Models

Previous attempts to model a term's distribution pattern have been based on the Poisson distribution. If the number of occurrences of a term in a document is denoted by k, then the model assumes:

$$p(k) \quad = \quad e^{-\lambda} \frac{\lambda^k}{k!}$$

for $k = 0, 1, 2, \ldots$ Estimates based on this model are good for non-content, non-informative terms, but not for the more informative content terms (Manning and Schütze, 1999).

The two-Poisson model is suggested as a variation of the Poisson distribution (Bookstein and Swanson, 1974; Church and Gale, 1995b). This model assumes that there are two classes of documents associated with a term, one class with a low average number of occurrences and the other with a high average number of occurrences.

$$p(k) \quad = \quad \alpha e^{-\lambda_1} \frac{\lambda_1^k}{k!} + (1-\alpha) e^{-\lambda_2} \frac{\lambda_2^k}{k!},$$

where α and $(1 - \alpha)$ denote the probabilities of a document in each of these classes. Often this model under-estimates the probability that a term will occur exactly twice in a document.

2.2 Burstiness

Burstiness is a phenomenon of content words, whereby they are likely to occur again in a text after they have occurred once. Katz (1996) describes *within-document burstiness* as the close proximity of all or some individual instances of a word within a document exhibiting multiple occurrences.

He proposes a model for within-document burstiness with three parameters as:

- the probability that a term occurs in a document at all (document frequency)
- the probability that it will occur a second time in a document given that it has occurred once
- the probability that it will occur another time, given that it has already occurred k times (where k > 1).

The drawbacks of this model are: (a) it cannot handle non-occurrence of a term in a document; (b) the model can handle only content terms, and is not suitable for high frequency function words or medium frequency terms; and (c) the rate of re-occurrence of the term or the length of gaps cannot be accounted for. We overcome these drawbacks in our model.

A measure of burstiness was proposed as a binary value that is based on the magnitude of average-term frequency of the term in the corpus (Kwok, 1996). This measure takes the value 1 (bursty term) if the average-term frequency value is large and 0 otherwise. The measure is too naive and incomplete to account for term burstiness.

2.3 Homogeneity Assumption

The popular “bag of words” assumption for text states that a term's occurrence is uniform and homogeneous throughout. A measure of homogeneity or self-similarity of a corpus can be calculated, by dividing the corpus into two frequency lists based on the term frequency and then calculating the χ^2 statistic between them (Kilgarriff, 1997). Various schemes for dividing the corpus were used (De Roeck et al., 2004a) to detect homogeneity of terms at document level, within-document level and by choosing text chunks of various sizes. Their work revealed that homogeneity increases by nullifying the within document term distribution pattern and homogeneity decreases when chunks of larger size are chosen as it incorporates more document structure in it. Other work based on the same methodology (De Roeck et al., 2004b) reveals that even very frequent function words do not distribute homogeneously over a corpus or document. These (De Roeck et al., 2004a; De Roeck et al., 2004b) provide evidence of the fact that the ”bag of words” assumption is invalid. Thus it sets the platform for a model that defies the independence assumption and considers the term distribution pattern in a document and corpus.

3 Modeling

3.1 Terminology and Notation

We build a single model for a particular term in a given corpus. Let us suppose the term under consideration is x as shown in Figure 1. We describe the notation for a particular document, i in the corpus.

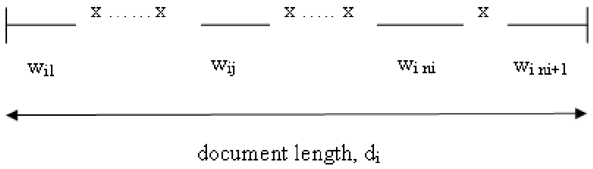

Figure 1: The document structure and the gaps between terms

- d_i denotes the number of words in document i (i.e. the document length).
- n_i denotes the number of occurrences of term x in document i.
- w_{i1} denotes the position of the first occurrence of term x in document i.
- $w_{i2}, \ldots, w_{in_i}$ denotes the successive gaps between occurrences of term x in document i.
- w_{in_i+1} denotes the gap for the next occurrence of x, somewhere after the document ends.
- cen_i is the value at which observation w_{in_i+1} is censored, as explained in section 3.2.2.

