Collocation Map for Overcoming Data Sparseness

Moonjoo Kim, Young S. Han, and Key-Sun Choi
Department of Computer Science
Korea Advanced Institute of Science and Technology
Taejon, 305-701, Korea
mj0712@eve.kaist.ac.kr, yshan@csking.kaist.ac.kr, kschoi@csking.kaist.ac.kr

Abstract

Statistical language models are useful because they can provide probabilistic information upon uncertain decision making. The most common statistic is n-grams measuring word cooccurrences in texts. The method suffers from data shortage problem, however. In this paper, we suggest Bayesian networks be used in approximating the statistics of insufficient occurrences and of those that do not occur in the sample texts with graceful degradation. Collocation map is a sigmoid belief network that can be constructed from bigrams. We compared the conditional probabilities and mutual information computed from bigrams and Collocation map. The results show that the variance of the values from Collocation map is smaller than that from frequency measure for the infrequent pairs by 48%. The predictive power of Collocation map for arbitrary associations not observed from sample texts is also demonstrated.

1 Introduction

In statistical language processing, n-grams are basic to many probabilistic models including Hidden Markov models that work on the limited dependency of linguistic events. In this regard, Bayesian models (Bayesian network, Belief network, Inference diagram to name a few) are not very different from HMMs. Bayesian models capture the conditional independence among probabilistic variables, and can compute the conditional distribution of the variables, which is known as a probabilistic inferencing. The pure n-gram statistic, however, is somewhat crude in that it cannot do anything about unobserved events and its approximation on infrequent events can be unreliable.

In this paper we show by way of extensive experiments that the Bayesian method that also can be composed from bigrams can overcome the data sparseness problem that is inherent in frequency counting methods. According to the empirical results, Collocation map that is a Bayesian model for lexical variables induced graceful approximation over unobserved and infrequent events.

There are two known methods to deal with the data sparseness problem. They are smoothing and class based methods (Dagan 1992). Smoothing methods (Church and Gale 1991) readjust the distribution of frequencies of word occurrences obtained from sample texts, and verify the distribution through held-out texts. As Dagan (1992) pointed out, however, the values from the smoothing methods closely agree with the probability of a bigram consisting of two independent words.

Class based methods (Pereira et al. 1993) approximate the likelihood of unobserved words based on similar words. Dagan and et al. (1992) proposed a non-hierarchical class based method. The two approaches report limited successes of purely experimental nature. This is so because they are based on strong assumptions. In the case of smoothing methods, frequency readjustment is somewhat arbitrary and will not be good for heavily dependent bigrams. As to the class based methods, the notion of *similar* words differs across different methods, and the association of probabilistic dependency with the similarity (class) of words is too strong to assume in general.

Collocation map that is first suggested in (Han 1993) is a sigmoid belief network with words as probabilistic variables. Sigmoid belief network is extensively studied by Neal (1992), and has an efficient inferencing algorithm. Unlike other Bayesian models, the inferencing on sigmoid belief network is not NP-hard, and inference methods by reducing the network and sampling are discussed in (Han 1995). Bayesian models constructed from local dependencies provide formal approximation among the variables, thus using Collocation map does not require strong assumption or intuition to justify the associations among words produced by the map.

The results of inferencing on Collocation map are probabilities among any combinations of words represented in the map, which is not found

in other models. One significant shortcoming of Bayesian models lies in the heavy cost of inferencing. Our implementation of Collocation map includes 988 nodes, and takes 2 to 3 minutes to compute an association between words. The purpose of experiments is to find out how gracefully Collocation map deals with the unobserved cooccurrences in comparison with a naive bigram statistic.

In the next section, Collocation map is reviewed following the definition in (Han 1993). In section 3, mutual information and conditional probabilities computed using bigrams and Collocation map are compared. Section 4 concludes the paper by summarizing the good and bad points of the Collocation map and other methods.

2 Collocation Map

In this section, we make a brief introduction on Collocation map, and refer to (Han 1993) for more discussion on the definition and to (Han 1995) on inference methods.

