A KEYWORD SELECTION STRATEGY FOR DIALOGUE MOVE RECOGNITION AND MULTI-CLASS TOPIC IDENTIFICATION

Philip N. Garner and Aidan Hemsworth

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

ABSTRACT

The concept of usefulness for keyword selection in topic identification problems is reformulated and extended to the multi-class domain. The derivation is shown to be a generalisation of that for the two class problem. The technique is applied to both multinomial and Poisson based estimates of word probability, and shown to outperform or compare favourably to various information theoretic techniques classifying dialogue moves in the map task corpus, and reports in the LOB corpus.

1. INTRODUCTION

Over the past few years, a general class of problem has arisen where inference is required about high level semantic meaning from some lower level feature set. The main manifestation of this problem is in topic identification, where a system is required to detect when a 'Wanted' topic is being discussed in a stream of largely 'Unwanted' material. The source data can be text, the word level output of a speech recogniser [1], or acoustic or phonetic level data, for instance [2].

Topic identification is traditionally a two class problem, but can easily be extended to multi-class by partitioning the 'Wanted' class into sub-classes, for example [3]. The same methods have been used to do dialogue move recognition by other authors, eg. [4] and [5]; here the problem is specified in terms of spoken language understanding, but the methodology is exactly the same as in topic identification.

In all of these problems, one approach is to identify a set of 'keywords' or 'key features' $\mathcal{W} = \{w_1, w_2, \ldots, w_W\}$, which are sufficient to distinguish the chosen classes. This reduced dictionary is then used to build language models each indicative of a particular class; the number of key features (dictionary size) is a trade off between complexity and performance. In the two class case, the decision rule is to assign the observation, $\boldsymbol{x} = w_1, w_2, \ldots, w_K$, to the Wanted class, C_W, iff

$$\prod_{k=1}^{K} \frac{P(w_k|C_W)}{P(w_k|C_U)} > \lambda,$$

where the w_k are the independent constituent features of the observation, and λ is some threshold. In this paper, the features are words.

The metric dictating the choice of features follows directly from the decision rule: choose features which maximise the probability ratio inside the product (weighted by the frequency of occurrence of those features). For this reason, this weighted ratio has been termed 'Usefulness' [6].

The decision rule in the multi-class case is more complex. If the set of M classes is $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$, then the decision rule is to maximise

$$\max_i \frac{P(\boldsymbol{x}|m_i)P(m_i)}{P(\boldsymbol{x})}.$$

It is clear that a simple inequality cannot be formed resulting in a simple ratio.

2. INFORMATION THEORETIC MEASURES

It is reasonable to assume that keywords should be chosen which maximise some measure of information. Less clear, though, is which measure; three possible measures can be identified as follows.

Quoting Gallager [9], if m is a sample from $\mathcal{M}$ and w is a sample from $\mathcal{W}$, *the information provided about the event* $m = m_i$ *by the occurence of the event* $w = w_k$ *is*

$$I(m_i; w_k) = \log \frac{P(m_i|w_k)}{P(m_i)}.$$

This is the mutual information between the two events. To extend the measure to apply over all classes, consider the expectation over classes:

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

Mutual information expressed in this way is very similar to the expression for the change in entropy (with one changed term):

$$\begin{aligned} I_E(\mathcal{M}; w_k) &= -\sum_{i=1}^{M} P(m_i) \log P(m_i) \\ &\quad + \sum_{i=1}^{M} P(m_i|w_k) \log P(m_i|w_k). \end{aligned}$$

This has the intuitively appealing quality of representing the increase in entropy of the ensemble $\mathcal{M}$ when word w_k is observed.

Salience has been used by Gorin [8] to rank words in order of importance to classify actions in a dialogue management system. Salience is defined as

$$S(\mathcal{M}; w_k) = \sum_{i=1}^{M} P(m_i|w_k) I(m_i; w_k).$$

Writing the three measures $I(\mathcal{M}; w_k)$, $I_E(\mathcal{M}; w_k)$ and $S(\mathcal{M}; w_k)$, which shall be referred to as mutual information, entropy and salience respectively, as

$$\sum_{i=1}^{M} P(m_i) \log P(m_i|w_k) - \sum_{i=1}^{M} P(m_i) \log P(m_i)$$

$$\sum_{i=1}^{M} P(m_i|w_k) \log P(m_i|w_k) - \sum_{i=1}^{M} P(m_i) \log P(m_i)$$

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

it is clear that they are intimately related, the only difference being whether the raw information term (the logarithm term) is weighted by $P(m_i)$ or $P(m_i|w_k)$.