3.2 The Model

We suppose we are looking through a document, noting when the term of interest occurs. Our model assumes that the term occurs at some low underlying base rate $1/\lambda_1$ but, after the term has occurred, then the probability of it occurring soon afterwards is increased to some higher rate $1/\lambda_2$. Specifically, the rate of re-occurrence is modeled by a mixture of two exponential distributions. Each of the exponential components is described as follows:

- The exponential component with larger mean (average), $1/\lambda_1$, determines the rate with which the particular term will occur if it has not occurred before or it has not occurred recently.
- The second component with smaller mean (average), $1/\lambda_2$, determines the rate of re-occurrence in a document or text chunk given that it has already occurred recently. This component captures the bursty nature of the term in the text (or document) i.e. the *within-document burstiness*.

The mixture model is described as follows:

$$\phi(w_{ij}) \quad = \quad p\lambda_1 e^{-\lambda_1 w_{ij}} + (1-p)\lambda_2 e^{-\lambda_2 w_{ij}}$$

for $j \in \{2, \ldots, n_i\}$. p and $(1-p)$ denote respectively, the probabilities of membership for the first and the second exponential distribution.

There are a few boundary conditions that the model is expected to handle. We take each of these cases and discuss them briefly:

3.2.1 First occurrence

The model treats the first occurrence of a term differently from the other gaps. The second exponential component measuring burstiness does not feature in it. Hence the distribution is:

$$\phi_1(w_{i1}) \quad = \quad \lambda_1 e^{-\lambda_1 w_{i1}}$$

3.2.2 Censoring

Here we discuss the modeling of two cases that require special attention, corresponding to gaps that have a minimum length but whose actual length is unknown. These cases are:

- The last occurrence of a term in a document.
- The term does not occur in a document at all.

We follow a standard technique from clinical trials, where a patient is observed for a certain amount of time and the observation of the study is expected in that time period (the observation might be the time until death, for example). In some cases it happens that the observation for a patient does not occur in that time period. In such a case it is assumed that the observation would occur at sometime in the future. This is called *censoring* at a certain point.

In our case, we assume the particular term would eventually occur, but the document has ended before it occurs so we do not observe it. In our notation we observe the term n_i times, so the $(n_i+1)^{th}$ time the term occurs is *after* the end of the document. Hence the distribution of w_{in_i+1} is censored at length cen_i. If cen_i is small, so that the n_i^{th} occurrence of the term is near the end of the document, then it is not surprising that w_{in_i+1} is censored. In contrast if cen_i is large, so the n_i^{th} occurrence is far from the end of the document, then either it is surprising that the term did not re-occur, or it suggests the term is rare. The information about the model parameters that is given by the censored occurrence is,

$$Pr(w_{in_i+1} > cen_i) \quad = \quad \int_{cen_i}^{\infty} \phi(x)dx$$

$= pe^{-\lambda_1 cen_i} + (1-p)e^{-\lambda_2 cen_i}$; where,

$$cen_i \quad = \quad d_i - \sum_{j=1}^{n_i} w_{ij}$$

Also when a particular term does not occur in a document, our model assumes that the term would eventually occur had the document continued indefinitely. In this case the first occurrence is censored and censoring takes place at the document length. If a term does not occur in a long document, it suggests the term is rare.

3.3 Bayesian formulation

Our modeling is based on a *Bayesian* approach (Gelman et al., 1995). The Bayesian approach differs from the traditional frequentist approach. In the frequentist approach it is assumed that the parameters of a distribution are constant and the data varies. In the Bayesian approach one can assign distributions to the parameters in a model. We choose non-informative priors, as is common practice in Bayesian applications. So we put,
$p \sim Uniform(0,1)$, and
$\lambda_1 \sim Uniform(0,1)$
To tell the model that λ_2 is the larger of the two λs, we put $\lambda_2 = \lambda_1 + \gamma$, where $\gamma > 0$, and
$\gamma \sim Uniform(0,1)$
Also cen_i depends on the document length d_i and the number of occurrences of the term in that document, n_i. Fitting mixture techniques is tricky and requires special methods. We use data augmentation to make it feasible to fit the model using Gibbs Sampling (section 4.2). For details about this, see Robert (1996) who describes in detail the fitting of mixture models in MCMC methods (section 4.2).

