CLUSTERING WORD CATEGORY BASED ON BINOMIAL POSTERIORI CO-OCCURRENCE DISTRIBUTION.

Masafumi Tamoto and Takeshi Kawabata

NTT Basic Research Laboratories
3-1, Morinosato Wakamiya, Atsugi-city
Kanagawa 243-01
Japan
e-mail: tamoto@av-sun2.ntt.jp

ABSTRACT

This paper describes a word clustering technique for stochastic language modeling and reports experimental evidence for its validity. The Binomial Posteriori Distribution (BPD) distance measurement between words is introduced. It is based on word co-occurrency and reliability. We plan to consider a practical application of this clustering technology by utilizing each cluster as a Markov state in the construction of a word prediction model.

1. INTRODUCTION

Stochastic language models are promising for dealing with practical language phenomena. However, for large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus. Many events are rarely or never seen. This makes their frequency counts unreliable estimates of their probabilities.To improve the reliability of model parameter estimation, word grouping is more effective.

In this paper, we define a distance measurement between words in terms of the co-occurrence probability and its estimation reliability. This measurement is based on the binomial posteriori distribution (BPD) of the occurrence probability. After that, this word distance is applied to the LBG-based clustering procedure. Several aspects on word clustering in a corpus are reported. Taking the estimation reliability into account, we have found that adequate and robust word clustering can be accomplished by means of estimation reliability.

1.1. Problem Setting

Consider a word set $\mathcal{N}$, containing nouns, pronouns, proper nouns and numerals; and a co-occurrence distribution that consists of the frequencies of what words follow immediately after these nouns in the training corpus. To cluster $\mathcal{N}$ according to co-occurrence distributions, we need a measurement of similarity between distributions. In many cases, mutual information (MI) between two distributions was used. However, this suffered from unreliable estimation due to lack of sufficient data. Our research addresses this problem and uses a distance measurement between words, which takes into account co-occurrence probability and estimation reliability.

2. BINOMIAL POSTERIORI DISTRIBUTION

A distance measurement between words based on their co-occurrency is defined in this section. Suppose the word i appears n times in the training corpus and co-occurs k times with the word l. Co-Occurrence probability of word i with word l is defined as the conditional probability $P(l|i)$. Let p be the true value of this conditional probability. The probability that the word appearing n times co-occurs k times with the other word is as follows (binomial distribution).

$$P(k|n,p) = {}_nC_k \cdot p^k(1-p)^{n-k} \tag{1}$$

Where ${}_nC_k$ is the number of combinations of n things taken k by k. When n and k are observed, the posteriori probability density of the true probability p is calculated by Bayes' theorem.

$$P(p|n,k) = \frac{{}_nC_k \cdot p^k(1-p)^{n-k}}{\int_0^1 {}_nC_k \cdot q^k(1-q)^{n-k}dq} \tag{2}$$

By reducing the fraction, we get

$$B_n^k(p) \equiv p^k(1-p)^{n-k} \Big/ \int_0^1 q^k(1-q)^{n-k}dq. \tag{3}$$

0-7803-2431-5/95 $4.00

This distribution is called the Binomial Posteriori Distribution. The mean value of the BPD is calculated as

$$\mu = \int_0^1 p \cdot B_n^k(p)dp = \frac{\int_0^1 (1-p)^{n-k}p^{k+1}dp}{\int_0^1 (1-p)^{n-k}p^k dp}. \quad (4)$$

Integrating the numerator by parts

$$\begin{aligned}\int_0^1 (1-p)^{n-k}p^{k+1}dp \\ = \frac{1}{n+2}\left(\left[(1-p)^{n-k+1}p^{k+1}\right]_0^1 + \right. \\ \left.(k+1)\int_0^1 (1-p)^{n-k}p^k dp\right), \quad (5)\end{aligned}$$

where $\left[(1-p)^{n-k+1}p^{k+1}\right]_0^1 = 0$. Consequently,

$$\mu = (k+1)/(n+2). \quad (6)$$

The BPD variance is calculated by

$$\begin{aligned}\sigma^2 &= \int_0^1 (p-\mu)^2 \cdot B_n^k(p)dp \\ &= \int_0^1 (p^2 - 2p\mu + \mu^2) \cdot B_n^k(p)dp \\ &= \frac{k+2}{n+3}\cdot\frac{k+1}{n+2} - \left(\frac{k+1}{n+2}\right)^2 + \left(\frac{k+1}{n+2}\right)^2 \\ &= (n-k+1)(k+1)/(n+3)(n+2)^2. \quad (7)\end{aligned}$$