Gorin [8] uses some standard smoothed relative frequencies to estimate the probabilities above. In this paper, we use the maximum likelihood estimate

$$P(m_i) = \frac{n_i}{N},$$

where n_i is the number of occurrences of class m_i in the training data, and N is the total number of occurrences. The posterior measure $P(m_i|w_k)$ is evaluated via Bayes's theorem:

$$P(m_i|w_k) = \frac{P(w_k|m_i)P(m_i)}{\sum_{i=1}^{M} P(w_k|m_i)P(m_i)}.$$

3. USEFULNESS

The decision rule itself can also indicate a measure of 'usefulness' for each possible word: The multi-class decision rule is to maximise

$$P(m_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|m_i)P(m_i)}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|m_i)P(m_i)}{\sum_{j=1}^{M} P(\boldsymbol{x}|m_j)P(m_j)}.$$

Denoting the reciprocal of this expression by $\mathcal{P}_i$, the problem is the same as minimising

$$\mathcal{P}_i = \frac{P(\boldsymbol{x}|m_1)P(m_1)}{P(\boldsymbol{x}|m_i)P(m_i)} + \frac{P(\boldsymbol{x}|m_2)P(m_2)}{P(\boldsymbol{x}|m_i)P(m_i)} + \cdots + \frac{P(\boldsymbol{x}|m_M)P(m_M)}{P(\boldsymbol{x}|m_i)P(m_i)},$$

which consists of easily differentiable parts. It is reasonable to assume that discriminative keywords will be those which lead to a high rate of change of this probability. Consider the expected rate of change of $\mathcal{P}_i$ when a new feature or word is considered: By definition,

$$E\left(\frac{\partial \mathcal{P}_i}{\partial x_k}\right) = \sum_{k=1}^{W} \frac{\partial \mathcal{P}_i}{\partial x_k} P(w_k|m_i),$$

where there are x_k words of type w_k in $\boldsymbol{x}$. The new feature is unknown, and this is accounted for by integrating over all possible features. The features or words which have maximum effect upon the decision rule are those which minimise this expectation (largest negative value). It is clear that the most useful words are those which minimise

$$\frac{\partial \mathcal{P}_i}{\partial x_k} P(w_k|m_i).$$

This can be evaluated with all the $x_k = 0$, embodying the assumption that the usefulness of the occurance of a word is independent of the number of times it has occurred already.

Thus far, the theory only addresses choosing keywords to discriminate one class from the others. A natural extension is to integrate over all classes:

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

This is actually slightly non-intuitive in that a change in probability of one class will be accompanied by an opposite change in that of other classes. One might feel happier adding squared rates of change to capture both large positive and negative gradients, but in practice this makes little difference.

If it is assumed that the underlying model for the word generation is a multinomial (dice throwing) distribution, the probability of a sequence of words $\boldsymbol{x}$ conditioned on the class, in a maximum likelihood sense, is

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

where there are n_{ik} words of type w_k and D_i words in total in class m_i of the training set. If $U(w_k)$ is defined to be the usefulness of word w_k, then this results in a usefulness for word w_k of

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

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 [6], 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, we feel justified in retaining the name usefulness. Curiously though, the term corresponding to class 1 is weighted by the probability of class 2 and vice-versa.

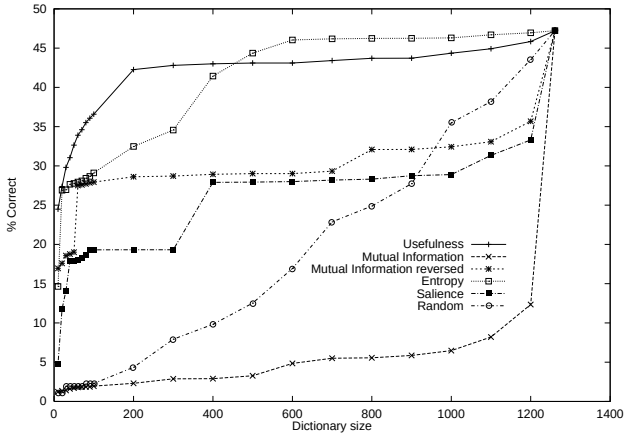

Figure 1: Map task corpus, multinomial

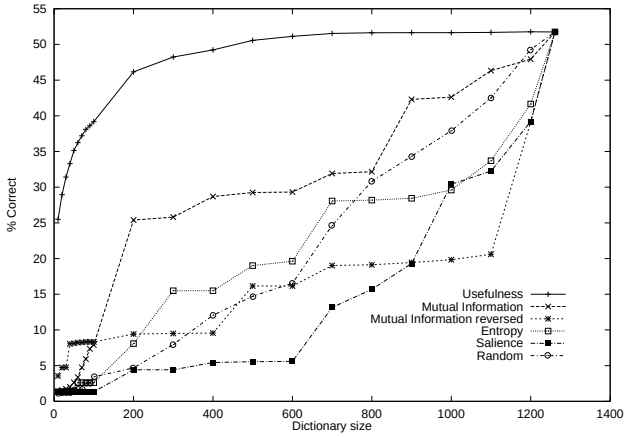

Figure 2: Map task corpus, absolute discounting

If it is assumed that the underlying model of word generation is Poisson, then from [5], the probability of the sentence is