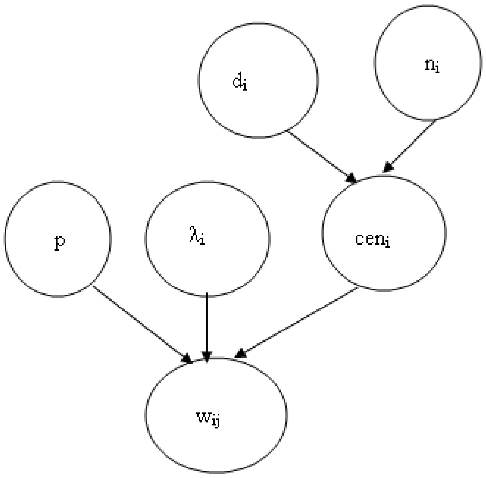

Figure 2: Bayesian dependencies between the parameters

4 Parameter Estimation

4.1 Bayesian Estimation

In the Bayesian approach of parameter estimation, the parameters are uncertain, and it is assumed that they follow some distribution. In our case the parameters and the data are defined as:
$\vec{\Theta} = \{p, \lambda_1, \lambda_2\}$ denote the parameters of the model.
$\vec{W} = \{w_{i1}, \ldots, w_{in_i}, w_{in_i+1}\}$ denotes the data.
Hence based on this we may define the following:

- $f(\vec{\Theta})$ is the **prior distribution** of $\vec{\Theta}$ as assigned in section 3.3. It summarizes everything we know about $\vec{\Theta}$ apart from the data $\vec{W}$.

- $f(\vec{W}|\vec{\Theta})$ is the **likelihood function**. It is our model for the data $\vec{W}$ conditional on the parameters $\vec{\Theta}$. (As well as the observed data, the likelihood also conveys the information given by the censored values)

- $f(\vec{\Theta}|\vec{W})$ is the **posterior distribution** of $\vec{\Theta}$, given $\vec{W}$. It describes our beliefs about the parameters given the information we have.

Deriving the density function for a parameter set $\vec{\Theta}$ after observing data $\vec{W}$, can be achieved by using **Bayes Theorem** as:

$$f(\vec{\Theta}|\vec{W}) = \frac{f(\vec{W}|\vec{\Theta})f(\vec{\Theta})}{f(\vec{W})} \qquad (1)$$

where $f(\vec{W})$ is simply a normalizing constant, independent of $\vec{\Theta}$. It can be computed in terms of the likelihood and prior as:

$$f(\vec{W}) = \int f(\vec{W}|\vec{\Theta})f(\vec{\Theta})d\vec{\Theta}$$

Hence equation 1 is reduced to:

$$f(\vec{\Theta}|\vec{W}) \propto f(\vec{W}|\vec{\Theta})f(\vec{\Theta})$$

So, once we have specified the posterior density function $f(\vec{\Theta}|\vec{W})$, we can obtain the estimates of the parameters $\vec{\Theta}$ by simply averaging the values generated by $f(\vec{\Theta}|\vec{W})$.

4.2 Gibbs Sampling

The density function of Θ_i, $f(\Theta_i|\vec{W})$ can be obtained by integrating $f(\vec{\Theta}|\vec{W})$ over the remaining parameters of $\vec{\Theta}$. But in many cases, as in ours, it is impossible to find a closed form solution of $f(\Theta_i)$.

In such cases we may use a simulation process based on random numbers, **Markov Chain Monte Carlo (MCMC)** (Gilks et al., 1996). By generating a large sample of observations from the joint distribution $f(\vec{\Theta}, \vec{W})$, the integrals of the complex distributions can be approximated from the generated data. The values are generated based on the Markov chain assumption, which states that the next generated value only depends on the present value and does not depend on the values previous to it. Based on mild regularity conditions, the chain will gradually *forget* its initial starting point and will eventually converge to a unique *stationary distribution*.

Gibbs Sampling (Gilks et al., 1996) is a popular method used for MCMC analysis. It provides an elegant way for sampling from the joint distributions of multiple variables: sample repeatedly from the distributions of one-dimensional conditionals given the current observations. Initial random values are assigned to each of the parameters. And then these values are updated iteratively based on the joint distribution, until the values settle down and converge to a stationary distribution. The values generated from the start to the point where the chain settles down are discarded and are called the *burn-in* values. The parameter estimates are based on the values generated thereafter.