Figure 1 shows some examples of BPD for simple n and k variations. Figure 1(a) is the BPD when the word i appears once in the corpus and does not occur with the other word j. The expected conditional probability is given as the maximum value of this distribution (*i.e.* $p = 0$). However, this value is not reliable because the distribution is broad. Figure 1(b) is the BPD when the word i appears twice and occurs once with the other word j. The expected conditional probability is 0.5. This value is also unreliable. Figure 1(c) is the distribution when the word appears 100 times and co-occurs 50 times. In this case, the expected conditional probability is 0.5, which is reliable. Figure 1(b) and (c) have the same expected value.

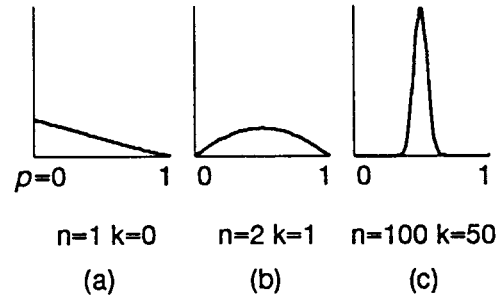

Figure 1. Examples of Binomial Posteriori Distributions

Reliability estimates very depending on the number of samples in the corpus. The reliability of conditional probability estimation is reflected in the BPD. We define the distance between the words i and j concerning co-occurrency with the word l as

$$d_{ij}^{(l)} = \iint_0^1 (x-y)^2 B_{n_i}^{k_i^{(l)}}(x) B_{n_j}^{k_j^{(l)}}(y)\,dxdy. \quad (8)$$

This equation gives the mean square distance between points x and y which are randomly and independently sampled from two BPDs. Equation 8 can be

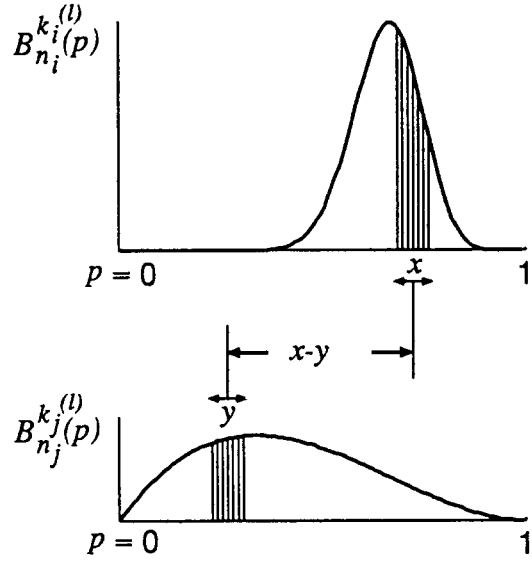

Figure 2. Stochastic distance between two distributions

simplified by the independency of sample variables x and y.

$$d_{ij}^{(l)} = \int_0^1 x^2 B_{n_i}^{k_i^{(l)}}(x)\,dx + \int_0^1 y^2 B_{n_j}^{k_j^{(l)}}(y)\,dy$$

$$-2\iint_0^1 xy B_{n_i}^{k_i^{(l)}}(x) B_{n_j}^{k_j^{(l)}}(y)\,dxdy$$
$$= (\mu_i - \mu_j)^2 + \sigma_i^2 + \sigma_j^2. \quad (9)$$

From quation 6, 7 and 9,

$$d_{ij}^{(l)} = \left(\frac{k_i^{(l)}+1}{n_i+2} - \frac{k_j^{(l)}+1}{n_j+2}\right)^2 + \frac{(n_i-k_i^{(l)}+1)(k_i^{(l)}+1)}{(n_i+3)(n_i+2)^2} + \frac{(n_j-k_j^{(l)}+1)(k_j^{(l)}+1)}{(n_j+3)(n_j+2)^2}. \quad (10)$$

Consequently, the total distance between words i and j in terms of co-occurrency with other words is defined as

$$d_{ij} = \sum_l d_{ij}^{(l)}, \quad (11)$$

where $\sum_l$ is the summation of all words.