$$P(\boldsymbol{x}|m_i) = \prod_{k=1}^{W} \left[\frac{\Gamma(n_{ik}+\alpha+x_k)}{\Gamma(n_{ik}+\alpha)} \frac{(D_i+\beta)^{n_{ik}+\alpha}}{(D_i+\beta+K)^{n_{ik}+\alpha+x_k}} \right],$$

where there are W distinct words in the vocabulary, and α and β are priors. By the same method as above, this results in a usefulness for word w_k of

$$\begin{aligned} U(w_k) \quad = \quad & \sum_{i=1}^{M} P(x_k|m_i) \sum_{\substack{j=1 \\ j\neq i}}^{M} n_j \big[\log(D_i+\beta) \\ & - \log(D_j+\beta) + \psi(n_{jk}+\alpha) - \psi(n_{ik}+\alpha) \big], \end{aligned}$$

where $P(x_k|m_i)$ is the the probability of a sentence consisting of the single word w_k, and ψ is the digamma function.

4. EXPERIMENTS

Two corpora were used: The HCRC Map Task Corpus [7], which is annotated at the dialogue move level, and the LOB corpus, which is divided into reports and essays classified into different topics. Each corpus was stripped of punctuation and annotation, and translated entirely to lower case. The map task corpus was split into training and testing sets of 64 dialogues each such that no map occured in both sets; this was to bias the discrimination against particular map features. There were 11799 moves in the training set and 10265 in the testing set. The LOB corpus was split by alternating reports into the training and testing sets; the training and testing sets both consisted of 250 reports.

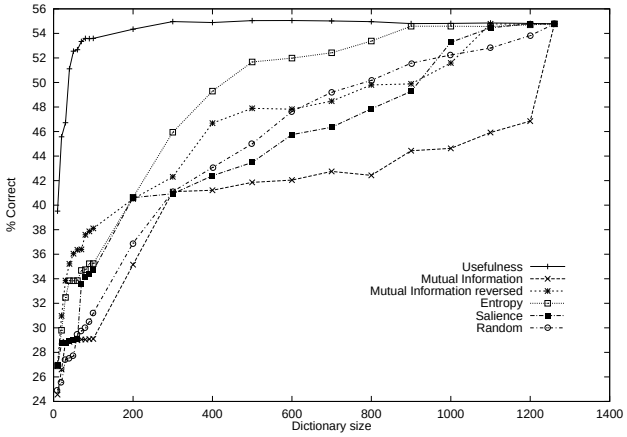

Figure 3: Map task corpus, poisson based

Classification experiments were performed using language models built from both Poisson based and multinomial based probability measures, and classification rate was plotted against dictionary size for various keyword selection methods. Each probability measure was also tested against three randomly ordered dictionaries, the results of which were averaged to provide a baseline.

For the multinomial, out of vocabulary (OOV) words were handled in two different ways. The first, after Nowell [2], involved simply scoring OOV words as if they had occured 0.5 times. The second was to use absolute discounting (for example [10]) to provide a smoothed estimate of word probabilities; this was only optimised for the largest dictionary size. In the Poisson based case, the hyperparameters α and β were set to 0.1 and 0 respectively after [5]. The experimental results are shown in figures 1-5 (Note the different ordinate scales), except those for the basic multinomial on the LOB corpus, which scored consistently below 14%, and were omitted after space considerations.

5. DISCUSSION

The Poisson based probability measure was developed specifically for this type of problem, indeed specifically to alleviate the OOV problems of the multinomial. It is gratifying, therefore, that the Poisson measure performs a good 5% better than the multinomial on the map task, and even better on the LOB corpus. In turn, the multi-class usefulness measure was developed specifically to complement the Poisson based probability, and performs consistently better than any other dictionary pruning method for the Poisson.

