Language Modeling using Dynamic Bayesian Networks

Murat Deviren, Khalid Daoudi and Kamel Smaïli

INRIA-LORIA, Parole team
B.P. 101 - 54602 Villers les Nancy, France
Tel.: +33 (0)3 83 59 20 22 - Fax: +33 (0)3 83 59 19 27
e-mail: daoudi,deviren,smaili@loria.fr - http://www.loria.fr/equipes/parole

Abstract

In this paper we propose a new approach to language modeling based on dynamic Bayesian networks. The principle idea of our approach is to find the dependence relations between variables that represent different linguistic units (word, class, concept, ...) that constitutes a language model. In the context of this paper the linguistic units that we consider are syntactic classes and words. Our approach should not be considered as a model combination technique. Rather, it is an original and coherent methodology that processes words and classes in the same model. We attempt to identify and model the dependence of words and classes on their linguistic context. Our ultimate goal is to devise an automatic mechanism that extracts the best dependence relations between a word and its context, i.e., lexical and syntactic. Preliminary results are very encouraging, in particular the model in which a word depends not only on previous word but also on syntactic classes of two previous words. This model outperforms the bi-gram model.

1. Introduction

The role of a statistical language model is to model different linguistic events in a language. This is achieved by assigning a probability to each word sequence hw_t (where h represent the history of word w_t). These probabilities are estimated using a Markov chain of some fixed order (usually 1 or 2). n-grams formulated as Markov chains are the most widely used language models. Based on the observation that certain words have similar behavior, one can think of clustering words into classes and construct n-class models to substitute or complement n-grams. This approach is used to reduce model complexity, improve generalization property, handle missing data and introduce high level linguistic information (syntactic, semantic etc.). Different complementary linguistic information sources can be exploited using simple linear ombination (Jelinek and Mercer, 1980) or maximum entropy (Rosenfeld, 1994) techniques.

As it will become clear in the following sections, n-gram and n-class models are actually particular cases of dynamic Bayesian networks (DBN). In this paper we propose to use the DBN formalism in order to achieve a better exploitation of each linguistic unit considered in modeling. We develop a unifying approach that processes each of these units in one model and construct new language models with improved performance. The principle of our approach is to construct DBNs in which a variable (word, class or any other linguistic unit) may depend on a set of context variables. These dependence links between linguistic units can be determined automatically or manually. Our ultimate goal is to propose an automatic scheme to learn the optimal DBN structure from a training corpus. With this goal in mind, in this paper we investigate the feasibility of our approach by first testing models for which the structure is specified manually. The advantage of our approach with respect to linear interpolation is that, instead of using a weighted average of models constructed over words and classes separately, our approach integrates all linguistic units in one model without any distinction between classes and words. Compared to the maximum entropy principle, we can say that our approach yields models that are much easier to interpret.

2. Dynamic Bayesian networks

Our approach is based on the framework of *dynamic Bayesian networks* (DBNs) which is a generalization of Bayesian networks (BNs) to dynamic processes. Briefly, the Bayesian networks formalism consists of associating a directed acyclic graph to the joint probability distribution (JPD) $P(X)$ of a set of random variables $X = \{X_1, ..., X_n\}$. The nodes of this graph represent the random variables, while the arrows encode the conditional independences (CI) which (are supposed to) exist in the JPD. A BN is completely defined by a graph structure S and the numerical parametrization Θ of the conditional probabilities of the variables given their parents. Indeed, the JPD can be expressed in a factored way as $P(X) = \prod_{i=1}^{n} P(X_i|\Pi_i)$, where Π_i denotes the parents of X_i in S.

A DBN encodes the joint probability distribution of a time evolving set $X[t] = \{X_1[t], \ldots, X_n[t]\}$ of variables. If we consider T time slices of variables, the DBN can be considered as a (static) BN with $T \times n$ variables. Using the factorization property of BNs, the joint probability density of $\mathbf{X}_1^T = \{X[1], \ldots, X[T]\}$ is written as:

$$P(X[1], \ldots, X[T]) = \prod_{t=1}^{T} \prod_{i=1}^{n} P(X_i[t]|\Pi_{it}) \qquad (1)$$

where Π_{it} denotes the parents of $X_i[t]$. In the BNs literature, DBNs are defined using the assumption that $X[t]$ is Markovian (Friedman et al., 1998). In this paper, we relax this hypothesis to allow non-Markov processes and we consider that the process $X[t]$ satisfies:

$$P(X_i[t]|\mathbf{X}_1^{t+\tau_f}) = P(X_i[t]|X[t-\tau_p], \ldots, X[t+\tau_f]) \quad (2)$$

for some integers τ_p and τ_f. Graphically, the above assumption states that a variable at time t can have parents in the interval $[t-\tau_p, t+\tau_f]$ (see (Deviren and Daoudi, 2001) for details).