5 Results

Parameter estimation was carried out using Gibb's Sampling on the WinBUGS software (Spiegelhalter et al., 2003). Values from the first 1000 iteration were discarded as burn-in. It had been observed that in most cases the chain reached the stationary distribution well within 1000 iterations. A further 5000 iterations were run to obtain the parameter estimates.

5.1 Interpretation of Parameters

The parameters of the model can be interpreted in the following manner:

- $\widetilde{\lambda_1} = 1/\lambda_1$ is the mean of an exponential distribution with parameter λ_1. $\widetilde{\lambda_1}$ measures the rate at which this term is expected in a running text corpus. $\widetilde{\lambda_1}$ determines the rarity of a term in a corpus, as it is the average gap at which the term occurs if it has not occurred recently. Thus, a large value of $\widetilde{\lambda_1}$ tells us that the term is very rare in the corpus and vice-versa.

- Similarly, $\widetilde{\lambda_2}$ measures the *within-document burstiness*, i.e. the rate of occurrence of a term given that it has occurred recently. It measures the term re-occurrence rate in a burst within a document. Small values of $\widetilde{\lambda_2}$ indicate the bursty nature of the term.

- $\widetilde{p}$ and $1 - \widetilde{p}$ denote, respectively, the probabilities of the term occurring with rate $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ in the entire corpus.

Table 1 presents some heuristics for drawing inference based on the values of the parameter estimates.

5.2 Data

We choose for evaluation, terms from the *Associated Press (AP)* newswire articles, as this is a standard corpus for language research. We picked terms which had been used previously in the literature (Church and Gale, 1995a; Church, 2000; Manning

	$\widetilde{\lambda_1}$ small	$\widetilde{\lambda_1}$ large
$\widetilde{\lambda_2}$ small	frequently occurring and common function word	topical content word occurring in bursts
$\widetilde{\lambda_2}$ large	comparatively frequent but well-spaced function word	infrequent and scattered function word

Table 1: Heuristics for inference, based on the parameter estimates.

Term	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
the	0.82	16.54	16.08	1.03
and	0.46	46.86	45.19	1.04
of	0.58	38.85	37.22	1.04
except	0.67	21551.72	8496.18	2.54
follows	0.56	80000.00	30330.60	2.64
yet	0.51	10789.81	3846.15	2.81
he	0.51	296.12	48.22	6.14
said	0.03	895.26	69.06	12.96
government	0.60	1975.50	134.34	14.71
somewhat	0.84	75244.54	4349.72	17.30
federal	0.84	2334.27	102.57	22.76
here	0.94	3442.34	110.63	31.12
she	0.73	1696.35	41.41	40.97
george	0.88	17379.21	323.73	53.68
bush	0.71	3844.68	53.48	71.90
soviet	0.71	4496.40	59.74	75.27
kennedy	0.78	14641.29	99.11	147.73
church	0.92	11291.78	70.13	161.02
book	0.92	17143.84	79.68	215.16
vietnam	0.92	32701.11	97.66	334.86
boycott	0.98	105630.08	110.56	955.42
noriega	0.91	86281.28	56.88	1516.82

Table 2: Parameter estimates of the model for some selected terms, sorted by the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ value

and Schütze, 1999; Umemura and Church, 2000) with respect to modeling different distribution, so as to present a comparative picture. For building the model we randomly selected 1% of the documents from the corpus, as the software (Spiegelhalter et al., 2003) we used is Windows PC based and could not handle enormous volume of data with our available hardware resources. As stated earlier, our model can handle both frequent function terms and rare content terms. We chose terms suitable for demonstrating this. We also used some medium frequency terms to demonstrate their characteristics.

5.3 Parameter estimates

Table 2 shows the parameter estimates for the chosen terms. The table does not show the values of $1 - \widetilde{p}$ as they can be obtained from the value of $\widetilde{p}$. It has been observed that the value $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ is a good indicator of the nature of terms, hence the rows in the table containing terms are sorted on the basis of that value. The table is divided into three parts. The top part contains very frequent (function) words. The second part contains terms in the medium frequency range. And the bottom part contains rarely occurring and content terms.

