Computer Speech and Language (1997) **11,** 275–306

On topic identification and dialogue move recognition

Philip N. Garner

Speech Research Unit, DERA Malvern, St. Andrews Rd, Malvern, Worcestershire WR14 3PS, U.K.

Abstract

Dialogue move recognition is cited as being representative of a class of problem which may be of concern in data driven natural language processing. The dialogue move recognition problem is formulated as a keyword-based topic identification problem, and is shown to be sensitive to the issue of unknown vocabulary. A model based on the multiple Poisson distribution is shown to alleviate the unknown vocabulary issue, subject to the assumption that the occurrence of keywords represents a small fraction of the data. A keyword selection strategy is derived to ensure this assumption is valid. It is shown that a modified version of Zipf's law provides a suitable prior probability distribution for keywords, and that its inclusion increases classification performance.

1. Introduction

A Spoken Language Understanding (SLU) system can be thought of as consisting of several different parts; signal processing, a speech recognizer, a language model and finally a natural language or dialogue management module. Generally speaking, current approaches to natural language processing (NLP) tend to be very hand crafted, requiring large amounts of prior knowledge about the structure of language. In stark contrast to this, current speech recognition technology is almost completely data driven. The hypothesis is that SLU technology could be improved by extending the use of data driven methods beyond the speech recognizer into the NLP and dialogue modules.

Several authors have made some progress in this area for specific applications: In the ATIS domain (Cohen, 1995), Schwartz, Miller, Stallard & Makhoul (1996) have developed a model they call a Hidden Understanding Model with the appealing symmetry of modelling higher order semantic features in a similar manner to the way the acoustic features are modelled. Pieraccini and Levin (1995) have developed a system called CHRONUS (Conceptual Hidden Representation Of Natural Unconstrained Speech), which also uses Markovian models to describe semantic meaning. The work of Gorin (1995) is also highly relevant. Several laboratories are also working on data driven dialogue modules for the VERBMOBIL project: Reithinger and Maier (1995)

0885–2308/97/040275 + 32 $25.00/0/la970032

describe a statistical dialogue model for predicting dialogue events, and Schmitz and Quantz (1996) show that knowledge of dialogue acts is necessary in a translation system.

This work is concerned with methods that may be useful in a data driven SLU scenario, without necessarily defining the scenario. Dialogue act recognition provides a convenient test-bed for such methods. The particular database we use is the HCRC map task corpus (Andersen *et al.*, 1991), which has been annotated at the dialogue move level. Dialogue moves are discussed by Kowtko, Isard & Doherty (1993). The basis of dialogue moves is that when two people engage in a conversation they play a series of games, with constituent moves, in order to impart some piece of information. In the particular case of the map task corpus, 12 distinct moves have been identified; many more are identified in the VERBMOBIL project (Jekat *et al.*, 1995).

Dialogue move recognition involves classification of an input utterance, be it acoustic or text (in this paper only text is considered), into one of M categories, and in this sense the problem is identical to that of topic identification. In its simplest form, topic identification is a two class problem, where the classes are referred to as "wanted" and "unwanted". The input can be the word level output of a speech recognizer (Carey & Parris, 1995), or acoustic features (Nowell & Moore, 1995). Recently, with the advent of the Switchboard corpus, the problem has been extended to the multi class domain (e.g. McDonough, Ng, Jeanrenaud, Gish & Rohlicek, 1994).

The purpose of this paper is to formalize the theory used for topic identification in the case of a closed set of M classes such that it can be applied to dialogue move recognition in a robust manner. The utility of the theory is demonstrated by applying it to the problem of dialogue move recognition on the map task corpus.

2. Probabilistic formulation of topic identification

2.1. Relationship with language modelling

Given an observation, $\mathbf{x}$, typically representing a sequence of words of a particular category, the problem is to infer the category from which it was sampled, also given some labelled training data, $\mathbf{D}$. Formally, the category is a sample m from the set $\mathscr{M} = \{m_1, m_2, \ldots, m_M\}$, and the solution is to assign $\mathbf{x}$ to the value of m resulting from

$$\max_i P(m = m_i | \mathbf{x}, \mathbf{D}).$$

This expression can be "inverted" via Bayes' theorem to yield

$$P(m = m_i | \mathbf{x}, \mathbf{D}) = \frac{P(\mathbf{x} | m = m_i, \mathbf{D}) P(m = m_i | \mathbf{D})}{\sum_{i=1}^{M} P(\mathbf{x} | m = m_i, \mathbf{D}) P(m = m_i | \mathbf{D})} \tag{1}$$

$$\propto P(\mathbf{x} | m = m_i, \mathbf{D}) P(m = m_i | \mathbf{D}).$$

Notice that $P(x | m = m_i, \mathbf{D})$ is a class dependent language model (LM); this can be made more clear by considering the speech recognition problem. In a speech recognizer, one is presented with an acoustic representation, $\mathbf{a}$, of a sequence of words to be recognized (an utterance). A probability, $P(\mathbf{w} | \mathbf{a}, \mathbf{D})$, must be attached to a hypothesized sequence,

MATRIX I. Loss matrix in topic identification

	Unwanted	Wanted
Treat as unwanted	L_{UU}(OK)	L_{WU}(False reject)
Treat as wanted	L_{UW}(False accept)	L_{WW}(OK)

w, of words which could have generated **a** originally. This probability can again be expanded using Bayes' rule:

$$P(\mathbf{w}|\mathbf{a}, \mathbf{D}) \propto P(\mathbf{a}|\mathbf{w}, \mathbf{D})P(\mathbf{w}|\mathbf{D}),$$

the final term being the (class independent) LM. Substituting **x** for **w**, and conditioning the LM term on some class highlights the similarity. Topic identification, then, can be thought of as discriminative language modelling.

2.2. The two class case

The two class case is worthy of particular mention as it is traditionally formulated from a decision theoretic point of view: Bayesian decision theory requires utility to be attached to combinations of classifications and actions, that is, define a loss matrix **L** with elements L_{ij} being some notional loss associated with performing action j when x belongs to class i (see Matrix I). If $j = W$ denotes "treat as wanted", (for instance, have an operator listen to a report), and $j = U$ denotes "treat as unwanted" (for instance, ignore the report) then L_{WU} is the loss associated with treating x as unwanted when it is actually wanted. For the time being, if $\mathscr{M}$ is redefined as $\mathscr{M} = \{W, U\}$; the expected loss when assigning x to class W is then

$$\begin{aligned} L_W &= L_{WW}P(m = W|\mathbf{x}, \mathbf{D}) + L_{UW}P(m = U|\mathbf{x}, \mathbf{D}) \\ &= L_{WW}\frac{P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} + L_{UW}\frac{P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} \end{aligned}$$

and similarly for L_U. For readability, $P(m = W)$ has been abbreviated to $P(W)$, and similarly for $P(U)$.

To minimize expected loss when there are only two classes, it follows that a decision rule is to classify **x** as W if and only if

$$\begin{aligned} &L_{WW}\frac{P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} + L_{UW}\frac{P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} \\ &< L_{WU}\frac{P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} + L_{UU}\frac{P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})}, \end{aligned}$$

or more simply:

$$(L_{UW} - L_{UU})P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D}) < (L_{WU} - L_{WW})P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D}).$$

It is generally assumed that the loss due to an incorrect classification is greater than that due to a correct classification, that is $L_{ij}>L_{ii}$, in which case the above expression reduces to:

$$\frac{P(\mathbf{x}|W,\mathbf{D})}{P(\mathbf{x}|U,\mathbf{D})}>\frac{P(U|\mathbf{D})}{P(W|\mathbf{D})}\cdot\frac{(L_{UW}-L_{UU})}{(L_{WU}-L_{WW})}. \tag{2}$$

In a real application, the L terms would be set by someone with some knowledge of the application, and the probabilities on the right hand side (the priors) would be inferred from the data. For evaluation purposes, however, the L terms are not known and the data is often weighted in favour of the wanted category, so the true prior is unknown; the whole right hand side is generally replaced by a single parameter, λ, which is varied over its range to produce a receiver operating characteristic (ROC) curve.

2.3. The multi-class case

The theory in the previous section assumes two classes, and hence can result in a single discrimination metric. Dialogue move recognition is a multi-class problem, and can be thought of as multi-class topic identification. It is tempting to try to use the likelihood ratio as a metric for discrimination of the classes, but alas, for more than two classes, an inequality cannot be formed with a single class on either side.

In the case where one of the classes corresponds to "none of the above", i.e. a babble topic, topic identification can be formulated as $M-1$ two class problems. These can be solved with likelihood ratios and combined into a single ROC curve. Dialogue move recognition however, clearly, corresponds to a "closed set" topic identification problem. Furthermore, in topic identification, one is generally interested in whether the subject is topic or non-topic, and it is correct, and useful, to attach utility at this point. If a car driver wishes to listen to traffic information, it is perfectly reasonable to attach a large loss to missing a report. In dialogue move recognition, however, the dialogue move is not the highest level question in the chain; that might be "Put me through to someone to complain to", in which case a large loss can be attached to being put through to the wrong telephone extension.

The move recognition can be thought of as being much deeper in the chain, and there is no way a utility can be justified in this problem other than to assign zero loss to a correct classification and equal loss to all possible misclassifications. This is the same as maximizing the likelihood of the move (class). The correct output of the move recognizer is simply a probability for each move, which can be interpreted by the next stage.

Without attaching utility to the various classifications, the correct strategy is to choose the class which maximizes the probability of the class, $P(m=m_i|\mathbf{x},\mathbf{D})$, i.e. to go back to Equation (1).

3. Calculation of probabilities

3.1. Standard maximum likelihood multinomial approach

Equation (1) requires the calculation of two probabilities: the likelihood of the particular class occurring, $P(m=m_i|\mathbf{D})$, and the likelihood that the observation was generated by the model for that class, $P(\mathbf{x}|m=m_i,\mathbf{D})$.

The easiest term to calculate is the prior, $P(m=m_i|\mathbf{D})$. Note that it is a prior in the sense that it is prior to seeing the observation, **x**; it is still posterior to the data, **D**. It simply says "What's the probability that a particular class occurs". Making the simplification that each class is independent of all previous classes, the intuitive thing to do is to divide the number of times class m_i occurred in the data by the total number of observations in the data.

The probability, $P(\mathbf{x}|m=m_i, \mathbf{D})$ is more involved. For the purpose of this paper, assume that the constituent features of **x** are words, generated by repeatedly sampling a variable $w \in \mathscr{W}$, where $\mathscr{W} = \{w_1, w_2, \ldots, w_w\}$. This model is a unigram language model.