From this perspective, it is obvious that classical language models can be represented as DBNs. Indeed, n-gram

models assume that the probability of a word sequence is factorized over the conditional probabilities of each word in the sequence given its recent history of $n-1$ words. That is, if W is the word vocabulary and $w_1^T = w_1...w_T \in W^T$ is a word sequence, one assumes that:

$$P(w_1^T) = \prod_{t=1}^{T} P(w_t|w_{t-1}, \ldots, w_{t-n+1}) \quad (3)$$

Thus, if W_t is a discrete random variable taking its values in W for every t, n-grams can be represented as the DBN shown in Fig. 1-(a) (for $n = 3$, i.e., trigram) which is a Markov chain of order n. Class-based approaches represent the history on word classes rather than words. That is, if $C = \{l_1, ..., l_m\}$ is the set of class labels and $c_1^T = c_1...c_T \in C^T$ is an observed class sequence, one assumes that:

$$P(w_1^T, c_1^T) = \prod_{t=1}^{T} P(w_t|c_t)P(c_t|c_{t-1}, \ldots, c_{t-n+1}). \quad (4)$$

Thus, if C_t is a discrete random variable taking its values in C for every t, n-class models can be represented as the DBN shown in Fig. 1-(b) (for $n = 2$, i.e., bi-class).

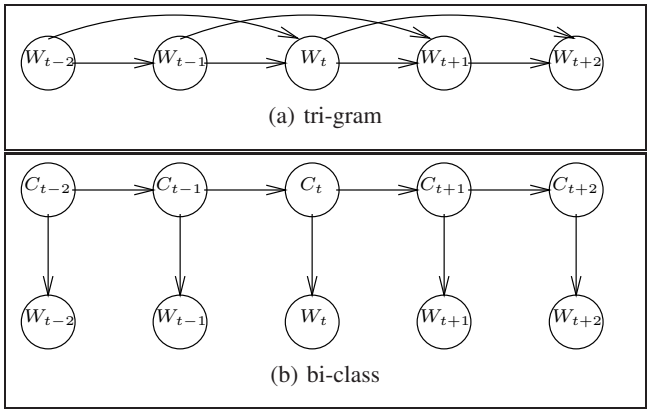

Figure 1: tri-gram and bi-class models

3. DBNs for language modeling

n-gram and n-class models are the most commonly used language models in state-of-the-art speech recognition systems. It has been shown that n-gram models are very efficient when they are trained on a sufficiently informative corpus. The principle inconvenience of these models is that they are computationally very demanding given that they require $|W|^n$ probabilities to be estimated, where n is the model order. Despite the "cut-off" techniques that reduce the complexity, these models remain extremely complex. On the other hand n-class models are less complex since the number of classes is generally much smaller than the vocabulary size. They also have better generalization properties. Nevertheless, they are less accurate and their perplexity is higher. For this reason they are generally combined with n-grams. In such circumstances the realization of a language model is performed first by specifying the model order (n) and then the two models (n-gram and n-class) are trained individually using maximum likelihood criteria. The last step is either a linear combination of these models or an integration of their respective characteristics in a single architecture using maximum entropy techniques. This approach yields quite interesting results, however if we want to better exploit the lexical and syntactic information, a solution would be to consider them in a unique model that is trained within a single procedure.

The DBN formalism provides a theoretical and computational framework to achieve our goal. Our principle idea is to impose no *a priori* hypothesis on the way a language should be represented but to consider all available data (words, classes, ...) as observations of the dynamic system $\{W_t, C_t\}$. Our goal then is to find the model that has the best description (in terms of perplexity) of these observations. In this way, we let data dictate what influences the pronunciation of a word. In Bayesian networks terminology this is the *structure learning* problem: find the graph structure (and its numerical parameterization) that explains the data at ''best". In BN literature there are algorithms that attempt to solve this (difficult) problem (Heckerman, 1995), their performances depend on the application. Our objective in this paper is not to use these algorithms (note that our ultimate goal is to develop a structure learning algorithm that is well adapted to language modeling). Our objective is rather to check whether our approach could effectively overcome classical n-gram and n-class models. Therefore we consider a rich set of graph structures that could be plausible for language modeling. We evaluate several DBNs in this set on a concrete application and compare them to n-gram and n-class models.