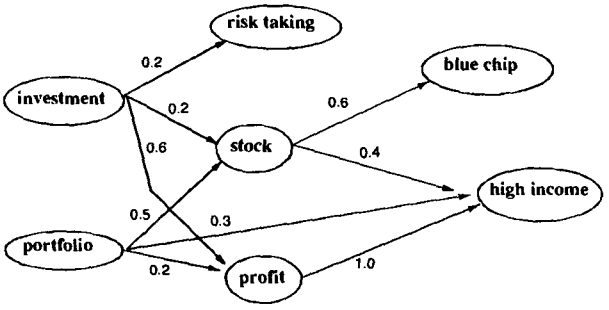
Bayesian model consists of a network and probability tables defined on the nodes of the network. The nodes in the network represent probabilistic variables of a problem domain. The network can compute probabilistic dependency between any combination of the variables. The model is well documented as subjective probability theory (Pearl 1988).

Collocation map is an application model of sigmoid belief network (Neal 1992) that belongs to belief networks which in turn is a type of Bayesian model. Unlike belief networks, Collocation map does not have deterministic variables thus consists only of probabilistic variables that correspond to words in this case.

Sigmoid belief network is different from other belief networks in that it does not have probability distribution table at each node but weights on the edges between the nodes. A node takes binary outcomes (1, -1) and the probability that a node takes an outcome given the vector of outcomes of its preceding nodes is a sigmoid function of the outcomes and the weights of associated edges. In this regard, the sigmoid belief network resembles artificial neural network. Such probabilities used to be stored at nodes in ordinary Bayesian models, and this makes the inferencing very difficult because the probability table can be very big. Sigmoid belief network does away with the NP-hard complexity by avoiding the tables at the loss of expressive generality of probability distributions that can be encoded in the tables.

One who works with Collocation map has to deal with two problems. The first is how to construct the network, and the other is how to compute the probabilities on the network.

Network can be constructed directly from a set of bigrams obtained from a training sample. Because Collocation map is a directed acyclic graph, cycles are avoided by making additional node of a word when facing cycle due to the node. No more than two nodes for each word are needed to avoid the cycle in any case (Han 1993). Once the network is setup, edges of the network are assigned with weights that are normalized frequency of the edges at a node.



P(profit | investment) = 0.644069
P(risk-taking | investment) = 0.549834
P(stock | investment) = 0.546001
P(high-income | investment) = 0.564798
P(investment | high-income) = 0.500000
P(high-income | risk-taking profit) = 0.720300
P(investment | portfolio high-income risk-taking) = 0.495988
P(portfolio | blue-chip) = 0.500000
P(portfolio stock | portfolio stock) = 1.000000

Figure 1: Example Collocation map and example inferences. Graph reduction method (Han 1995) is used in computing the probabilities.

The inferencing on Collocation map is not different from that for sigmoid belief network. The time complexity of inferencing by reducing graph on sigmoid belief networks is $O(N^3)$ given N nodes (Han 1995). It turned out that inferencing on networks containing more than a few hundred nodes was not practical using either node reduction method or sampling method, thus we adopted the hybrid inferencing method that first reduces the network and applies Gibbs sampling method (Han 1995). Using the hybrid inferencing method, computation of conditional probabilities took less than a second for a network with 50 nodes, two seconds for a network with 100 nodes, about nine seconds for a network with 200 nodes, and about two minutes for a network with about 1000 nodes.

Conditional and marginal probabilities can be approximated from Gibb's sampling. Some conditional probabilities computed from a small network are shown in figure 1. Though the network may not be big enough to model the domain of finance, the resulting values from the small network composed of 9 dependencies seem useful and intuitive.