3. CLUSTERING EXAMPLES

We used this method to classify 1,340 nouns with 10,022 co-occurrences appearing in the corpus. In this corpus, the chosen nouns appear as heads of a total of 213 distinct suffixes, so each noun is represented by a density over the 213 words.

3.1. Text Corpus

The evaluation described below was performed on a data set having 31,700 partial sentences from the ATR Dialogue Database (ADD). This correction process yielded 10,022 noun-suffix pairs. Table 1 shows the contents of the dataset.

Table 1. Noun - Suffix pairs in ADD Corpus

	words	frequency
noun	598	3254
pronoun	21	671
numeral	416	5266
proper noun	305	831
suffix	213	10022

Table 2. Example of numeral co-occurrence

word	suffix (frequency)			amount
1	日 (135)	泊 (112)	時間 (61)	(774)
一 ('1')	人 (167)	つ (154)	度 (139)	(674)
2	泊 (102)	日 (73)	人 (48)	(516)
二 ('2')	人 (58)	つ (22)	日 (21)	(154)
3	日 (141)	泊 (60)	人 (29)	(390)
三 ('3')	日 (19)	つ (12)		(64)
4	日 (93)	泊 (36)	人 (26)	(252)
四 ('4')	日 (13)	時 (6)		(42)
5	日 (59)	名 (36)	時 (34)	(205)
6	時 (34)	日 (33)	人 (10)	(114)
五 ('5')	日 (5)	分 (4)	時 (3)	(17)
7	時 (35)	日 (35)	名 (4)	(79)
8	日 (38)	時 (27)	名 (3)	(79)
9	時 (34)	日 (11)	名 (4)	(52)
1 1	時 (18)	日 (8)	人 (4)	(32)
1 2	時 (36)	日 (4)		(45)
1 3	時 (8)	日 (7)		(17)
1 0	日 (59)	時 (52)	名 (38)	(208)

Table 2 shows example noun clusters and co-occurrence distribution of frequent suffixes. These four clusters associate numerals resulting from 248 clusters.

Consider two suffixes "日" (meaning a "day"), "時" ("hour"), there are differences about co-occurrence distribution and frequency among four clusters. In the first cluster, words are followed by "日" but "時", and relatively high frequency. Arabic numeral from 1 to 6 and correspoding Chinese character ("一","二","三","四" meaning "one","two","three","four") are associated. Words in the other cluster allow both "日" and "時". "五"("five") of the second cluster is relatively high variance due to infrequent occurrence. Contrary, "10" is less variance. These results lead that the BPD distance measurement classified nouns according to their distribution and validity of estimation.

4. MODEL EVALUATION

The preceding discussion provided some indication of what aspects of distributional relationships may be discovered by clustering. We need to evaluate clustering as a basis for models of distributional relationships. So far, we have looked at two kinds of measurements of model quality: (i) mutual information of resulting clusters, and (ii) perplexity on the task of evaluation of test set data.