3.1. Proposed set of structures

In order to define a set of DBN structures plausible for language modeling, we need to specify CI assertions that are linguistically informative and easy to interpret. We also want n-gram and n-class models to be included in this set in order to be able to exploit their linguistic properties. To do so we start by relaxing the CI assumptions of n-class models. We assume that a word does not only depend on its class but also on the classes of words in a limited past and/or future context. To incorporate the properties of n-grams we authorize the dependence of the word to its lexical history. And finally we consider that a class may not only depend on its history but also on the current word and/or other words in a limited past and/or future context. Figure 2 shows an example of a DBN in the set of structures we consider. In this model each word depends on the previous, present and future class and each class depends on previous class.

The joint probability for a specific model is expressed as:

$$P(w_1^T, c_1^T) = \prod_t P(w_t|\pi_{w_t})P(c_t|\pi_{c_t}) \quad (5)$$

where π_{w_t} (resp. π_{c_t}) is a realization of the parents Π_{W_t} (resp. Π_{C_t}) of W_t (resp. C_t). The numerical parameterization of the model is given by conditional probability tables (independent of time t) $P(w_t|\pi_{w_t})$ and $P(c_t|\pi_{c_t})$ that we

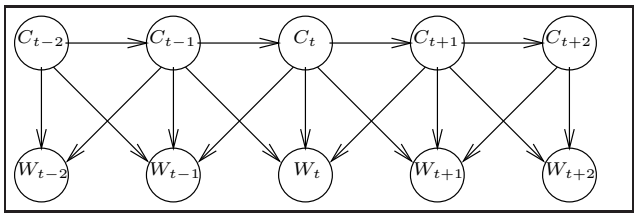

Figure 2: An example DBN structure.

note, for $X_t \in \{W_t, C_t\}$, by:

$$\theta_{x,j,\mathbf{k}} = P(X_t = j | \Pi_{X_t} = \mathbf{k}). \quad (6)$$

These parameters are estimated using maximum likelihood criterion that yields :

$$\theta_{x,j,\mathbf{k}} = \frac{N_{x,j,\mathbf{k}}}{\sum_j N_{x,j,\mathbf{k}}} \quad (7)$$

where $N_{x,j,\mathbf{k}}$ is the number of realizations of $X_t = j, \Pi_{X_t} = \mathbf{k}$. The classical "smoothing" techniques are still applicable to these models.

4. Experiments and evaluation

The training and test corpus are extracted from *Le monde* newspaper. We use 22M words for training and a test corpus of 2M words. The vocabulary consists of the most frequent 5000 words. The training corpus has been labeled automatically by a set of 200 syntactic classes set by hand (Smaïli et al., 1999). All models used in the experiments are smoothed using absolute discounting method (Ney et al., 1994)

Table 1 shows perplexity results of 14 different Bayesian network language models. The models DBN1, DBN2 and DBN3 correspond to bi-gram, bi-class and tri-class models respectively. In order to achieve our objective to find the best model we set the bi-class model as baseline and extend it incrementally by incorporating additional lexical and/or syntactic context. We also introduce the concept of right context of a word. DBN6 is a typical example of this case that integrates not only the left class context of a word but also its right syntactic context. We obtain a 16.6% improvement with respect to DBN4 that proves the importance of right context. It is true that linguistically this is not a surprising result. On the other hand, it is difficult to realize the use of right context in speech recognition, but this could be achieved with a multi-pass decoding scheme. DBN5, on the other hand, shows that left context is quite important. That is why its removal reduces the results by 7.8%. A significant perplexity reduction is observed if a word not only depends on its syntactic but also lexical context. Indeed, DBN11 yields an improvement of 24.6% with respect to DBN4. This results confirms that lexical history is indispensable and that syntactic history provides a significant improvement.