The general approach in the literature is to express the likelihood term as the joint probability of the constituent features of **x**.

$$P(\mathbf{x}|m=m_i, \mathbf{D}) = P(w=w_1, w=w_2, \ldots, w=w_K|m=m_i, D), \tag{3}$$

where $P(w=w_k)$ in this context is taken to mean the probability that w takes the value of the kth word in **x**. Given the independence assumption between words, Equation (3) can be expressed as

$$\begin{aligned} P(\mathbf{x}|m_i, \mathbf{D}) &= P(w_1, w_2, \ldots, w_K|m=m_i, \mathbf{D}) \\ &= P(w_1|m_i, \mathbf{D})P(w_2|m_i, \mathbf{D}) \cdots P(w_K|m_i, \mathbf{D}), \end{aligned}$$

the notation abbreviated slightly.

Taking "given the move type and the data" to mean "consider only the data that is of that move type", it is now possible to work out these probabilities. The intuitive method is simply to use the same method as the prior: $P(w_k|m_i, \mathbf{D})$ can be estimated by taking the number of times that word w_k occurred in the data of move type m_i, and dividing by the total number of words in all moves of that type.

3.2. An experiment

The HCRC map task corpus (Andersen *et al*., 1991) has been annotated at the dialogue move level; there are 12 move types in all. The corpus was split into a training and testing set such that no map featured in both sets; in this way, the discrimination could be attributed to the semantic qualities of the text, not the map features.

The training data were used to calculate probabilities as described in the previous section, and these were used to classify the utterances (observations) of the testing data. A confusion matrix is shown in Matrix II.

The horizontal axis represents classification bins, the vertical is the actual class of the utterances. All axes are totalled, so as an example, there were 2459 "Acknowledge" moves in the testing data, 1795 of which were correctly classified. In total 2687 moves were classified as "Acknowledge".

The classification accuracy is just over 47%; Kowtko *et al*. (1993) state that 70–80% of the moves can be correctly identified by a human (though using context too). The model accuracy is believable given that the model has independence assumptions in the move sequence and in the word sequence.

MATRIX II. An initial confusion matrix for the map task data

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1795	17	32	2	17	66	4	18	119	61	5	323	2459
Align	390	125	19	11	6	33	14	28	114	3	9	8	760
Check	29	38	273	38	46	251	40	40	209	21	37	15	1037
Clarify	7	11	54	35	15	135	7	5	111	8	31	4	423
Explain	23	23	52	15	172	43	11	9	277	82	77	2	786
Instruct	10	30	122	159	35	639	41	20	425	11	49	2	1543
Query-W	5	8	19	4	5	29	186	11	47	1	0	0	315
Query-YN	3	28	47	10	28	36	29	401	144	12	13	3	754
Ready	82	0	4	0	1	11	0	0	9	0	0	0	107
Reply-N	3	1	4	1	1	3	0	1	4	301	3	0	322
Reply-W	11	13	45	20	28	82	10	10	108	21	51	4	403
Reply-Y	329	14	25	4	21	32	3	14	35	11	8	860	1356
Total	2687	308	696	299	375	1360	345	557	1602	532	283	1221	10265
Accuracy = 47·22%													

Assuming the test set accuracy is binomially distributed (Bedworth, 1992), the 95% confidence limits for 10 265 independent testing samples are around $\pm 1\%$.

The matrix as a whole is reasonably distributed, with large values on the leading diagonal, and smaller values off it. There is a tendency, however, for a lot of utterances to be classified as "Ready". The "Ready" move is generally played at the start of the conversation, and consists of words like "right" and "okay". Given that "Acknowledge" also consists of exactly the same words, but is far more frequent, one would expect all the "Ready" moves to actually be classified as "Acknowledge". It is hypothesized in the next section that this is a symptom of the unknown word or out of vocabulary (OOV) problem.

3.3. The unknown word problem

When a new utterance is to be classified, a probability must be attached to each word in that utterance. For instance, if the utterance *Go to the left* is to be classified, the probability of each of the words must be evaluated for each of the classes. In a move such as "Instruct", which is both frequent and has a "rich" language model, most of the words *go*, *to*, *the*, and *left* are likely to have occurred in the training data, and will be given finite probabilities. In a move such as "Clarify", however, the language model is still rich but the move itself does not occur very often; in such a case, one or more of the words in the utterance to be classified may not have occurred in the training data.

Following the intuitive method, the probability of an unknown word is zero, so the probability of the utterance is zero; the utterance clearly happened, so the model is wrong. In fact, intuition can be updated: it is clear that unknown words will occur, and that their probability ought to be non-zero and will probably be less than that of the least frequent word in that category. The least frequent word that did occur will have occurred once, and a common strategy (e.g. Nowell and Moore, 1995) is to count unknown words as having occurred 0·5 times (some justification for this is hinted at in section 3.4); this is how the confusion matrix in Matrix I was generated.

This ad hoc approach to unknown words explains the bias towards "Ready": "Ready" is the least frequent move, so the probability attached to an unknown word would be 0·5 divided by some small number being the number of words in that move type in the data. Compare this with a move like "Instruct", where the unknown word probability is 0·5 divided by a much larger number, due to the rich and frequent nature of that move type. Now imagine that a completely new word occurs in the utterance to be classified: a new map feature, or a nuance of a new talker. The new word will be given a much larger probability by the least frequent class.

3.4. Dice throwing

The OOV phenomenon is one of the main problems in language modelling, and there is a large amount of literature on the subject. In general, the solution is to apply a statistical smoothing function to the word probabilities, although this can involve a lengthy cross-validation procedure to determine parameters; a recent reference is Ney, Essen and Kneser (1995). The following sections show that for the task of discrimination, a mathematically more attractive approach is available.

When one calculates a probability by dividing the number of occurrences of interest

by the total number of occurrences, one is implicitly assuming a dice throwing model. Statistically, the problem is the same as that of coin tossing or drawing coloured balls from an urn and replacing them. If the number of occurrences of interest is represented by n and the total number of occurrences by N, then n/N is the maximum likelihood (ML) estimate of the true probability. ML estimates traditionally get more accurate as n and N get large, and fall over completely as n tends to zero.

With small databases, especially cases where n is ever close to zero, it becomes necessary to incorporate prior knowledge in some way. This is traditionally done in classical statistics by assuming some distribution and smoothing the observations to that distribution. In Bayesian statistics, the prior knowledge can be incorporated explicitly, although quantifying prior knowledge is often a problem in itself.

It is instructive to consider the Bayesian formulation of the dice throwing model. Formally, such a model is a multinomial distribution. Appendix A details a Bayesian analysis of this formulation using a flat prior (that is, all combinations of bias are initially estimated to have equal probability), and proves that the result can be applied by replacing the n/N estimate with

$$\frac{n+1}{N+W},$$

where W is the total number of possible outcomes (2 for a coin, 6 for a die).

Whilst there is little justification for using a flat prior, this result is useful in that it highlights a fundamental problem: in language modelling, W is the total vocabulary of the task in question. It can be thought of as the total vocabulary of all the speakers who took part in the task. W cannot possibly be known; a study by Efron and Thisted (1976), on estimating Shakespeare's vocabulary simply proved that it depends strongly on initial assumptions. The problem has also been tackled by Fisher, Corbet and Williams (1943), Goodman (1949), Good and Toulmin (1956) and McNeil (1973). W however, is clearly large, and suggests that all probabilities calculated by the simple maximum likelihood model will be grossly overestimated.

Note that in the Bayesian "estimate", the probability when $n=0$ is half that when $n=1$, which justifies in part the $n=0{\cdot}5$ estimate in the maximum likelihood case.

3.5. The multinominal distribution and topic identification

In fact, the multinominal distribution has other problems when applied to topic identification. Without breaking the sentence down into constituent features, the two class discrimination metric is to classify $\mathbf{x}$ as wanted if and only if

$$\log\left(\frac{P(\mathbf{x}|W)}{P(\mathbf{x}|U)}\right)>\lambda.$$

Where λ represents the product of the prior ratio and the loss function ratio of Equation (2). The logarithm is generally used for practical convenience. If the words in $\mathbf{x}$ are considered to be independent, $P(\mathbf{x})$ can be broken down into the product of the word likelihoods, and the classfication rule becomes

$$\sum_{i=1}^{K} n_i \log\left(\frac{P(w_i|W)}{P(w_i|U)}\right) > \lambda \tag{4}$$

where w_i represents the ith word in **x**, and **x** is K words in length. This equation is often referred to as "accumulated usefulness".

For the purpose of topic identification, it is clear that only discriminative words need be considered, and these words are termed keywords. The keywords are immediately apparent, being those that result in extreme values of the likelihood ratio. In conventional topic identification, the assumption is made that it is only necessary to compute probabilities for these discriminative words, simply ignoring the others. In fact, it is the probability of the whole utterance that is required.

A more subtle, but very important failure of the multinomial distribution in keyword identification is best demonstrated by example. Consider the problem of spotting weather reports in radio broadcasts. Words that maximize the likelihood ratio are likely to be *rain*, *snow*, *north* and *south*, whilst words that minimize it might be *minister*, *stockmarket* and *Ambridge*. Which class does *the cat sat on the mat* belong in? The purely keyword-based accumulated usefulness equation falls down here, having no evidence whatsoever to make a decision upon. The proper language modelling solution uses default probabilities for unknown words, but these probabilities will be higher for the least frequent class, hence favouring weather forecast given a properly representative database.

What should be acknowledged here is the absence of keywords. The multinomial model correctly applied does this by noticing the presence of other words that are not keywords, but it cannot do this correctly as it does not know the vocabulary. What is needed is a model that explicitly acknowledges zero occurrences of something, whilst ignoring words that it has no knowledge of.

4. Removing the unknown vocabulary problem

4.1. The multiple Poisson distribution

The Poisson distribution was originally derived as an approximation to the binomial distribution. The following brief derivation shows how the multiple Poisson distribution can be derived from the multinomial distribution.

If ρ_i is defined to be the underlying probability of the event $w = w_i$, then the multinomial distribution is

$$P(\mathbf{n}|\boldsymbol{\rho}) = \frac{N!}{n_1! n_2! \cdots n_W!} \rho_1^{n_1} \rho_2^{n_2} \cdots \rho_W^{n_W}.$$

where $\boldsymbol{\rho}$ is the vector $(\rho_1, \rho_2, \ldots, \rho_W)$, and **n** is the vector $(n_1, n_2, \ldots, n_W)$. The sum of the components of $\boldsymbol{\rho}$ is constrained to be unity.