The evaluation described below was performed on the whole data set we have worked with so far, extracted from 10,022 ADD co-occurrences of ADD. We

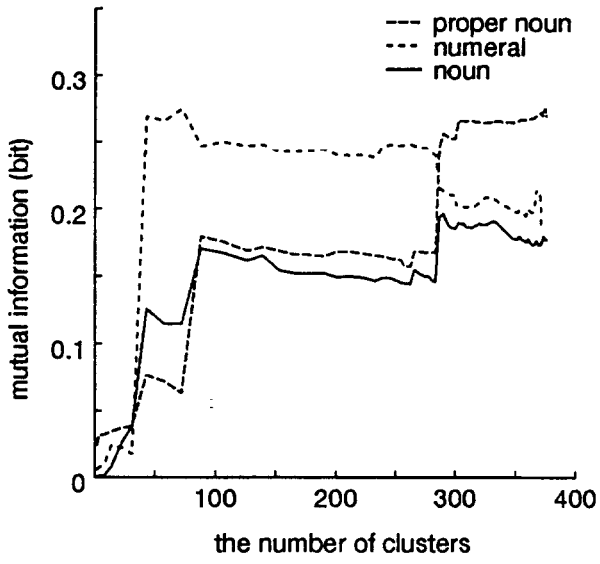

Figure 3. Mutual Information

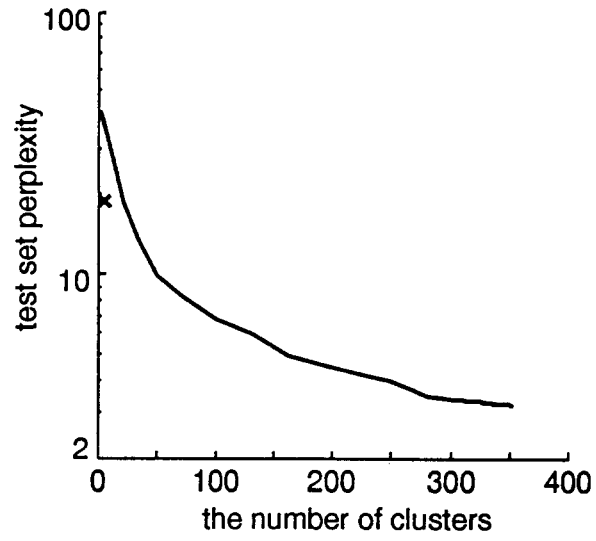

Figure 4. Word Perplexity

randomly divided it into a training set of 9,020 co-occurrences and a test set of 1,000.

4.1. Mutual Information

Figure 3 plots the mutual information, in bits, between clusters, given by $I = -\sum_i \sum_j P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$ where i is a part of speech and j is a cluster. Mutual information between grammatical category and word clusters according to their co-occurrences indicates the ability of organizing syntactic category by word clustering. The mutual information increases as the clusters are divided and saturates at 248 clusters.

4.2. Word Perplexity

We also evaluated our method on word perplexity. Figure 4 plots the word perplexity. Perplexity is given by

$$H = -\sum_{i,k} P(k,i) \log P(k|i),$$

where $P(k|i) = \sum_j P(k|j) \cdot P(j|i)$, i is a suffix, j is a word class and k is a noun.

Perplexity decreases monotonically according to the number of clusters. Perplexity based on the grammatical 4 categories for nouns, pronouns, proper nouns and numerals, is 19.4 (marked '×'). Perplexity was reduced to 3.96 using the 248 generated clusters.

5. CONCLUSIONS

We have demonstrated that a divisive clustering procedure for binomial posteriori distributions can be used to group words according to their participation in co-occurrency with other words. The results are intuitively informative and can be used to construct class-based word co-occurrence models with substantial predictive power. Using word bigram language models, this word grouping improved word perplexity by 20% on a spoken dialog corpus containing 31,700 partial sentences.

6. REFERENCES

[1] D Hindle. Noun classification from predicate argument structures. In *ACL 90*, pp. 168–275, 1990.

[2] Eric Brill, David Magerman, Mitchell Marcus, and Beatrice Santorini. Deducing linguistic structure from the statistics of language corpora. In *Proceedings of Speech and Natural Language Workshop*, pp. 275–281. DARPA, 1990.

[3] Kenneth Ward Church. Word association norms, mutual information and lexicography. *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29, 1 1990.

[4] Ute Essen and Volker Steinbess. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of ICASSP Speech Processing*, pp. 161–164. IEEE, 1992.

[5] Fernando Pereira. Distributional clustering of english words. In *ACL 93*, pp. 183–190, 1993.