Pushing forwards this strategy, we achieve a model that is not only much better than the bi-class but also better than the bi-gram. Indeed, the model DBN14 that is shown in Fig. 3, reduces the perplexity by 57.9% with respect to DBN2 and 2.4% with respect to DBN1 (bi-gram). This new approach outperforms the classical bi-gram. These results show that our approach leads indeed to a new category of language models that are able to achieve better performance than classical ones.

Table 1: Training is performed on 11 months of "Le Monde 87". Perplexity is computed using 1 month containing more than 1M words.

	DBN structure (or JPD)	Perplexity
DBN1	$\prod_t P(w_t\|w_{t-1})$	65.24
DBN2	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1})$	151.31
DBN3	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1}c_{t-2})$	130.00
DBN4	$\prod_t P(w_t\|c_t, c_{t-1})P(c_t\|c_{t-1})$	113.13
DBN5	$\prod_t P(w_t\|c_t, c_{t+1})P(c_t\|c_{t-1})$	121.98
DBN6	$\prod_t P(w_t\|c_{t-1}, c_t, c_{t+1})P(c_t\|c_{t-1})$	94.35
DBN7	$\prod_t P(w_t\|c_t, c_{t-1})P(c_t\|c_{t-1}, c_{t-2})$	97.19
DBN8	$\prod_t P(w_t\|c_t, c_{t+1})P(c_t\|c_{t-1}, c_{t-2})$	104.8
DBN9	$\prod_t P(w_t\|c_{t-1}, c_t, c_{t+1})P(c_t\|c_{t-1}, c_{t-2})$	81.06
DBN10	$\prod_t P(w_t\|w_{t-1}, c_{t-1}, c_t, c_{t+1})P(c_t\|c_{t-1})$	78.00
DBN11	$\prod_t P(w_t\|w_{t-1}, c_t)P(c_t\|c_{t-1})$	85.20
DBN12	$\prod_t P(w_t\|w_{t-1}, c_t)P(c_t\|c_{t-1}, c_{t-2})$	73.20
DBN13	$\prod_t P(w_t\|w_{t-1}, c_{t-1})P(c_t\|w_t)$	70.86
DBN14	$\prod_t P(w_t\|w_{t-1}, c_{t-1}, c_{t-2})P(c_t\|w_t)$	**63.67**

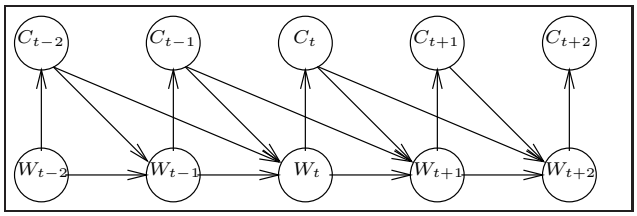

Figure 3: DBN structure that yields smaller perplexity w.r.t. bi-gram.

5. Conclusion and perspectives

We presented a new approach for language model construction based on dynamic Bayesian networks formalism. This approach has several advantages with respect to classical techniques. First of all, it infers the best model for language from training data in such a way that the resulting model is the best explanation of the corpus and it is not constrained by *a priori* assumptions. Another advantage is that all linguistic units considered in modeling are handled in one procedure. Hence the resulting models are consistent and easy to interpret. In this paper we tested different DBN structures in order to evaluate the potential of our approach. The results are promising and further detailed experiments are necessary to obtain better performance. On the other hand our main objective is to develop an algorithm

that infer *automatically* the best DBN for the language under consideration from training data. This will be the main direction of our future work.

6. References

Deviren, M. and K. Daoudi, 2001. Structural learning of dynamic Bayesian networks in speech recognition. In *Eurospeech 2001*, volume 1. Aalborg, Denmark.

Friedman, Nir, Kevin Murphy, and Stuart Russell, 1998. Learning the structure of dynamic probabilistic networks. In *UAI'98*, volume 1. Madison, Wisconsin.

Heckerman, David, 1995. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division.

Jelinek, F. and R.L. Mercer, 1980. Interpolated estimation of markov source parameters from sparse data. In *Pattern Recognition in Practice*. Amsterdam, Holland.

Ney, H., U. Essen, and R. Kneser, 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38.

Rosenfeld, R., 1994. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Ph.D. thesis, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213.

Smaïli, K., A. Brun, I. Zitouni, and J.P. Haton, 1999. Automatic and manual clustering for large vocabulary speech re cognition: A comparative study. In *European Conference on Speech Communication and Technology*, volume 4. Budapest, Hungary.