Making the substitution $\lambda_i = N\rho_i$, and replacing ρ_W with the sum to unity constraint,

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{N!}{n_1!n_2!\cdots n_W!}\left(\frac{\lambda_1}{N}\right)^{n_1}\left(\frac{\lambda_2}{N}\right)^{n_2}\cdots\left(\frac{\lambda_{W-1}}{N}\right)^{n_{W-1}}$$

$$\times\left(1-\frac{\lambda_1}{N}-\frac{\lambda_2}{N}-\cdots-\frac{\lambda_{W-1}}{N}\right)^{n_W}.$$

rearranging yields

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{\lambda_1^{n_1}\lambda_2^{n_2}\cdots\lambda_{W-1}^{n_{W-1}}}{n_1!n_2!\cdots n_{W-1}!}\frac{N(N-1)(N-2)\cdots(n_W+1)}{N^{N-n_W}}$$

$$\times\left(1-\frac{\lambda_1+\lambda_2+\cdots+\lambda_{W-1}}{N}\right)^{n_W}.$$

it can be shown that as $n\to\infty$

$$\left(1-\frac{x}{n}\right)^n\to e^{-x},$$

so the limiting case turns out to be

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{\lambda_1^{n_1}\lambda_2^{n_1}\cdots\lambda_{W-1}^{n_{W-1}}}{n_1!n_2!\cdots n_{W-1}!}e^{-\lambda_1-\lambda_2-\cdots-\lambda_{W-1}},$$

where n does not contain n_W, and $\boldsymbol{\lambda}$ is the obvious thing. Note that λ_W and n_W have disappeared. This is just the product of $W-1$ independent Poisson distributions—the multiple Poisson distribution. The approximation is valid if N is large and n_W is also large compared to the sum of the other n.

In fact, the key point here is that ρ_W does not exist in the Poisson distribution. In the multinomial case there is a certain amount of redundancy in that a d dimensional multinomial actually has the constraint that all the d probabilities add to one; it is actually a $d-1$ dimensional distribution. The redundancy in the p terms is mirrored in the n terms, in that if the sum of the n (N) is known, one of the n is consequently redundant. The Poisson distribution ties down N to a fixed (infinite) value, so n_W is redundant. In turn, this is mirrored in the λ terms.

The fact that one term disappears is useful, for example: in a keyword based system, all of the non-keywords can be grouped together and referred to as a single word, the unknown word. If it is this unknown word that is dropped in the above derivation, the result is an expression that is independent of unknown words. Given that all unknown words are grouped together, regardless of how many there are, the result is also vocabulary independent. Re-interpreting the approximations in the derivation, the multiple Poisson distribution is valid for a large training database where the number of occurrences of keywords is small.

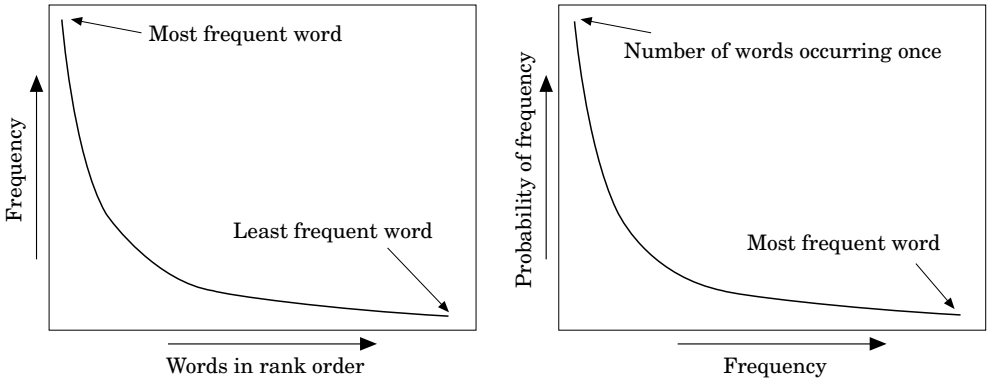

Figure 1. The Zipf plot and how to modify it to relate to probability.

The multiple Poisson is clearly the better method to use if the approximations in its derivation are valid. This distribution has two clear advantages prior to running an experiment:

(1) The absence of a word has a finite probability, that is, if any or all of the test observation word frequencies are zero then $P(\mathbf{x}|\mathbf{D})$ is finite. This means the absence of keywords can be penalized.
(2) There is a default probability for unknown words, that is, if any or all of the training word frequencies are zero then $P(\mathbf{x}|\mathbf{D})$ is still non-zero.

5. Prior information

5.1. Zipf's law

The parameters of the Poisson distribution are unknown, but information about them is available via the training data. Classically, the training data would be used to estimate the values of these parameters in a maximum likelihood sense. In more recent years the Bayesian approach has found favour, resulting in either integrating out the unknown parameters or calculating a Maximum a posteriori (MAP) estimate. Practically, the two approaches tend to produce similar results in the absence of prior information; in this case though, prior information exists in the form of Zipf's law (Zipf, 1935).

Zipf's law itself is an empirical law relating frequencies of words. If a graph is plotted of frequency as ordinate, and the words rank ordered on the abscissa, that is, the most frequent word on the left and the least frequent on the right, the points will form a smooth curve with approximately reciprocal square root form; the actual analytical form is discussed by McNeil (1973). Further, this law will hold no matter which database is used.

Such a graph is not very useful in that form, but integrating up the vertical axis produces a graph which, suitably normalized, can be interpreted as "Probability of Frequency", which in turn is the prior on the λ terms in the Poisson distribution. This is illustrated in Fig. 1, where the graph on the left is a traditional Zipf plot, and the one on the right is modified as described.

The graph on the right of Fig. 1 can be estimated with a histogram from a large

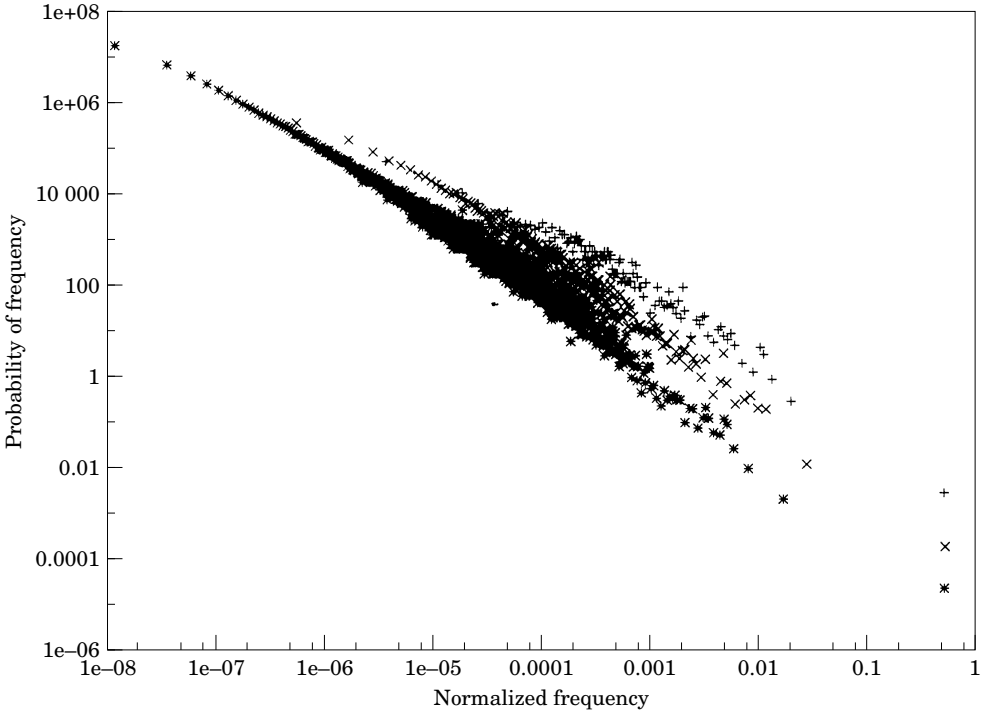

Figure 2. Modified Zipf plot for various data sources; Map task (+), King James Bible (×), Wall Street Journal (*).

dataset, and this is depicted in Fig. 2. The plots refer to the 35 million word ARPA Wall Street Journal corpus, the King James version of the Bible (less than 1 million words), and the entire HCRC map task corpus (less than 200 000 words). Three things are apparent from this plot:

(1) All the plots are straight lines with (approximately) the same gradient on log–log axes. If the gradients are indeed the same, then Zipf's law holds, and one dataset can be used as a prior for another.
(2) The smaller data sets have higher tails (the right hand end in this case). This is a well known effect, and suggests that the large dataset is a better approximation to the true distribution.
(3) The fact that they are straight lines on a double logarithmic scale implies that the real curve is of the form $y = Ax^m$, where A is some normalizing term and m is the gradient of the line.

Note that the map task plot is only shown for reference. The information in the plots is supposed to be prior information, and looking at any of the map task data is cheating, never mind looking at all of it.

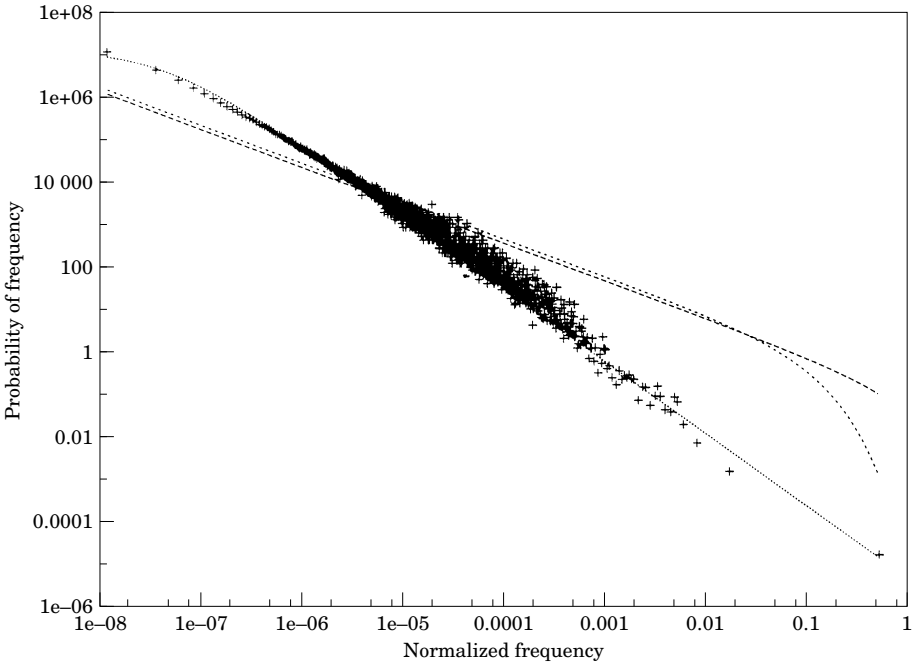

Figure 3. Various fits to the Wall Street Journal data; Wall Street Journal (+), Gamma 1 (– – –), Gamma 2 (- - -), Line Fit (· · ·).