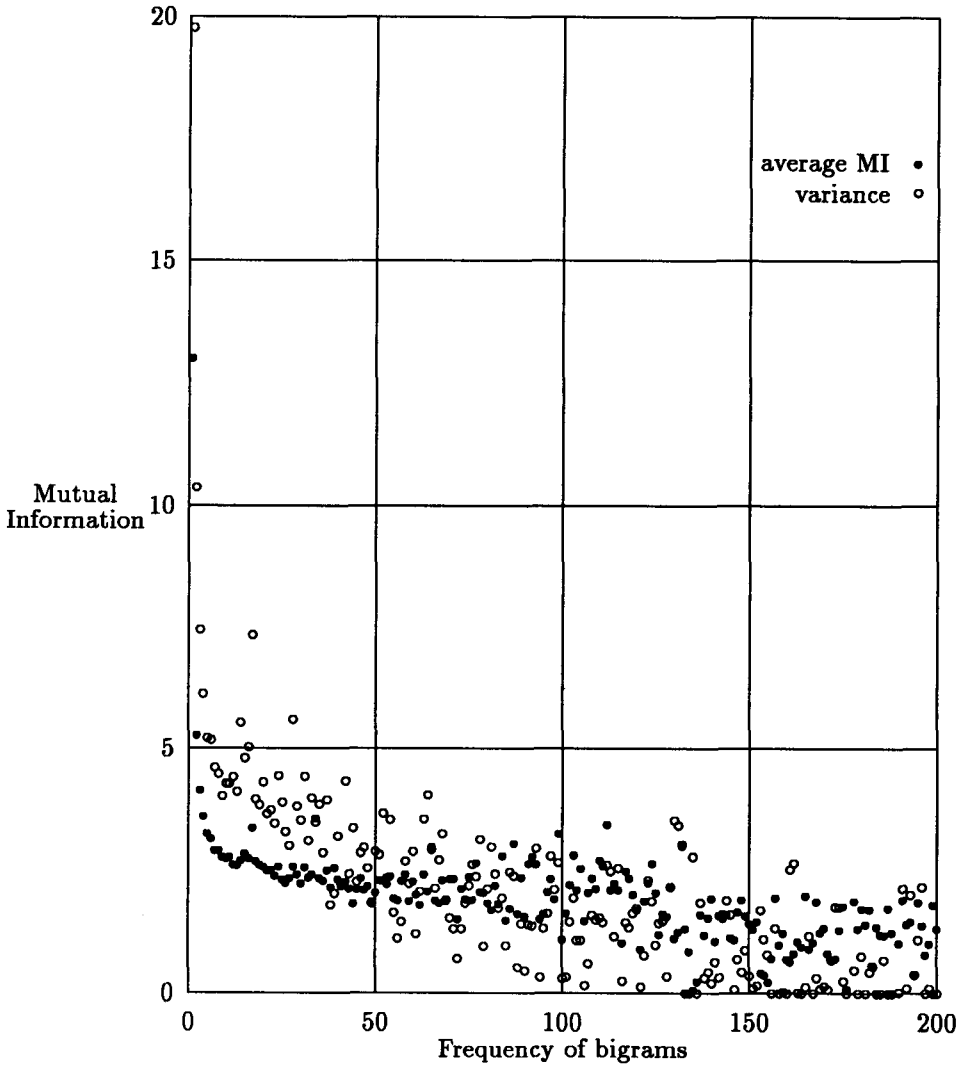


Figure 2: Average MI's and variances. 378,888 unique bigrams are classified according to frequency.

The computation in figure 1 was done by using graph reduction method. As it is shown in the example inferences, the association between any combination of variables can be measured.

3 Experiments

The goal of our experiment is first to find how data sparseness is related with the frequency based statistics and to show Collocation map based method gives more reliable approximations. In particular, from the experiments we observed the variances of statistics might suggest the level of data sparseness. The less frequent data tended to have higher variances though the values of statistics (mutual information for instance) did not distinguish the level of occurrences. The predictive account of Collocation map is demonstrated by observing the variances of approximations on the infrequent events.

The tagged Wall Street Journal articles of Penn Tree corpus were used that contain about 2.6 million word units. In the experiments, about 1.2 million of them was used. Programs were coded in C language, and run on a Sun Sparc 10 workstation.

For the first 1.2 million words, the bigrams consisting of four types of categories (*NN, NNS, IN, JJ*) were obtained, and mutual information of each bigram (order insensitive) was computed. The bigrams were classified into 200 sets according to their occurrences. Figure 2 summarizes the the average MI value and the variance of each frequency range. From figure 3 that shows the occurrence distribution of 378,888 unique bigrams, about 70% of them occur only one time. One interesting and important observation is that those of 1 to 3 frequency range that take about 90% of the population have very high MI values. This results also agree with Dunning's argument about overestimation on the infrequent occurrences in which many infrequent pairs tend to get higher estimation (Dunning 1993). The problem is due to the assumption of normality in naive frequency based statistics according to Dunning (1993). Approximated values, thus, do not indicate the level of data quality. Figure 3 shows variances can suggest the level of data sufficiency. From this observation we propose the following definition on the notion of data sparseness.

