A THEORY OF WORD FREQUENCIES AND ITS APPLICATION TO DIALOGUE MOVE RECOGNITION

P. N. Garner, S. R. Browning, R. K. Moore and M. J. Russell

Defence Research Agency, St Andrews Rd, Malvern, WORCS. WR14 3PS, UK
Email: garner@signal.dra.hmg.gb

ABSTRACT

Dialogue move recognition is taken as being representative of a class of spoken language applications where inference about high level semantic meaning is required from lower level acoustic, phonetic or word based features. Topic identification is another such application. In the particular case of inference from words, the multinomial distribution is shown to be inadequate for modelling word frequencies, and the multivariate Poisson is a more reasonable choice. Zipf's law is used to model a prior distribution. This more rigorous mathematical formulation is shown to improve dialogue move classification both subjectively and quantitatively.

1. INTRODUCTION

It has been suggested [5] that a dialogue, that is, the interaction between two or more people in a conversation, can be represented as a series of moves (as a game of chess consists of alternate moves). These moves follow a natural sequence, with alternatives and counter moves. The dialogue moves dictate portions of speech that can be classified into the different move types, and may in turn dictate sensible bounds between which processing can be carried out.

The dialogue moves also form a natural part of the progression from raw acoustic data to natural language processing. Inference can proceed in either direction: down towards the acoustic recogniser or up towards the natural language processor. This paper is concerned with the latter, and in particular with the question of whether it may be possible to construct a data driven natural language processor. Dialogue move recognition can be viewed as a metric against which the contribution of dialogue moves to natural language processing can be judged.

2. AN INITIAL EXPERIMENT

2.1. Data

The HCRC map task corpus [1] has been annotated at the dialogue move level, and this database was used as an experimental vehicle. Only utterances which could be identified as belonging to one move category were used, and all non-word annotation was stripped out. Punctuation was removed, and upper case letters were converted to lower case.

The 128 dialogues were then split into training and testing sets of 64 dialogues each such that no map appeared in both sets. This was to prevent discrimination occuring on particular map features, hence forcing the use of other words more indicative of semantic meaning. The training and testing sets contained 11799 utterances and 10265 utterances respectively.

2.2. Methodology

The methodology was essentially that used in word based topic identification, outlined as follows:

The moves were assumed to be samples from a random variable $\mathcal{M} \in \{m_1, m_2, \ldots, m_M\}$; in this case, the number of possible moves, M, was 12. Given an utterance $\boldsymbol{x}$, and training data $\boldsymbol{D}$, the problem is to maximise the likelihood of the move m_i. Using Bayes's theorem,

$$P(\mathcal{M} = m_i|\boldsymbol{x}, \boldsymbol{D}) = \frac{P(\boldsymbol{x}|\mathcal{M} = m_i, \boldsymbol{D})P(\mathcal{M} = m_i|\boldsymbol{D})}{P(\boldsymbol{x}|\boldsymbol{D})}.$$

The denominator, $P(\boldsymbol{x}|\boldsymbol{D})$, is independent of the move and can be ignored.

Assuming $P(m_i)$ to be an abbreviation for $P(\mathcal{M} = m_i)$, $P(m_i|\boldsymbol{D})$ is the prior (prior to the utterance but posterior to the data), and was calculated as the number of moves of type m_i in $\boldsymbol{D}$ divided by the total number of moves in $\boldsymbol{D}$.