5.2. Parameterization of prior information

To be useful as a prior distribution, some convenient parameterized form must be made to fit the Zipf plot. The gamma distribution, defined as

$$P(\lambda|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

has an x^m term, so it ought to be possible to fit a gamma distribution to this database. Fig. 3 illustrates this. The line labelled "Gamma 1" is a gamma distribution with parameters $\alpha = 0.1$ and $\beta = 0$; "Gamma 2" is the same with $\beta = 10$. Shrinking α any more has the effect of moving the whole line downwards.

There is clearly nothing to be gained from setting β to be anything other than 0; even a value of 10 introduces more curvature than can be justified. Setting α to some small value may clearly be of benefit though.

A gamma distribution has the advantage of mathematical convenience, being a conjugate prior for a Poisson distribution. Rather than insisting on conjugacy and going out of the way to make a gamma distribution fit the prior information, it ought to be possible to find a distribution that fits the prior information, but is not necessarily conjugate. In Fig. 3, it is clear that the line labelled "Line fit" fits the data much better than the gamma distributions. This is simply the line $y = Ax^{-1.7}$, where A was chosen to make the line go through the data rather than above or below it.

The gamma distribution is not proper (does not integrate to 1) if the α term falls below -1, so the line fit is out of the range of the gamma distribution. It is possible, however, to alter $y \propto x^{-1.7}$ such that it is proper by moving the whole graph to the left by an amount δ such that it actually crosses the y axis. This is equivalent to modelling the prior as

$$P(x) \propto (x+\delta)^{-\gamma},$$

where γ is the (negative) gradient of the line on double logarithmic axes and δ is some small number. The value of δ can be obtained by evaluating the normalizing constant

$$P(x) = \frac{\gamma - 1}{\delta^{1-\gamma}}(x+\delta)^{-\gamma},$$

and fitting to the histogram. In fact, δ controls the general "height" of the line on the graph.

Appendix B shows that assuming a Poisson model, where the parameters follow a gamma distribution, yields the probability of a sequence of words to be

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i-1}^{V} \frac{\Gamma(x_i + n_i + \alpha)}{\Gamma(n_i + \alpha)} \frac{(D+\beta)^{n_i+\alpha}}{(D+\beta+K)^{x_i+n_i+\alpha}}, \tag{5}$$

where $\mathbf{x} = \{\mathbf{x}_1, x_2, \ldots, x_K\}$, x_k is the number of times word w_k occurred in the observation, n_i is the number of times word w_i occurred in the data, **D**, and D is the total number of words in **D**. In practice, all of these terms are conditioned on the class too. Note that the definition of **x** has been slightly overloaded here to refer to a vector of word counts.

Matrix III shows a confusion matrix for the data with the prior set from line "Gamma 1" in Fig. 3 ($\alpha = 0{\cdot}1$, $\beta = 0$). Note that only one move is now categorized as "Ready", as was the problem with the ML multinomial. There is no category that scoops up all the unclear observations either. As a result, the overall accuracy is higher than that for the ML multinomial.

Appendix C proves that the equivalent of Equation (5) for a "log–linear" prior is

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i=1}^{V} \frac{(x_i+n_i)!}{n_i!} \frac{U(\gamma, \gamma - x_i - n_i, (D+K)\delta)}{U(\gamma, \gamma - n_i, D\delta)} \frac{D^{1+n_i-\gamma}}{(D+K)^{1+x_i+n_i-\gamma}}, \tag{6}$$

where $U(a, b, z)$ is Kummer's confluent hypergeometric function sometimes known as $\mathbf{\Psi}(a; b; z)$.

Matrix IV shows a confusion matrix for the "log–linear" prior. The classification accuracy is a little less than for a gamma prior, but within the 95% confidence limits. There is a slight bias towards classifying moves as "Ready", and this is detrimental (more wrongly than correctly classified). On the whole, though, no clear conclusions can be drawn about the relative benefits of the two priors.

In the case of the multinomial, it was clear how to assign a "flat" prior to the distribution by simply setting all the hyperparameters to 1. In this case, however, a flat

MATRIX III. Confusion matrix for measured gamma prior

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1851	25	39	2	37	86	4	23	1	58	9	324	2459
Align	397	171	28	9	24	59	14	38	0	3	9	8	760
Check	41	42	326	28	109	359	28	53	0	11	23	17	1037
Clarify	12	13	69	28	37	212	4	9	0	4	30	5	423
Explain	42	37	101	12	379	86	9	23	0	35	58	4	786
Instruct	21	36	164	74	88	1052	27	34	0	6	39	2	1543
Query-W	9	15	34	3	9	39	187	17	0	0	2	0	315
Query-YN	12	32	70	3	74	81	25	438	0	3	13	3	754
Ready	87	1	4	0	2	12	0	0	0	0	1	0	107
Reply-N	6	1	8	1	10	3	0	1	0	289	3	0	322
Reply-W	22	18	56	16	83	130	6	14	0	9	44	5	403
Reply-Y	343	15	32	2	40	38	3	14	0	3	9	857	1356
Total	2843	406	931	178	892	2157	307	664	1	421	240	1225	10265

Accuracy = 54·77%

MATRIX IV. Confusion matrix for measured "log–linear" prior

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1838	21	41	3	45	91	6	23	5	56	10	320	2459
Align	392	157	27	14	38	58	10	39	1	3	13	8	760
Check	37	41	312	37	121	343	34	56	0	10	30	16	1037
Clarify	12	17	73	32	37	199	5	11	0	4	30	3	423
Explain	42	34	96	14	395	83	6	23	0	29	60	4	786
Instruct	14	36	161	70	111	1043	30	34	3	3	36	2	1543
Query-W	9	16	33	3	9	34	193	16	0	0	2	0	315
Query-YN	11	32	65	4	105	78	26	417	0	1	14	1	754
Ready	82	1	4	0	2	13	0	0	4	0	1	0	107
Reply-N	7	1	11	2	13	2	0	3	0	280	3	0	322
Reply-W	21	19	58	14	90	128	6	12	0	6	45	4	403
Reply-Y	341	16	44	2	43	35	8	11	0	1	10	845	1356
Total	2806	391	925	195	1009	2107	324	645	13	393	254	1203	10265

Accuracy = 54·17%

prior is less clear cut. By inspection, the gamma distribution can be made flat by setting $\alpha = 1$ and $\beta = 0$. This prior essentially says that all the λ_i have an equal probability of lying anywhere from zero to infinity. This is plainly riduculous; even if all the λ_i were 1, then each observation on average would be expected to contain the entire vocabulary of the task.

Matrix V shows a confusion matrix for classification using a "flat" gamma prior and probabilities calculated using Equation (5). Considering the prior, the results are remarkably good.

A flat prior is a mathematical convenience, though. A prior should either be non-informative or represent real prior information. The next section shows that the addition of Zipf's law can be even more beneficial when training data is scarce.

6. Evaluation

In order to evaluate the different methods of assigning probabilities to observations of sequences of words, classification experiments were performed on various amounts of training data. Of the original 64 dialogues in the training set, 10 were used as a "burn in" set to ensure that at least some data from each move type was present. Classifications were then performed, adding another dialogue to the training data each time.

The hypothesis was that the approaches using prior information should perform better than those without for small amounts of training data. In addition, the use of a log–linear prior should improve on the conjugate gamma prior.

Figure 4 shows the classification performance for the various methods for the range of training data sizes used. The behaviour is broadly as predicted: all of the Poisson based measures outperform the standard multinomial, and the inclusion of prior information increases performance for small amounts of training data.

The log–linear prior does not perform as well as expected, though. In fact, the gamma prior is consistently better. The reason for this is most likely to be that the log–linear fit is only a somewhat *ad hoc* attempt to fit the Zipf plot. Whilst it fits the visible part of the plot, there is no reason to believe that it fits the unseen part to the extreme left. In fact, the log–linear curve bends downwards to cross the axis in this region, and the unseen plot is unlikely to do this. In turn, it is this region which is most important from the point of view of reverting to prior information because it contains the unknown words. The gamma distribution has two advantages here: it does not actually cross the axis, and for larger amounts of data it does not dictate a particular functional form, i.e., the functional form with a gamma prior is the same as for a flat prior.

The evaluation as shown is somewhat biased in that certain moves (notably "Acknowledge") are very easy to classify, and are very prevalent. A more objective evaluation should use a test set with equal probability of occurrence of any particular move. This is reflected in Fig. 5: A test set was constructed by randomly sampling 100 observations of each type of move from the original test set, and this knowledge was reflected by ignoring $P(m_i|\mathbf{D})$. The effect of this is to increase "performance resolution". The overall performance is lower reflecting the lower frequency of easy to classify moves, but the curves are now separated, emphasizing the importance of prior information. The 95% confidence limits on the classification rate for this smaller test set are around $\pm 3\%$.