> A set of units belonging to a sample of ordered word units (texts) is α *data-sparse*
> if and only if the variance of measurements on the set is greater than α.

The definition sets the concept of sparseness within the context of a focused set of linguistic units. For a set of units unoberved from a sample, the given sample text is for sure *data-sparse*. The above definition then gives a way to judge with respect to observed units. The measurement of data sparseness can be a good issue to study where it may depend on the contexts of research. Here we suggest a simple method perhaps for the first time in the literature.

Figure 4 compares the results from using Collocation map and simple frequency statistic. The variances are smaller and the pairs in frequency 1 class have non zero approximations. Because computation on Collocation map is very high, we have chosen 2000 unique pairs at random. The network consists of 988 nodes. Computing an approximation (inferencing) took about 3 minutes. The test size of 2000 pairs may not be sufficient, but it showed the consistent tendency of graceful degradation of variances. The overestimation problem was not significant in the approximations by Collocation map. The average value of zero frequency class to which 50 unobserved pairs belong was also on the line of smooth degradation, and figure 4 shows only the variance.

Table 1 summarizes the details of performance gain by using Collocation map.

4 Conclusion

Corpus based natural language processing has been one of the central subjects gaining rapid attention from the research community. The major virtue of statistical approaches is in evaluating linguistic events and determining the relative importance of the events to resolve ambiguities. The evaluation on the events (mostly cooccurrences) in many cases, however, has been unreliable because of the lack of data.

Data sparseness addresses the shortage of data in estimating probabilistic parameters. As a result, there are too many events unobserved, and even if events have been found, the occurrence is not sufficient enough for the estimation to be reliable.

In contrast with existing methods that are based on strong assumptions, the method using Collocation map promises a logical approximation since it is built on a thorough formal argument of Bayesian probability theory. The powerful feature of the framework is the ability to make use of the conditional independence among word units and to make associations about unseen cooccurrences based on observed ones. This naturally induces the attributes required to deal with data sparseness. Our experiments confirm that Collocation map makes predictive approximation and avoids overestimation of infrequent occurrences.

One critical drawback of Collocation map is the time complexity, but it can be useful for applications of limited scope.

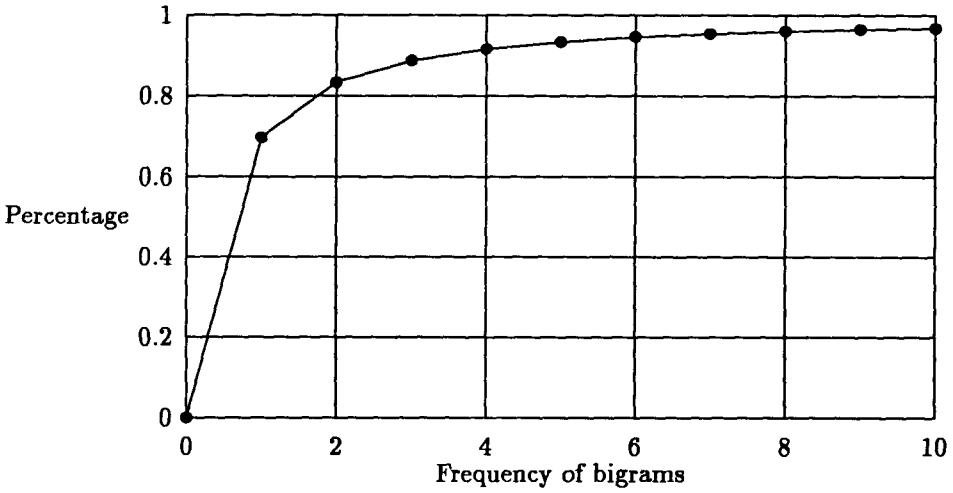


Figure 3: The distribution of 378,888 unique bigrams. First ten classes are shown.

Freq	Collocation map	Frequency Based	Improvement
1	5.1	12.2	57%
10	2.28	4.28	46%
20	1.29	5.29	75%
30	1.51	3.51	56%
40	2.18	3.18	31%
50	1.52	2.87	47%
average	2.04	4.5	45%

Table 1: Comparison of variances between frequency based and Collocation map based MI computations.