The comparitive results are still informative though. In

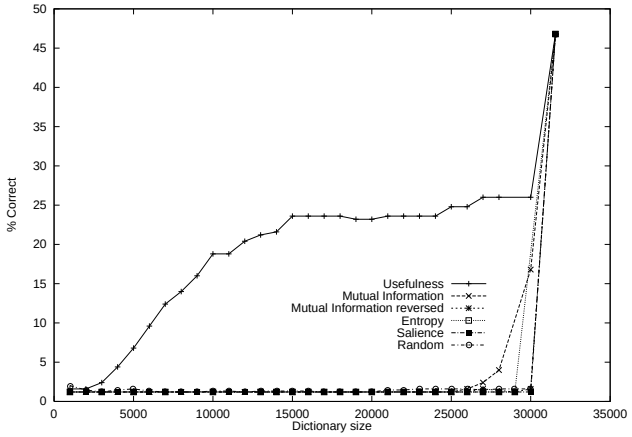

Figure 4: LOB corpus, absolute discounting

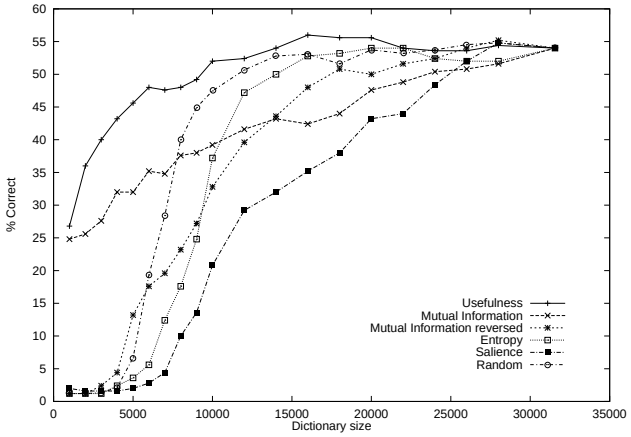

Figure 5: LOB corpus, poisson based

the case of the multinomial, the new usefulness measure bears a striking similarity to the information based measures, no doubt connected with the original derivation of information theory. This resemblance is reflected in the experimental performance: the entropy measure performs better than usefulness for large numbers of keywords.

The behaviour of mutual information is erratic. In particular, the words 'yes' and 'no' corresponding to positive and negative replies in the map task appear as the most useful when ranked by usefulness, but least useful when ranked by mutual information, which produces a word list that is intuitively 'upside down'. The graphs show the effect of simply reversing this list, though with a dubious improvement. In fact, there is no theoretical reason to invert the list. The problems with mutual information are presumably what prompted the invention of salience. Salience, however, still appears from these experiments to perform erratically; sometimes even worse than random. These experiments suggest that entropy would be a better information theoretic measure.

6. CONCLUSIONS

The best results in this study have been obtained with the combination of Poisson based probability estimates for words, and the new multi-class usefulness measure. In this case, performance has been shown to improve when the dictionary size is reduced.

It is not clear that there is any theoretically justifiable reason to choose any particular information theoretic measure over another, although experimentally, entropy has been shown to choose good keywords consistently. It is better to derive a measure specifically to maximise discriminability, and in the case of the multinomial, this derivation yields an expression very similar to information theoretic ones.

7. REFERENCES

[1] Michael J. Carey and Eluned S. Parris. Topic spotting using task independent models. In *Proceedings Eurospeech 95, Madrid*, pages 2133–2137, 1995.

[2] Peter Nowell and Roger K. Moore. The application of dynamic programming techniques to non-word based topic spotting. In *Proceedings Eurospeech '95*, volume 2, pages 1355–1358, Madrid, Spain, September 1995.

[3] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek. Approaches to topic identification on the switchboard corpus. In *Proceedings ICASSP'94*, volume 1, pages 385–388. IEEE, April 1994.

[4] Stuart Bird, Sue R. Browning, Roger K. Moore, and Martin J. Russell. Dialogue move recognition using topic spotting techniques. In *Proceedings ESCA Workshop on Spoken Dialogue Systems*, pages 45–48, May 1995.

[5] Philip N. Garner, Sue R. Browning, Roger K. Moore, and Martin J. Russell. A theory of word frequencies and its application to dialogue move recognition. In *Proceedings ICSLP96*, pages 1880–1883, October 1996.

[6] Eluned S. Parris and Michael J. Carey. Discriminative phonemes for speaker identification. In *Proceedings ICSLP 94*, volume 4, pages 1843–1846, Yokohama, Japan, September 1994.

[7] 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.

[8] Allen L. Gorin. On automated language acquisition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.

[9] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley and sons, Inc., 1968.

[10] Hermann Ney, Ute Essen, and Reinhard Kneser. On the estimation of 'small' probabilities by leaving-one-out. *IEEE Transactions on pattern analysis and machine intelligence*, 17(12):1202–1212, December 1995.