In the latter figure, the curves for the two informative priors are coincident for a while, but separate when there is a large amount of data, although they still lie within each other's 95% confidence limits. It can be concluded at this stage that there is

MATRIX V. Confusion matrix for flat gamma prior

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1916	15	36	0	36	107	3	22	0	4	3	317	2459
Align	404	159	23	2	23	93	8	39	0	0	1	8	760
Check	55	24	303	6	101	464	12	50	0	1	6	15	1037
Clarify	15	10	73	7	32	265	1	9	0	0	6	5	423
Explain	43	18	117	2	430	134	2	28	0	1	8	3	786
Instruct	23	20	117	17	82	1233	10	30	0	0	10	1	1543
Query-W	11	19	46	0	11	50	157	21	0	0	0	0	315
Query-YN	11	25	80	0	59	118	9	447	0	0	4	1	754
Ready	88	1	4	0	1	13	0	0	0	0	0	0	107
Reply-N	188	1	10	0	35	8	0	3	0	75	2	0	322
Reply-W	19	9	51	6	105	181	2	11	0	0	16	3	403
Reply-Y	352	8	30	0	41	50	0	14	0	2	4	855	1356
Total	3125	309	890	40	956	2716	204	674	0	83	60	1208	10265

Accuracy = 54·53%

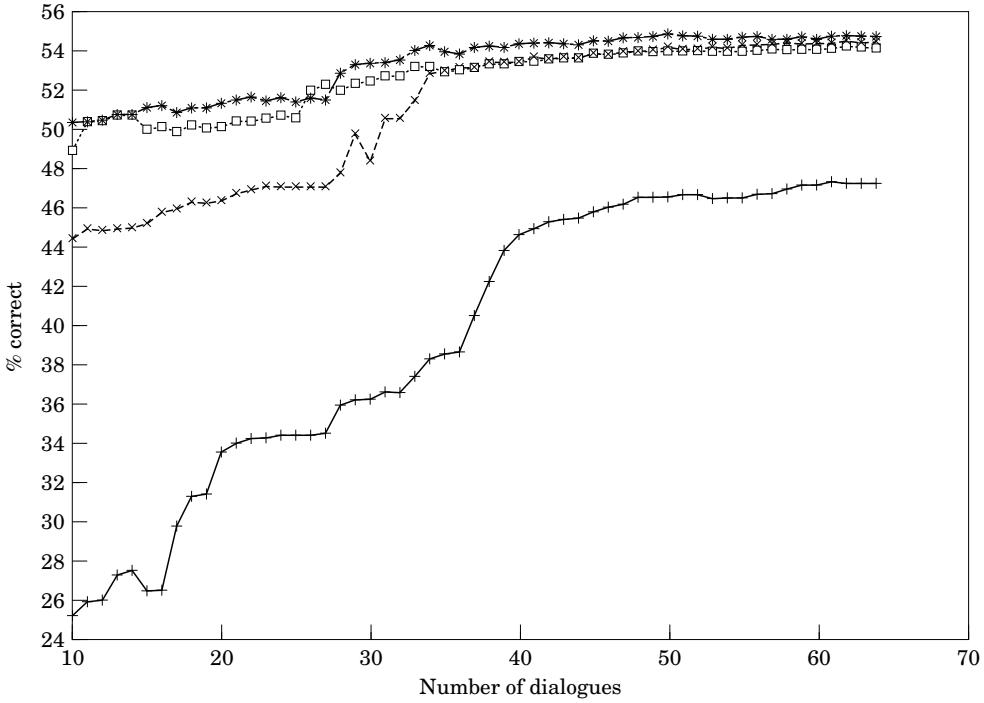

Figure 4. Classification performance vs. amount of training data for the four different probability measures. ML multinomial (+), Poisson, flat prior (×), Poisson, gamma prior (*), Poisson, log–linear prior (□).

nothing to be gained from using the log–linear prior, especially since the functional form is unnecessarily complicated.

7. Pruning the vocabulary—keyword identification

7.1. Discussion

In the preceding sections, no attempt has been made to choose those words that are discriminative. The vocabulary size has been defined as the complete vocabulary of the training data. In fact, it is clear that some words will have a much greater discriminative effect than others, and even that some words will have no discriminative effect at all. Further, the multiple Poisson approximation to the multinomial becomes more valid as the combined rate of occurrence of vocabulary words decreases. Pruning the vocabulary should therefore increase the performance of the model. One can imagine some optimal vocabulary that is small enough to allow the Poisson approximation to be valid, yet large enough to retain discriminability.

In the traditional topic identification scenario, the discriminative words are chosen as those that maximize the ratio (4), and are referred to as keywords. This ratio is referred to as usefulness because it identifies those words that are useful. In the multi-

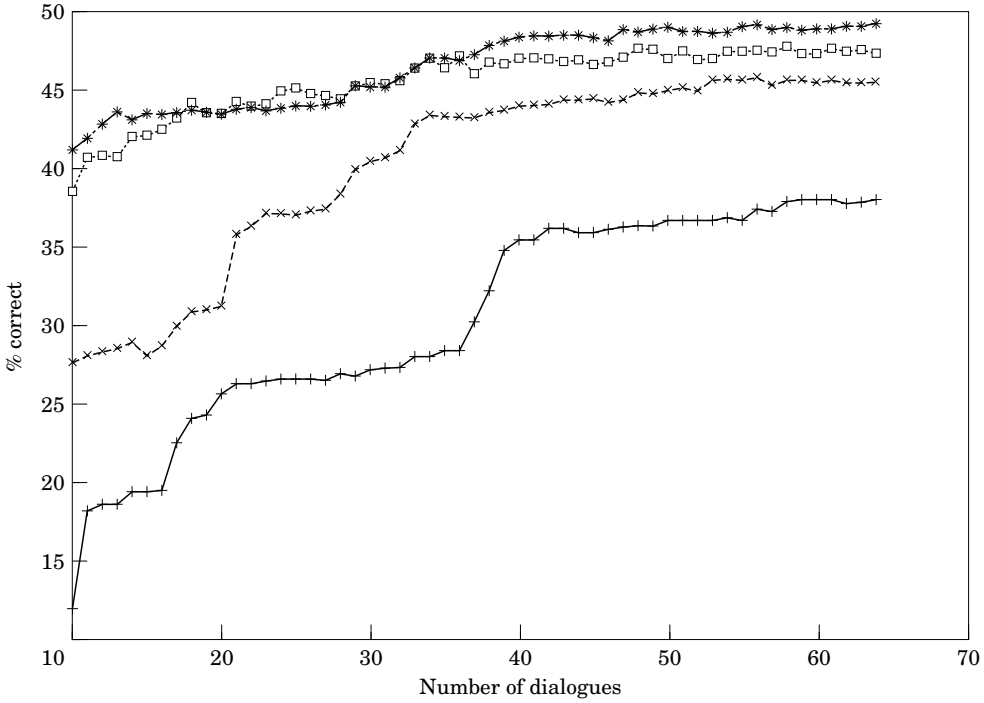

Figure 5. Classification performance vs. amount of training data for the four different probability measures using a test set with equal move distribution; ML multinomial (+), Poisson, flat prior (×), Poisson, gamma prior (*), Poisson, log–linear prior (□).

class case, however, the single ratio does not apply and it is less clear how to attach a discriminability measure to words.

7.2 A multi-class discriminability measure

The decision rule itself can be used to indicate the measure of discriminability for each word in the vocabulary: the overall decision rule is to maximize

$$P(m_i|\mathbf{x}, \mathbf{D}) = \frac{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})}$$

$$= \frac{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})}{\sum_{j=1}^{M} P(\mathbf{x}|m_j, \mathbf{D})P(m_j|\mathbf{D})}$$

over all moves in $\mathscr{M}$. This is the same as minimizing the reciprocal, in which case the summation appears in the numerator and the expression breaks down into a sum of ratios:

$$\mathscr{P}_i = \frac{P(\mathbf{x}|m_1, \mathbf{D})P(m_1|\mathbf{D})}{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})} + \frac{P(\mathbf{x}|m_2, \mathbf{D})P(m_2|\mathbf{D})}{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})}$$

$$+ \cdots + \frac{P(\mathbf{x}|m_M, \mathbf{D})P(m_M|\mathbf{D})}{P(\mathbf{x}|m_i, \mathbf{D})P(\mathrm{m}_i|\mathbf{D})},$$

which consists of easily differentiable parts.

When choosing a feature set, it is desirable to choose features that have maximum effect upon the decision rule. Consider an observation, **x**, consisting of a single word w_k. The difference in $\mathscr{P}_i$ after obsering **x** is likely to be proportional to

$$\left.\frac{\partial \mathscr{P}_i}{\partial x_k}\right|_{x_k=0}.$$

where x_k is the number of times words w_k occured in **x**. It is natural to use the expectation of this expression over all words in the vocabulary:

$$E\left(\frac{\partial \mathscr{P}_i}{\partial x_k}\right) = \sum_{k=1}^{V} \frac{\partial \mathscr{P}_i}{\partial x_k} P(w_k|m_i, \mathbf{D}),$$

and since the problem is multi-class, an expectation can also be taken over classes.

$$E\left(\frac{\partial \mathscr{P}}{\partial x_k}\right) = \sum_{i=1}^{M} E\left(\frac{\partial \mathscr{P}_i}{\partial x_k}\right) P(m_i|\mathbf{D}).$$

Interchanging the order of summation, the contribution of a particular word w_k to this expression is

$$U(w_k) = \sum_{i=1}^{M} \frac{\partial \mathscr{P}_i}{\partial x_k} P(w_k|m_i, \mathbf{D}) P(m_i|\mathbf{D}).$$

It follows that, since $\mathscr{P}_i$ is to be minimized for a correct classification, words should be chosen which minimize $U(w_k)$.

Consider first the case where the words are assumed to be distributed multinomially. The probability of a sequence of words **x** conditioned on the class, in a maximum likelihood sense, is

$$P(\mathbf{x}|m_i) = \prod_{k=1}^{K} \frac{n_{ik}}{D_i},$$

where, with a change of notation to allow conditioning on the class, there are n_{ik} words of type w_k and D_i words in total in class m_i of the training set. Differentiating as prescribed and setting $x_k = 0$ results in

$$U(w_k) = \sum_{i=1}^{M} \frac{n_{ik}}{D_i} \frac{n_i}{N} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{n_j}{n_i} \log \frac{n_{jk} D_i}{n_{ik} D_j} ,$$

where there are n_j examples of class m_j in the training data. In practice, the two n_i terms cancel, and the N is unnecessary.

In the special case of two classes, this expression can be written

$$U(w_k) = -P(m_2) P(w_k|m_1) \log \frac{P(w_k|m_1)}{P(w_k|m_2)}$$

$$-P(m_1) P(w_k|m_2) \log \frac{P(w_k|m_2)}{P(w_k|m_1)} .$$

Each of these terms is exactly the same as that given by Parris and Carey (1994), though from a much more general view, and corresponds to combining features indicative of the wanted class with features indicative of the unwanted class. For this reason, the name usefulness is retained. Curiously though, the term corresponding to class 1 is weighted by the probability of class 2 and vice-versa.