$P(\boldsymbol{x}|m_i, \boldsymbol{D})$ is the likelihood. Here, it was assumed that $\boldsymbol{x}$ was generated by sequentially sampling from a random variable $\mathcal{W} \in \{w_1, w_2, \ldots, w_V\}$, where V is the vocabulary of the task, and samples from $\mathcal{W}$ are independent. Hence, if $\boldsymbol{x}$ is K words in length,

$$\begin{aligned} P(\boldsymbol{x}|m_i, \boldsymbol{D}) &= P(\mathcal{W} = w_1, \mathcal{W} = w_2, \ldots, \mathcal{W} = w_K|m_i, \boldsymbol{D}), \\ &= P(w_1|m_i, \boldsymbol{D}) \\ &\quad \times P(w_2|m_i, \boldsymbol{D}) \times \cdots \\ &\quad \times P(w_K|m_i, \boldsymbol{D}). \end{aligned}$$

$P(w_k|m_i, \boldsymbol{D})$ was calculated as the number of words of type w_k in move m_i in $\boldsymbol{D}$ divided by the total number of words in move m_i in $\boldsymbol{D}$. Where the count for a word in $\boldsymbol{x}$ was zero, that word was assumed to have occured 0.5 times.

2.3. Results

Table 1 shows a confusion matrix for the classification problem so far described. The overall accuracy is 47.22%, and assuming the test set accuracy is binomially distributed [2], the 95% confidence limits for 10265 independent testing samples are around ±1%.

Note that a disproportionate number of utterances have been classified as 'Ready'. This is counter intuitive; one would expect utterances about which the system was unsure to be classified as 'Acknowledge', since that is the most frequent class. Further, 'Acknowledge', 'Ready' and 'Reply-Y' are all basically affirmative utterances ("yes"), and one would expect them to be indistinguishable at this level.

3. PROBABILITY DISTRIBUTIONS

3.1. The Multinomial

When probabilities are calculated as a relative frequency as described, one is implicitly assuming a multinomial (dice throwing) distribution. That is, if the number of words of type w_i in a move is n_i, and $N = \sum_{i=1}^{V} n_i$, then $P(w_i|\boldsymbol{D}) = n_i/N$. In fact, this is the maximum likelihood estimator of the true probability; it becomes more accurate as $N \to \infty$. In this case, though, some of the n_i are actually zero and the maximum likelihood estimator breaks down completely.

More light can be shed on the situation by considering a Bayesian formulation of the word probability problem [4]. Using a multinomial distribution with a flat Dirichlet prior, the probability of a single word w_i being drawn from $\mathcal{W}$ is

$$P(w_i|\boldsymbol{D}) = \frac{n_i + 1}{N + V}.$$

The formula now depends on V, the vocabulary of the task. This can be thought of intuitively too: Given a biassed die, but no data upon which to base an approximation, most people would agree that a good starting point would be to assume a probability of throwing any particular number to be 1/6. This is implicitly based on the prior knowledge that a die has 6 sides.

This explains the reason for assuming $n_i = 0.5$ for unseen words: the probability for $n_i = 0$ is half that for $n_i = 1$. V is large, though, and whilst it is unknown it suggests that the maximum likelihood estimate is consistently an overestimate of the true posterior probability. The largest overestimates of this word probability will occur in the class for which N is smallest; the least frequent class is 'Ready'.

3.2. The Multivariate Poisson

If the underlying probability of drawing word w_i from $\mathcal{W}$ is ω, then the multinomial distribution is

$$P(\boldsymbol{n}|\boldsymbol{\omega}) = \frac{N!}{n_1! \ldots n_V!} \omega_1^{n_1} \ldots \omega_V^{n_V}$$

where $\boldsymbol{n} = \{n_1, n_2, \ldots, n_V\}$ and $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_V\}$. Consider what would happen if this model were used to generate an infinite amount of data: It can be proved that if the ω_i are constrained to be small enough such that $N\omega_i \to \lambda_i$ as $N \to \infty$,

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

where $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_{V-1}\}$. This is the multivariate Poisson distribution.

Note that one of the ω terms has disappeared. More correctly, any of the ω terms can be made to disappear by simply grouping them into one term; the useful approach is to group all unknown words into a single ω, and have that disappear. The result is a distribution which is independent of vocabulary; indeed it can be tailored to any arbitrarily sized vocabulary.