5.4 Discussion

The top part of the table consists of the very frequently occurring function words occurring frequently throughout the corpus. These statements are supported by the low values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$. These values are quite close, indicating that the occurrence of these terms shows low burstiness in a running text chunk. This supports our heuristics about the value of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$, which is small for such terms. Moderate, not very high values of $\widetilde{p}$ also support this statement, as the term is then quite likely to be generated from either of the exponential distributions (*the* has high value of $\widetilde{p}$, but since the values of λ are so close, it doesn't really matter which distribution generated the observation). We observe comparatively larger values of $\widetilde{\lambda_1}$ for terms like *yet*, *follows* and *except* since they have some dependence on the document topic. One may claim that these are some outliers having large values of both $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$. The large value of $\widetilde{\lambda_1}$ can be explained, as these terms are rarely occurring function words in the corpus. They do not occur in bursts and their occurrences are scattered, so values of $\widetilde{\lambda_2}$ are also large (Table 1). Interestingly, based on our heuristics these large values nullify each other to obtain a small value of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$. But since these cases are exceptional, they find their place on the boundary region of the division.

The second part of the table contains mostly *non-topical content terms* as defined in the literature (Katz, 1996). They do not describe the main topic of the document, but some useful aspects of the document or a nearby topical term. Special attention may be given to the term *george*, which describes the topical term *bush*. In a document about *George Bush*, the complete name is mentioned possibly only once in the beginning and further references to it are made using the word *bush*, leading to *bush* being assigned as a topical term, but not *george*. The term *government* in the group refers to some newswire article about some government in any state or any country, future references to which are made using this term. Similarly the term *federal* is used to make future references to the *US Government*. As the words *federal* and *government* are used frequently for referencing, they exhibit comparatively small values of $\widetilde{\lambda_2}$. We were surprised by the occurrence of terms like *said*, *here* and *she* in the second group, as they are commonly considered as function words. Closer examination revealed the details. *Said* has some dependence on the document genre, with respect to the content and reporting style. The data were based on newswire articles about important people and events. It is true, though unfortunate, that the majority of such people are male, hence there are more articles about men than women (*he* occurs $757,301$ times in $163,884$ documents as the 13^{th} most frequent term in the corpus, whereas *she* occurs $164,030$ times in $48,794$ documents as the 70^{th} frequent term). This explains why *he* has a smaller value of $\widetilde{\lambda_1}$ than *she*. But the $\widetilde{\lambda_2}$ values for both of them are quite close, showing that they have similar usage pattern. Again, newswire articles are mostly about people and events, and rarely about some location, referenced by the term *here*. This explains the large value of $\widetilde{\lambda_1}$ for *here*. Again, because of its usage for referencing, it re-occurs frequently while describing a particular location, leading to a small value of $\widetilde{\lambda_2}$. Possibly, in a collection of "travel documents", *here* will have a smaller value of $\widetilde{\lambda_1}$ and thus occur higher up in the list, which would allow the model to be used for characterizing genre.

Terms in the third part, as expected, are *topical content terms*. An occurrence of such a term defines the topic or the main content word of the document or the text chunk under consideration. These terms are rare in the entire corpus, and only appear in documents that are about this term, resulting in very high values of $\widetilde{\lambda_1}$. Also low values of $\widetilde{\lambda_2}$ for these terms mean that repeat occurrences within the same document are quite frequent; the characteristic expected from a topical content term. Because of these characteristics, based on our heuristics these terms have very high values of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$, and hence are considered the most informative terms in the corpus.

5.5 Case Studies

Here we study selected terms based on our model. These terms have been studied before by other researchers. We study these terms to compare our findings with previous work and also demonstrate the range of inferences that may be derived from our model.

5.5.1 somewhat *vrs* boycott

These terms occur an approximately equal number of times in the AP corpus, and *inverse document frequency* was used to distinguish between them (Church and Gale, 1995a). Our model also gives approximately similar rates of occurrence ($\widetilde{\lambda_1}$) for these two terms as shown in Table 2. But the reoccurrence rate, $\widetilde{\lambda_2}$, is 110.56 for *boycott*, which is very small in comparison with the value of 4349.72 for *somewhat*. Hence based on this, our model assigns *somewhat* as a rare function word occurring in a scattered manner over the entire corpus. Whereas *boycott* is assigned as a topical content word, as it should be.