In the case of the Poisson based estimate for word probability, consider one of the terms of $\mathscr{P}_i$, comparing move j with move i:

$$\frac{\prod_{k=1}^{V} \left[\frac{\Gamma(n_{jk} + \alpha + x_k)}{\Gamma(n_{jk} + \alpha)} \frac{(D_j + \beta)^{n_{jk} + \alpha}}{(D_j + \beta + K)^{n_{jk} + \alpha + x_k}} \right] \frac{n_j}{N}}{\prod_{k=1}^{V} \left[\frac{\Gamma(n_{ik} + \alpha + x_k)}{\Gamma(n_{ik} + \alpha)} \frac{(D_j + \beta)^{n_{ik} + \alpha}}{(D_i + \beta + K)^{n_{ik} + \alpha + x_k}} \right] \frac{n_i}{N}} ,$$

rearranging yields

$$\prod_{k=1}^{V} \left[\frac{\Gamma(n_{jk} + \alpha + x_k)}{\Gamma(n_{ik} + \alpha + x_k)} \frac{\Gamma(n_{ik} + \alpha)}{\Gamma(n_{jk} + \alpha)} \frac{(D_j + \beta)^{n_{jk} + \alpha}}{(D_i + \beta)^{n_{ik} + \alpha}} \frac{(D_i + \beta + K)^{n_{ik} + \alpha + x_k}}{(D_j + \beta + K)^{n_{jk} + \alpha + x_k}} \right] \frac{n_j}{n_i} .$$

Differentiating with respect to a single x_k yields the same expression multipled by

$$\log(D_i + \beta + K) - \log(D_j + \beta + K) + \psi(n_{jk} + \alpha + x_k) - \psi(n_{ik} + \alpha + x_k) ,$$

where ψ is the digamma function. Setting all the $x_k = 0$ as before, and $K = 0$, the expression for the usefulness of word w_k becomes

$$U(w_k) = \sum_{i=1}^{M} P(w_k|m_i) \sum_{\substack{j=1 \\ j \neq i}}^{M} n_j [\log(D_i + \beta) - \log(D_j + \beta)$$

$$+ \psi(n_{jk} + \alpha) - \psi(n_{ik} + \alpha)] ,$$

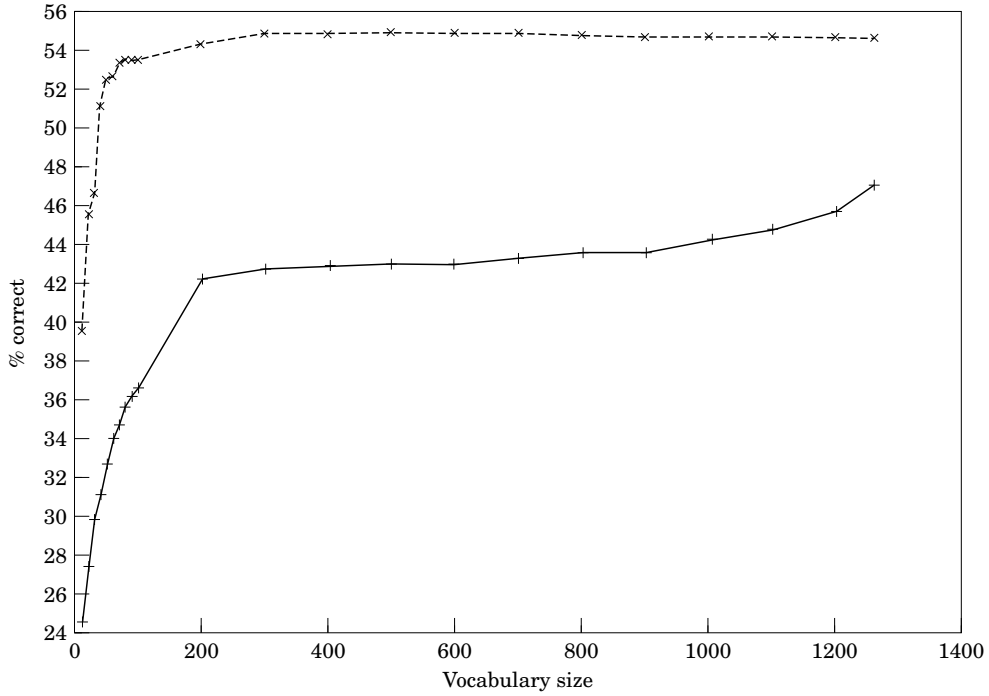

Figure 6. The effect on the classification rate of pruning the vocabulary; ML multinomial (+), Poisson, gamma prior (×).

where $P(w_k|m_i)$ is the probability of an observation consisting of the single word w_k.

The usefulness in the Poisson case is essentially the same form as that for the multinomial (the two logarithm terms can be written as the logarithm of a ratio), with the addition of the digamma functions. Digamma functions simply relate gamma functions to their first derivatives. Practically, the expression is more complicated to compute as the $P(w_k|m_i)$ term is a product over all keywords, but mathematically the result is reassuringly simple.

7.3. Evaluation

The 64 dialogues of the same training set as before were used to generate ordered lists of words for the ML multinomial and Poisson with gamma prior probability measures. Classification experiments were then performed using all 64 training dialogues and the same test situations as before, but with various vocabulary sizes. The results for the full test set are shown in Fig. 6.

Figure 7 shows the same results, but for the equally distributed test set used before. The features of the Poisson based curve are enhanced in the latter figure.

There is a definite peak in the Poisson curve at 300 keywords which corresponds to an optimal vocabulary size. To the right of this point, the performance of the multinomial continues to increase as the unknown vocabulary becomes less of a problem. The performance of the Poisson based system deteriorates though. One reason for this is

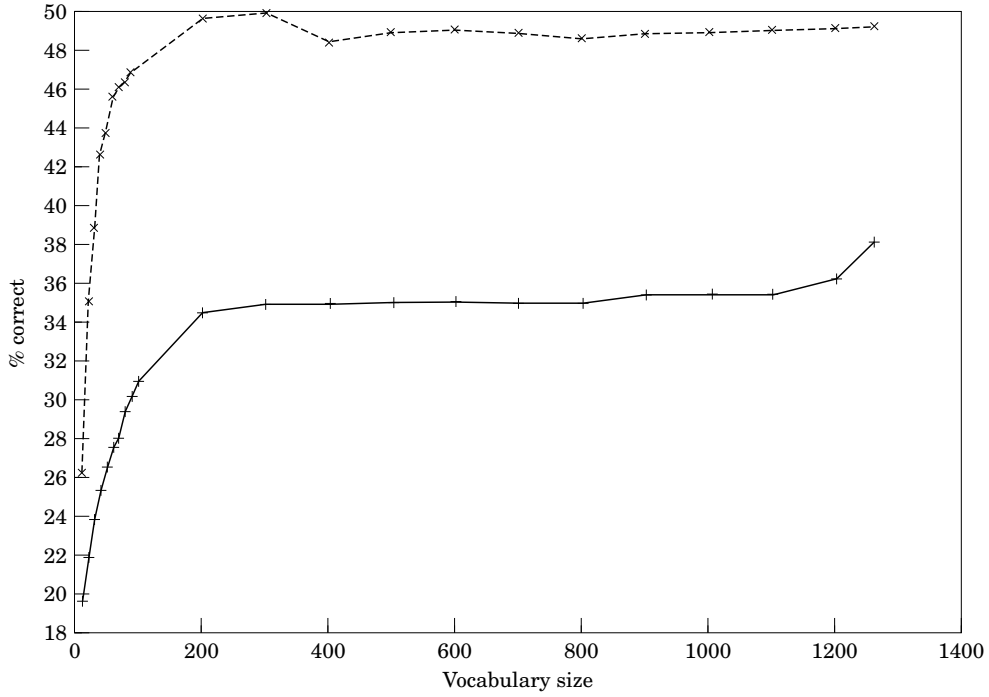

Figure 7. The effect on the classification rate of pruning the vocabulary for equal move probability; ML multinomial (+), Poisson, gamma prior (×).

clearly the failing nature of the approximation in the derivation of the multiple Poisson distribution.

The other reason is that the system is including words which only occurred once in the training set and are not discriminative. If a word only appears once, it will be treated as positively discriminative for the move in which it appears. With fewer than 300 keywords both systems deteriorate, and with fewer than 100 words there is simply not enough information to retain performance.

Results were reported by Garner and Hemsworth (1997), comparing several other methods of pruning the vocabulary. These results are summarized in Fig. 8, which shows the effect of pruning the vocabulary for a Poisson based model, using various pruning strategies. The key labels refer to measures as follows: usefulness is discussed in this paper, and the line is identical to that in Fig. 6. Mutual information,

$$I(\mathscr{M}; w_k) = \sum_{i=1}^{M} \log \frac{P(m_i|w_k)}{P(m_i)} P(m_i),$$

is the information provided by word w_k about the set of moves $\mathscr{M}$. Mutual information reversed can be thought of as $-I(\mathscr{M};w_k)$, and was used because it was not clear whether to maximize large positive or negative values. Entropy is defined as

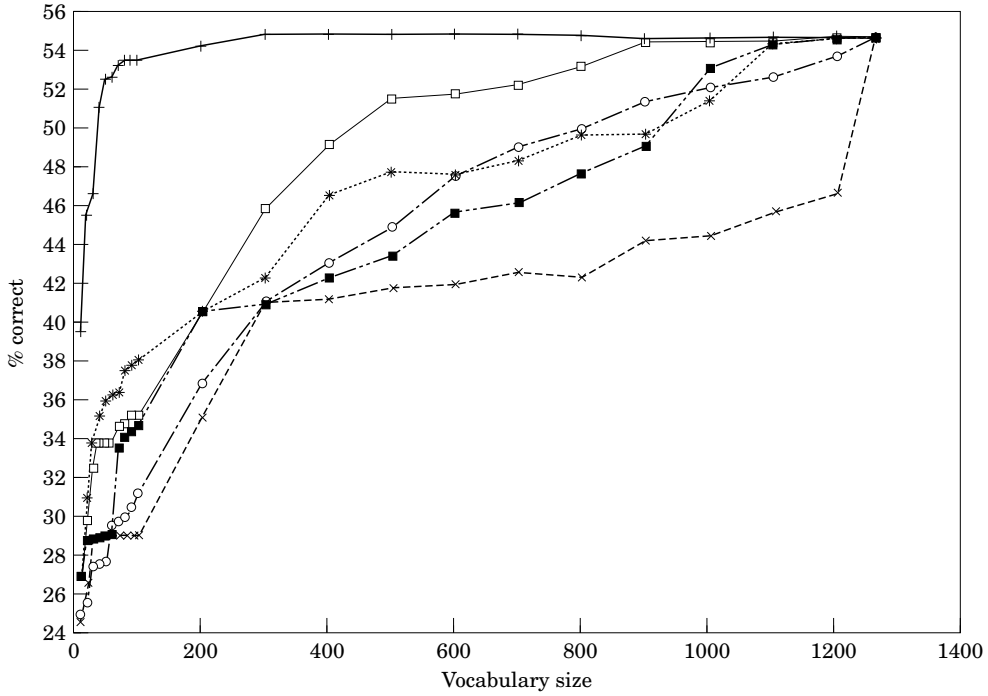

Figure 8. Comparison of various vocabulary pruning methods for the Poisson based model; Usefulness (+), Mutual information (×), Mutual information reversed (*), Entropy (□), Salience (■), Random (○).

$$I_E(\mathscr{M}; w_k) = -\sum_{i=1}^{M} P(m_i) \log P(m_i) + \sum_{i=1}^{M} P(m_i|w_k) \log P(m_i|w_k)\,,$$

and represents the increase in entropy of the ensemble $\mathscr{M}$ when word w_k is observed. Salience is defined as

$$S(\mathscr{M}; w_k) = \sum_{i=1}^{M} P(m_i|w_k) \log \frac{P(m_i|w_k)}{P(m_i)}\,,$$

and is used by Gorin (1995) in his language acquisition work. The line labelled "Random" is simply a random pruning of the vocabulary. It is clear that usefulness outperforms all other methods considered in this experiment.