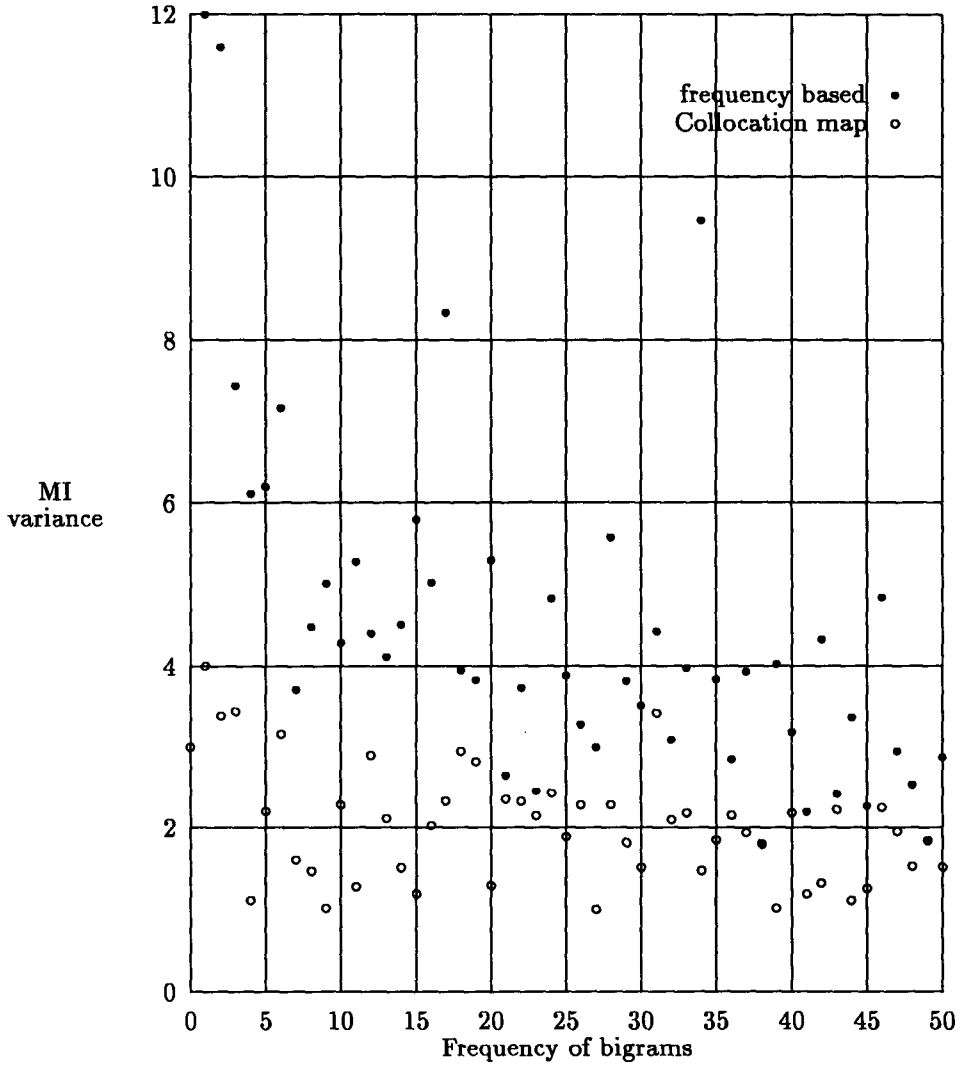


Figure 4: Variances by frequency based and Collocation map based MI computations for 2000 unique bigrams.

References

Kenneth W. Church, and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language.* **5**. 19-54.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics.* **19** (1). 61-74.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1992. Contextual word similarity and estimation from sparse data. In *Proceedings of AAAI fall symposium*, Cambridge, MI. 164-171.

Young S. Han, Young G. Han, and Key-sun Choi. 1992. Recursive Markov chain as a stochastic grammar. In *Proceedings of a SIGLEX workshop*, Columbus, Ohio. 22-31.

Young S. Han, Young C. Park, and Key-sun Choi. 1995. Efficient inferencing for sigmoid Bayesian networks. to appear in *Applied Intelligence.*

Radford M. Neal. 1992. Connectionist learning of belief networks. *J of Artificial Intelligence.* **56**. 71-113.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the Annual Meeting of the ACL.*