The intuitive approach to the above derivation is to consider several throws of a die. ω_i relates to each individual throw, whereas λ_i is concerned with the rate of occurence of the feature of interest.

The probability of an utterance of K words in length using a multivariate Poisson distribution and a gamma prior can be shown [4] to be

$$P(\boldsymbol{x}|\boldsymbol{D}) = \prod_{i=1}^{W} \left(\frac{(N+\beta)^{n_i+\alpha}}{(N+\beta+K)^{n_i+\alpha+x_i}} \frac{\Gamma(n_i+\alpha+x_i)}{\Gamma(n_i+\alpha)} \right),$$

where n_i and N are the same as in the multinomial, x_i is the number of words of type w_i in $\boldsymbol{x}$, W is the number of 'keywords' and α and β are the parameters of the gamma prior. Note that this calculation refers to the probability of the whole utterance, not the product of the probabilities of the individual words.

4. PRIOR INFORMATION

4.1. Zipf's Law

Whilst it is convenient to attach a flat prior to a distribution and simply let the data decide what to do, it must be acknowledged that prior information exists in the form of Zipf's law[7]. Zipf's law itself is an empirical law relating relative frequencies. 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[6]. 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 normalised, 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 figure 1, where the graph on the left is a traditional Zipf plot, and the one on the right is modified as described.

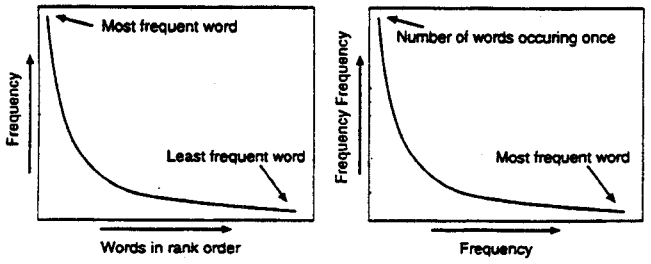

Figure 1: The Zipf plot, and how to modify it to relate to probability.

The graph on the right of figure 1 can be estimated with a histogram from a large dataset, and this is depicted in figure 2. The scatter plots refer to the King James version of the Bible, the entire radio 4 weather forecast spotting database [3], and the entire HCRC Map Task corpus. Two things are apparent from this plot:

1. All the plots are straight lines with the same gradient. If they are indeed the same, then Zipf's law holds, and one dataset can be used as a prior for another.
2. 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 normalising term and m is the gradient of the line.

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

The gamma distribution has a x^m term, so it ought to be possible to fit a gamma distribution to this database. The lines on Figure 2 illustrate this. The line labelled 'Gamma 1' is a gamma distribution with parameters $\alpha = 0.1$ and $\beta = 1$; '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. It only acts as a prior on the number of observations, which is of the order of several thousand. Even a value of 10 introduces more curvature than can be justified. Setting α to some small value may clearly be of benefit though.

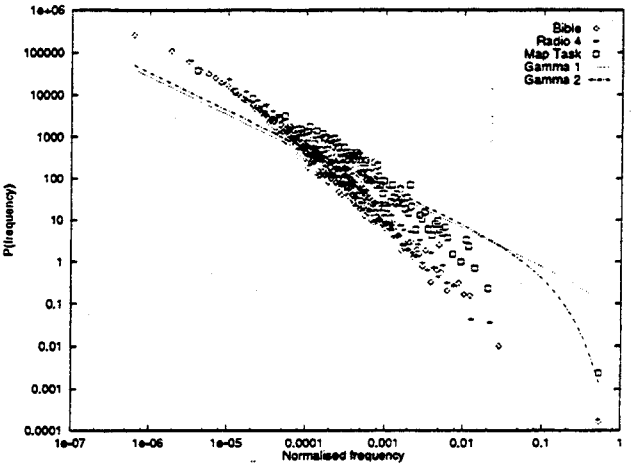

Figure 2: Modified Zipf plot for various data sources, with approximate gamma distribution fits.