8. Conclusions

This paper has outlined a consistent and rigorous approach to keyword-based topic identification, resulting in a robust enough theory to give good results when applied to dialogue move recognition. This leads to the practical result that dialogue moves can be inferred using a unigram language model to an accuracy of around 50%. The

approach, however, is more important than the actual result. Dialogue move recognition can clearly be much improved using dialogue context and acoustic intonation: Reithinger, Engel, Kipp and Klesen (1996) report an accuracy of around 40% predicting 18 intentional dialogue acts from a VERBMOBIL corpus using purely dialogue context, and Taylor, Shimodaira, Isard and Kowtko (1996) report approximately 55% on the map task corpus using purely intonation. In a more genuine experiment on a similar corpus again using intonation (Taylor, King, Isard, Wright and Kowtko, 1997), a move accuracy of around 39% is reported, which goes up to around 44% when dialogue history is considered.

In short, it is suggested that the multiple Poisson distribution is a better distribution with which to model words than a multinomial, since it alleviates the unknown vocabulary problem. This advantage far outweighs the approximate nature of the distribution. When the approximation is taken into account and only discriminative words are chosen, the multiple Poisson distribution performs even better.

Zipf's law provides a convenient subjective linguistic prior to incorporate into the posterior probability in a Bayesian sense. Its inclusion further improves performance.

This paper only goes as far as suggesting that there is some optimal vocabulary to use for a particular task; it does not suggest how to find that vocabulary, other than the obvious use of a validation set.

An assumption taken throughout is that the word and dialogue move boundaries are known, which is not the case in the context of, for instance, automatic speech recognition (ASR). Any extension to ASR would need to acknowledge the uncertain nature of the transcription, and one possible approach would be the use of lattices. The probability of a single utterance could then be evaluated as the sum of the probabilities of the words in each path through a lattice, weighted by the probability of the path. This, however, remains a subject for future research. The problem of detection of move boundaries has been addressed by Cettolo and Corazza (1997).

In addition to thanking the anonymous referees for their time and comments, I should like to extend my gratitude to the following colleagues for their help at various stages of this work. From DERA: Sue Browning, Martin Russell, Roger Moore, Mark Bedworth, Anthony Brown, Wendy Holmes and Richard Glendinning, from Forum Technology: Jörg Ueberla, and from University of Edinburgh: Steve Isard, Jacqueline Kowtko and Cathy Sotillo.
We acknowledge HCRC for the use of the dialogue game annotations for the map task corpus.

References

Anderson, A. H., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C. & Thompson, H. S. (1991). The HCRC map task corpus. *Language and Speech* **34(4)**, 351–366.

Bedworth, M. D. (1992). On the quality and quantity of data and pattern recognition. Memorandum, Defence Research Agency, St Andrews Rd, Malvern, WORCS, WR14 3PS, UK.

Carey, M. J. & Parris, E. S. (1995). Topic spotting using task independent models. In *Proceedings Eurospeech 1995, Madrid*, pp. 2133–2137.

Cettolo, M. & Corazza, A. (1997). Automatic detection of semantic boundaries. In *Proceedings Eurospeech '97*, **5**, Rhodes, pp. 919–922.

Cohen, J. R. (ed.) (1995). *Proceedings of the Spoken Language Systems Technology Workshop, Barton Creek Resort Conference Center, Austin, Texas*. ARPA, Morgan Kaufmann Publishers, Inc., San Francisco.

Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrica*, **63(3)**, 435–47.

Fisher, R. A., Corbet, A. S. & Williams, C. B. (1943). The relation between the number of species and the

number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.
Garner, P. N. & Hemsworth, A. (1997). A keyword selection strategy for dialogue move recognition and multi-class topic identification. In *Proceedings ICASSP 1997.* IEEE, **3**, 1823–1826.
Good, I. J. & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrica*, **43**, 45–63.
Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*, **20**, 572–579.
Gorin, A. L. (1995). On automated language acquisition. *Journal of the Acoustical Society of America*, **97(6)**, 3441–3461.
Gradshteyn, I. S. & Ryzhik, I. M. (1980). *Table of Integrals, Series, and Products.* Academic Press, 5th Edn.
Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M. & Quantz, J. J. (1995). Dialogue acts in VERBMOBIL. VERBMOBIL report 65.
Kowtko, J. C., Isard, S. D. & Doherty, G. M. (1993). Conversational games within dialogue. Technical report, Human Communication Research Centre, University of Edinburgh, 2 Buccleugh Place, Edinburgh EH8 9LW SCOTLAND.
McDonough, J., Ng, K., Jeanrenaud, P., Gish, H. & Rohlicek, J. R. (1994). Approaches to topic identification on the switchboard corpus. In *Proceedings ICASSP 1994*, volume **1**, pp. 385–388. IEEE.
McNeil, D. R. (1973). Estimating an author's vocabulary. *Journal of the American Statistical Association*, **68(341)**, 92–96.
Ney, H., Essen, U. & Kneser, R. (1995). On the estimation of 'small' probabilities by leaving-one-out. *IEEE transactions on pattern analysis and machine intelligence*, **17(12)**, 1202–1212.
Nowell, P. & Moore, R. K. (1995). The application of dynamic programming techniques to non-word based topic spotting. In *Proceedings Eurospeech 1995*, volume **2**, pp. 1355–1358, Madrid, Spain.
O'Hagan, A. (1994). *Bayesian Inference,* volume 2B of *Kendall's Advanced Theory of Statistics.* Edward Arnold.
Parris, E. S. & Carey, M. J. (1994). Discriminative phonemes for speaker identification. In *Proceedings ICSLP 1994*, volume **4**, pp. 1843–1846, Yokohama, Japan.
Pieraccini, R. & Levin, E. (1995). A spontaneous-speech understanding system for database query applications. In *Proceedings ESCA Workshop on Spoken Dialogue Systems*, pp. 85–88.
Reithinger, N. & Maier, E. (1995). Utilizing statistical dialogue act processing in VERBMOBIL. VERBMOBIL report 80, Reprint from *ACL-95 Proceedings*.
Reithinger, N., Engel, R., Kipp, M. & Klesen, M. (1996). Predicting dialogue acts for a speech-to-speech translation system. VERBMOBIL report 151, also in *Proceedings of ICSLP 1996*, pp 654–657.
Schmitz, B. & Quantz, J. J. (1996). Dialogue acts in automatic dialogue interpreting. VERBMOBIL report 173, also in *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, TMI-95, Leuven, pp. 33–47.
Schwartz, R., Miller, S., Stallard, D. & Makhoul, J. (1996). Language understanding using hidden understanding models. In *Proceedings ICSLP 1996*, pp. 997–1000.
Taylor, P., Shimodaira, H., Isard, S., King, S. & Kowtko, J. (1996). Using prosodic information to constrain language models for spoken dialogue. In *Proceedings ICSLP 1996*, volume **1**, pp. 216–219.
Taylor, P., King, S., Isard, S., Wright, H. & Kowtko, J. (1997). Using intonation to constrain language models in speech recognition. In *Proceedings Eurospeech 1997*, **5**, Rhodes, pp. 2763–2768.
Zipf, G. K. (1935). *The Psycho-Biology of Language*. Houghton-Mifflin, Boston.

(*Received 20 May 1997 and accepted for publication 11 August 1997*)

Appendix A: probability "estimates" from the multinomial

The following proof is by no means new, but is included in an abbreviated form for reference. For a more complete discussion of the techniques involved, see any text on Bayesian statistics.

A predictive distribution, $P(\mathbf{x}|m_i, D)$, is sought, where $\mathbf{x}$ is a sequence of words. To simplify the notation, assume that all of the calculations in this section are conditioned on $m=m_i$, i.e. $P(\mathbf{x}|\mathbf{D}) \equiv P(\mathbf{x}|m=m_i, \mathbf{D})$. Assume that $\mathbf{x}$ was generated by repeatedly sampling a variable w from the set $\mathscr{W}=\{w_1, w_2, \ldots, w_W\}$. If there are K words in $\mathbf{x}$,

$$P(\mathbf{x}|\mathbf{D}) = P(w=w_1, w=w_2, \ldots, w=w_k|\mathbf{D}).$$

If there are n_k words of type k in **D**, and N words altogether, and the unconditional probability $P(w=w_k)$ of each word is ρ_k, the core problem is to find

$$P(\mathbf{x}|\mathbf{D}) = P(\mathbf{x}|\mathbf{n}, N) = \int_0^1 \cdots_W \int_0^1 d\boldsymbol{\rho} P(\mathbf{x}|\boldsymbol{\rho}, \mathbf{n}, N) P(\boldsymbol{\rho}|\mathbf{n}, N),$$

where $\boldsymbol{\rho} = \{\rho_1, \rho_2, \ldots, \rho_W\}$, and the notation is meant to mean "integrate w.r.t. each ρ_k".

With reference to, for instance O'Hagan (1994), applying Bayes' theorem to the final term and assuming $\boldsymbol{\rho}$ follows a Dirichlet distribution, if there are x_i words of type w_i in **x**, then

$$P(\mathbf{x}|\mathbf{D}) = \int_0^1 \cdots_W \int_0^1 d\boldsymbol{\rho}\, \rho_1^{x_1}\rho_2^{x_2}\cdots\rho_W^{x_W} \frac{\rho_1^{n_1+\alpha_1-1}\rho_2^{n_2+\alpha_2-1}\cdots\rho_W^{n_W+\alpha_W-1}}{\iint d\boldsymbol{\rho}\, \rho_1^{n_1+\alpha_1-1}\rho_2^{n_2+\alpha_2-1}\cdots\rho_W^{n_W+\alpha_W-1}}$$

$$= \frac{B(n_1+\alpha_1+x_1, n_2+\alpha_2+x_2, \ldots, n_W+\alpha_W+x_W)}{B(n_1+\alpha_1, n_2+\alpha_2, \ldots, n_W+\alpha_W)},$$

where $B(a, b, \ldots)$ is the multivariate beta function.