5.5.2 follows *vrs* soviet

These terms were studied in connection with fitting Poisson distributions to their term distribution (Manning and Schütze, 1999), and hence determining their characteristics[1]. In our model, *follows* has large values of both $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ (Table 2), so that it has the characteristics of a rare function word. But *soviet* has a large $\widetilde{\lambda_1}$ value and a very small $\widetilde{\lambda_2}$ value, so that it has the characteristics of a topical content word. So the findings from our model agree with the original work.

5.5.3 kennedy *vrs* except

Both these terms have nearly equal *inverse document frequency* for the AP corpus (Church, 2000; Umemura and Church, 2000) and will be assigned equal weight. They used a method (Kwok, 1996) based on average-term frequency to determine the nature of the term. According to our model, the $\widetilde{\lambda_2}$ value of *kennedy* is very small as compared to that for *except*. Hence using the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ measure, we can correctly identify *kennedy* as a topical content term

[1] The original study was based on the *New York Times*, ours on the *Associated Press* corpus

and *except* as an infrequent function word. This is in agreement with the findings of the original analysis.

5.5.4 noriega *and* said

These terms were studied in the context of an adaptive language model to demonstrate the fact that the probability of a repeat occurrence of a term in a document defies the "bag of words" independence assumption (Church, 2000). The deviation from independence is greater for content terms like *noriega* as compared to general terms like *said*. This can be explained in the context of our model as *said* has small values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$, and their values are quite close to each other (as compared to other terms, see Table 2). Hence *said* is distributed more evenly in the corpus than *noriega*. Therefore, *noriega* defies the independence assumption to a much greater extent than *said*. Hence their findings (Church, 2000) are well explained by our model.

6 Conclusion

In this paper we present a model for term re-occurrence in text based on gaps between successive occurrences of a term in a document. Parameter estimates based on this model reveal various characteristics of term use in a collection. The model can differentiate a term's dependence on genre and collection and we intend to investigate use of the model for purposes like genre detection, corpus profiling, authorship attribution, text classification, etc. The proposed measure of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ can be appropriately adopted as a means of feature selection that takes into account the term's occurrence pattern in a corpus. We can capture both within-document burstiness and rate of occurrence of a term in a single model.

References

A. Bookstein and D.R Swanson. 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25:312–318.

K. Church and W. Gale. 1995a. Inverse document frequency (idf): A measure of deviation from poisson. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130.

K. Church and W. Gale. 1995b. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

K. Church. 2000. Empirical estimates of adaptation: The chance of two noriega's is closer to p/2 than p^2. In *COLING*, pages 173–179.

Anne De Roeck, Avik Sarkar, and Paul H Garthwaite. 2004a. Defeating the homogeneity assumption. In *Proceedings of 7th International Conference on the Statistical Analysis of Textual Data (JADT)*, pages 282–294.

Anne De Roeck, Avik Sarkar, and Paul H Garthwaite. 2004b. Frequent term distribution measures for dataset profiling. In *Proceedings of the 4th International conference of Language Resources and Evaluation (LREC)*, pages 1647–1650.

Alexander Franz. 1997. Independence assumptions considered harmful. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 182–189.

A. Gelman, J. Carlin, H.S. Stern, and D.B. Rubin. 1995. *Bayesian Data Analysis*. Chapman and Hall, London, UK.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice.* Interdisciplinary Statistics Series. Chapman and Hall, London, UK.

Slava M. Katz. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–60.

A Kilgarriff. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of ACL-SIGDAT Workshop on very large corpora*, Hong Kong.

K. L. Kwok. 1996. A new method of weighting query terms for ad-hoc retrieval. In *SIGIR*, pages 187–195.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Christian. P. Robert. 1996. Mixtures of distributions: inference and estimation. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 441–464.

D.J. Spiegelhalter, A. Thomas, N. G. Best, and D. Lunn. 2003. Winbugs: Windows version of bayesian inference using gibbs sampling, version 1.4.

K. Umemura and K. Church. 2000. Empirical term weighting and expansion frequency. In *Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 117–123.