5. EVALUATION

Table 2 shows a confusion matrix for the classification experiment using the Poisson based estimate with a gamma prior with α set to 0.1. The classification rate is better than the maximum likelihood case, but more importantly, the misclassifications are much better distributed. No one class seems to mop up the ambiguous observations in a disproportionate manner. In fact, nothing is classified as 'Ready', but that is understandable since that category is indistinguishable from 'Acknowledge'.

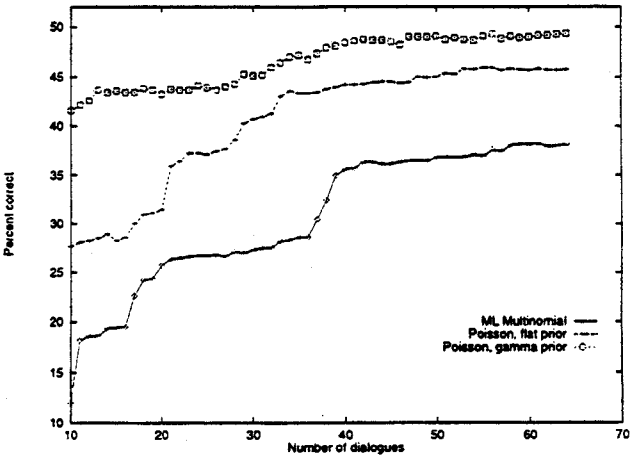

Figure 3: Classification rate as a function of amount of training data.

To evaluate the performance of the Poisson technique more fully, a test data set was constructed by randomly sampling 100 observations of each category from the test data previously described. With the classifier suitably modified for equal class membership priors, experiments were performed on training set sizes ranging from 10 to 64 dialogues. The results are shown in figure 3, confidence limits for 1200 test

samples are around ±3%. This plot is very gratifying, showing that the Poisson based estimate performs better than the maximum likelihood multinomial, and that incorporation of a Zipf's law based prior further improves performance, especially for small amounts of training data.

6. CONCLUSIONS

It has been shown that the multivariate Poisson distribution is a justifiable and more suitable distribution to model word frequencies for dialogue move recognition. Incorporation of Zipf's law as a prior follows naturally and further improves performance.

Dialogue moves can be inferred from their constituent words to an accuracy of around 50% using a very simple unigram model, implying that better performance should be possible using a more involved N-gram Markov model.

	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

Table 1: Confusion matrix for the initial experiment, Accuracy = 47.22%

	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

Table 2: Confusion matrix for the Poisson based classification, Accuracy = 54.77%

7. REFERENCES

1. Anne H. Anderson, Miles Bader, Ellen Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jaqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, and Henry S. Thompson. The HCRC map task corpus. *Language and Speech*, 34(4):351–366, 1991.
2. Mark D. Bedworth. On the quality and quantity of data and pattern recognition. Memorandum (unpublished), Defence Research Agency, St Andrews Rd, Malvern, WORCS, WR14 3PS, UK, 1992.
3. Mike J. Carey and E. S. Parris. Topic spotting using task independent models. In *Proceedings Eurospeech 95, Madrid*, pages 2133–2137, 1995.
4. Philip N. Garner. On topic spotting and dialogue move recognition. Memorandum (unpublished), Defence Research Agency, St Andrews Rd, Malvern, WORCS, WR14 3PS, UK, 1996.
5. Jaqueline C. Kowtko, Stephen D. Isard, and Gwyneth M. Doherty. Conversational games within dialogue. Technical report, Human Communication Research Centre, University of Edinburgh, 2 Buccleugh Place, Edinburgh EH8 9LW SCOTLAND, November 1993.
6. Donald R. McNeil. Estimating an author's vocabulary. *Journal of the American Statistical Association*, 68(341):92–96, March 1973.
7. G. K. Zipf. *The Psycho-Biology of Language*. Houghton-Mifflin, Boston, 1935.