It is actually more informative to look at this equation for a specific sequence: the terms for any word that does not appear in **x** simply cancel, leaving terms for the words that do occur, so the probability of the sequence $\{w_1, w_1, w_2, w_2\}$ is

$$\frac{n_1+1}{N+W}\,\frac{n_1+2}{N+W+1}\,\frac{n_2+1}{N+W+2}\,\frac{n_2+2}{N+W+3}.$$

A flat prior has been assumed by setting $\alpha_1=\alpha_2=\alpha_W=1$. Notice that the expression is equivalent to adding each word of the observation, **x**, to the data, **D**, before evaluating the next word. This effect is sometimes known as Laplace's rule.

Appendix B: word probabilities from the multiple Poisson

Given an observation **x** containing x_k words of type w_k, the probability $P(\mathbf{x}|\mathbf{D})$ is required. This depends upon the parameters of the Poisson distribution and is given by

$$P(\mathbf{x}|\mathbf{D}) = \int_0^\infty \cdots_V \int_0^\infty d\boldsymbol{\lambda} P(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{D}) P(\boldsymbol{\lambda}|\mathbf{D}). \tag{B.1}$$

This integral is actually a lot easier than it looks because the multiple Poisson distribution is simply the product of V independent Poisson distributions.

An important consideration here is that of "window size". In the case of the multinomial, the probability of generating n sets of l words is the same as the probability

of generating a single set of $n \times l$ words. For the multiple Poisson, the concept of window size is more important, and the two cases are different. One approach would be to choose a window size natural to the application such as length of observation. Observations are generally of different length, though, so the approach taken here is to normalize the window to be one word long, and to treat an observation length l as l separate observations.

Consider the univariate version of $P(\boldsymbol{\lambda}|\mathbf{D})$: the univariate Poisson distribution is defined to be

$$P(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}.$$

If a sequence of D trials results in observations $\mathbf{n} = \{n_1, n_2, \ldots, n_D\}$, then

$$\begin{aligned} P(\mathbf{n}|\lambda) &= \frac{\lambda^{n_1} e^{-\lambda}}{n_1!} \frac{\lambda^{n_2} e^{-\lambda}}{n_2!} \cdots \frac{\lambda^{n_D} e^{-\lambda}}{n_D!} \\ &= \frac{\lambda^{n_1+n_2+\cdots+n_D} e^{-D\lambda}}{n_1! n_2! \cdots n_{\mathrm{D}}!} \\ &= \lambda^n e^{-D\lambda}. \end{aligned}$$

The n_k can be either 1 or 0 corresponding to a word either appearing or not appearing, whereas the n refers to the number of occurrences in the D trials. Using Bayes' theorem to obtain the posterior,

$$P(\lambda|\mathbf{D}) = \frac{P(\mathbf{n}|\lambda)P(\lambda)}{\int_0^\infty d\lambda P(\mathbf{n}|\lambda)P(\lambda)}.$$

Assuming that $P(\lambda)$ is a gamma distribution, and that the normalizing terms cancel,

$$\begin{aligned} P(\lambda|\mathbf{D}) &= \frac{\lambda^n e^{-D\lambda} \lambda^{\alpha-1} e^{-\beta\lambda}}{\int_0^\infty d\lambda\, \lambda^n e^{-D\lambda} \lambda^{\alpha-1} e^{-\beta\lambda}} \\ &= \frac{\lambda^{n+\alpha-1} e^{-(D+\beta)\lambda}}{\int_0^\infty d\lambda\, \lambda^{n+\alpha-1} e^{-(D+\beta)\lambda}} \\ &= \frac{(D+\beta)^{n+\alpha}}{\Gamma(n+\alpha)} \lambda^{n+\alpha-1} e^{-(D+\beta)\lambda}. \end{aligned}$$

Assuming that all the λ_i are drawn from the same gamma distribution, the multivariate case is simply the product of these over all words, that is

$$P(\boldsymbol{\lambda}|\mathbf{D}) = \prod_{i=1}^{V} \frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \lambda_i^{n_i+\alpha-1} e^{-(D+\beta)\lambda_i}.$$

The likelihood term of (B.1) is simply the raw multivariate Poisson distribution,

$$P(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{D}) = P(\mathbf{x}|\boldsymbol{\lambda}) = \prod_{k=1}^{K} \left(\prod_{i=1}^{V} \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{\mathbf{x_{ik}}!} \right),$$

where x_{ik} is the number of words of type w_i in position k in $\mathbf{x}$, $\mathbf{x}$ being K words in length.

Equation (B.1) can be rearranged and evaluated as V independent integrals thus:

$$P(\mathbf{x}|\mathbf{D}) = \int_0^\infty \cdots \int_0^\infty d\boldsymbol{\lambda} \prod_{k=1}^{K} \left(\prod_{i=1}^{V} \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \prod_{i=1}^{V} \frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \lambda_i^{n_i+\alpha-1} e^{-(D+\beta)\lambda_i}$$

$$= \prod_{i=1}^{V} \left[\frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \int_0^\infty d\lambda_i \prod_{k=1}^{K} \left(\frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \lambda_i^{n_i+\alpha-1} e^{-(D+\beta)\lambda_i} \right]$$

$$= \prod_{i=1}^{V} \left[\frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \prod_{k=1}^{K} \frac{1}{x_{ik}!} \int_0^\infty d\lambda_i \lambda_i^{x_i+n_i+\alpha-1} e^{-(D+K+\beta)\lambda_i} \right]$$

Since x_{ik} is always either 1 or 0, and the integral is now just another gamma integral, the final form is

$$P(x|D) = \prod_{i=1}^{V} \left[\frac{\Gamma(x_i+n_i+\alpha)}{\Gamma(n_i+\alpha)} \frac{(D+\beta)^{n_i+\alpha}}{(D+\beta+K)^{x_i+n_i+\alpha}} \right].$$

In practice, this equation simplies in that if $V \gg K$, x_i will mostly be zero and the gamma functions cancel. Further, $\prod_V (\cdot)^{n_i+\alpha} = (\cdot)^{N+V_\alpha}$, so the product is only over K terms.

This distribution is related to the negative binomial distribution. Consider for the moment the terms inside the product, which can be written

$$\frac{(x_i+n_i+\alpha-1)!}{x_i!(n_i+\alpha-1)!} (1-p)^{n_i+\alpha} p^{x_i}$$

where $p = (D+\beta-1)^{-1}$. The x_i disappeared in the derivation since it was always 0 or 1. This expression is of the form

$$P(x|r, p) = \binom{r+x-1}{x} p^r q^x$$

which is the negative binomial distribution that Fisher used to count butterflies (Fisher, Corbet & Williams, 1943) and that Efron & Thisted (1976) used to model Shakespeare's output. The derivation differs from Fisher in its use of a prior distribution, and since the α terms are not necessarily integer, the normalizing term cannot be written using factorials.

Appendix C: Poisson distribution with "log–linear" prior

Following the notation and argument in Appendix B, $P(\boldsymbol{\lambda}|\mathbf{D})$ is required. This is the product of all the univariate cases, where a single univariate case is given by

$$P(\lambda|n) = \frac{\lambda^N e^{-D\lambda}(\lambda+\delta)^{-\gamma}}{\int_0^\infty d\lambda \lambda^N e^{-D\lambda}(\lambda+\delta)^{-\gamma}},$$

assuming the normalizing constants cancel.

The integral in the denominator can be solved by noticing the similarity with the integral definition of the confluent hypergeometric function (Gradshteyn & Ryzhik, 1980):

$$\Gamma(a)U(a, b, z) = \int_0^\infty e^{-zt} t^{\alpha-1}(1+t)^{b-a-1} dt.$$

Making the change of variable $t = \lambda/\delta$, the integral in the denominator becomes

$$I = \int_0^\infty dt\, \delta(\delta t)^n e^{-D\delta t}(\delta+\delta t)^{-\gamma}$$

$$= \delta^{N+1-\gamma} \int_0^\infty dt\, t^N e^{-Dt}(1+t)^{-\gamma}$$

$$= \delta^{N+1-\gamma}\Gamma(N+1)U(N+1, N+2-\gamma, D\delta).$$

Changing notation to allow for the multivariate case, an proceeding as in Appendix B,

$$P(\mathbf{x}|\mathbf{D}) = \int_0^\infty \cdots \int_V^\infty d\lambda \prod_{k=1}^{K} \left(\prod_{i=1}^{V} \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right)$$

$$\times \prod_{i=1}^{V} \frac{\lambda_i^{n_i} e^{-D\lambda_i} (\lambda_i + \delta)^{-\gamma}}{\delta^{n_i+1-\gamma} \Gamma(n_i+1) U(n_i+1, n_i+2-\gamma, D\delta)}$$

$$= \prod_{i=1}^{V} \left[\frac{\int_0^\infty d\lambda_i \prod_{k=1}^{K} \left(\frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \lambda_i^{n_i} e^{-D\lambda_i} (\lambda_i + \delta)^{-\gamma}}{\delta^{n_i+1-\gamma} \Gamma(n_i+1) U(n_i+1, n_i+2-\gamma, D\delta)} \right]$$

$$= \prod_{i=1}^{V} \left[\frac{\prod_{k=1}^{K} \frac{1}{x_{ik}!} \int_0^\infty d\lambda_i \lambda_i^{x_i + x_i} e^{-(D+K)\lambda_i} (\lambda + \delta)^{-\gamma}}{\delta^{n_i+1-\gamma} \Gamma(n_i+1) U(n_i+1, n_i+2-\gamma, D\delta)} \right].$$

Again, x_{ik} can only ever be 0 or 1. The whole expression can be simplified using the Kummer transformation

$$U(\mathrm{a}, b, z) = z^{1-b} U(1+a-b, 2-b, z).$$

In addition, some of the δ terms cancel, and the arguments to the gamma functions are always integer so factorials can be used, yielding

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i=1}^{V} \frac{(x_i+n_i)!}{n_i!} \frac{U[\gamma, \gamma - x_i - n_i, (D+K)\delta]}{U(\gamma, \gamma - n_i, D\delta)} \frac{D^{1+n_i-\gamma}}{(D+K)^{1+x_i+n_i-\gamma}}.$$

The relationship with the gamma prior expression is now evident; this expression is like that for a flat prior, but with γ as a notional "initial count" for n_i, and the addition of the ratio of confluent hypergeometric functions.