Une nouvelle approche de modélisation du langage par des réseaux Bayésiens dynamiques

Murat Deviren, Khalid Daoudi et Kamel Smaïli

INRIA-LORIA, équipe Parole
B.P. 101 - 54602 Villers les Nancy, France
Tél. : +33 (0)3 83 59 20 22 - Fax : +33 (0)3 83 59 19 27
Mél : 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

Le rôle d'un modèle de langage statistique est de modéliser les différents événements langagiers de la langue. Cette modélisation est assurée par l'affectation d'une probabilité à chaque séquence de mots hw_t (où h désigne l'historique du mot w_t). Cette probabilité est estimée à l'aide d'une chaîne de Markov d'un certain ordre (souvent 1 ou 2). Les n-grammes issus des chaînes de Markov sont les modèles les plus couramment utilisés. En partant du constat que certains mots ont un comportement similaire, leur regroupement en classes est envisageable et des modèles dits n-classes se substituent parfois aux n-grammes, et se combinent dans d'autres cas. L'objectif étant de réduire la complexité des modèles de base, de les généraliser, de passer outre les limites des données manquantes, et d'introduire des connaissances linguistiques plus importantes (morphologiques, syntaxiques, sémantiques et autres). L'exploitation de différentes connaissances linguistiques peut se faire par une simple combinaison linéaire de modèles de langage différents [4] ou par l'utilisation du maximum d'entropie [7].

Comme nous le verrons par la suite, les n-grammes et n-classes sont en fait des réseaux Bayésiens dynamiques (RBDs) très particuliers. Dans cet article, Nous nous proposons d'utiliser le formalisme des RBDs afin de mieux exploiter chacune des unités linguistiques considérées dans la modélisation. Nous développons une approche qui permet de traiter toutes ces unités dans une seule et même procédure et de fournir de nouveaux modèles de langage performants. Le principe de cette approche est de considérer des RBDs où une variable (qui peut être un mot ou une classe, ou tout autre unité linguistique) peut dépendre d'un sous-ensemble de variables. Ces liens de dépendance inter-unités linguistiques peuvent être déterminés automatiquement ou manuellement. Notre objectif ultime est de proposer une méthode d'identification automatique à partir du corpus d'apprentissage du RBD optimal pour modéliser le langage. Pour atteindre cet objectif, nous étudions dans cet article la faisabilité de cette approche en testant d'abord des modèles pour lesquels la topologie du RBD correspondant est fixé manuellement. Soulignons qu'un avantage de notre approche par rapport à l'interpolation linéaire est qu'elle ne fait pas de moyenne pondérée entre modèles fondés sur ces unités, mais intègre l'ensemble des unités linguistiques au sein d'une même structure qui ne fait pas la différence entre classes et mots. Comparé au principe du maximum d'entropie, on peut dire que notre approche conduit à des modèles qui ont une interprétation plus simple.

2. RÉSEAUX BAYÉSIENS DYNAMIQUES

Notre approche est fondée sur le formalisme des réseaux Bayésiens dynamiques (RBDs) qui est une généralisation des réseaux Bayésiens statiques aux processus dynamiques. Brièvement, un réseau Bayésien (statique) consiste à associer un graphe acyclique orienté à une distribution jointe de probabilités (DJP) $P(X)$ d'un ensemble de variables aléatoires $X = \{X_1..., X_n\}$. Les noeuds de ce graphe représentent les variables aléatoires, et les flèches codent les indépendances conditionnelles (IC) (supposées) existantes dans $P(X)$. Un réseau Bayésien ainsi est complètement défini par une structure graphique S et une paramétrisation numérique Θ de probabilités conditionnelles des variables sachant leurs parents. En effet, la DJP se factorise sous la forme :

$$P(X) = \prod_{i=1}^{n} P(X_i|\Pi_i),$$

où Π_i dénote les parents de X_i dans S.

Un RBD code la distribution jointe de probabilités d'un ensemble de variables $X[t] = \{X_1[t], \ldots, X_n[t]\}$ évoluant dans le temps. Si on se fixe T pas de temps, le RBD peut être considéré comme un réseau Bayésien (statique) avec $T \times n$ variables. En utilisant la propriété de factorisation, la densité jointe de probabilités de $\mathbf{X}_1^T =$

$\{X[1], \ldots, X[T]\}$ est :

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

où Π_{it} dénote les parents de $X_i[t]$. Dans la littérature, les RBDs sont définis en faisant l'hypothèse que $X[t]$ est un processus markovien [3]. Dans cet article, nous affaiblissons cette hypothèse pour permettre des processus non-markovien et pour que le processus $X[t]$ satisfasse :

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

pour des entiers positifs τ_p et τ_f donnés. Graphiquement, l'hypothèse ci-dessus signifie qu'une variable au temps t peut avoir des parents dans l'intervalle $[t-\tau_p, t+\tau_f]$ (voir [2] pour les détails).

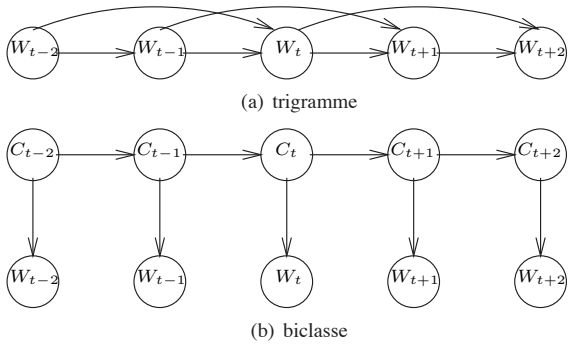
De ce point de vue, il est clair que les modèles de langage classiques peuvent être représentés comme des RBDs. En effet, les modèles n-gramme supposent que la probabilité d'une séquence de mots se factorise comme le produit de probabilités conditionnelles de chaque mot sachant son historique récent de $n-1$ mots. Précisément, si V est l'ensemble des mots et $w_1^T = w_1 ... w_T \in V^T$ est une séquence, on suppose que :

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

Ainsi, si W_t est une variable aléatoire discrète prenant ses valeurs dans V pour tout t, un n-grammes peut être représenté par le RBD de la figure 1-(a) (avec $n = 3$, i.e., un trigramme). Les modèles à base de classes représentent l'historique sur les classes de mots plutôt que sur les mots. Précisément, si $C = \{l_1, ..., l_m\}$ est l'ensemble des étiquettes de classe et $c_1^T = c_1 ... c_T \in C^T$ est une séquence de classes, on suppose que :

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

Ainsi si C_t est une variable aléatoire discrète prenant ses valeurs dans C pour tout t, un n-classe peut être représenté par le RBD de la figure 1-(b) (avec $n = 2$, i.e., un bi-classe).



(a) trigramme

(b) biclasse

FIG. 1: modèles trigramme et biclasse

3. RBDS POUR LA MODÉLISATION DU LANGAGE

Les n-grammes et n-classes sont les modèles les plus utilisés actuellement. Les n-grammes ont prouvé leur efficacité lorsqu'ils sont appris sur de gros corpus de données. Ces modèles ont comme inconvénient principal d'être algorithmiquement gourmand étant donné que l'on doit estimer $|V|^n$ probabilités où V est le vocabulaire et n l'ordre du modèle de langage. Malgré l'élagation opérée par les techniques de "cut-off", ces modèles demeurent complexes. Les modèles n-classes quant à eux sont moins complexes car le nombre de classes utilisés est souvent très inférieur à celui du vocabulaire. Ils permettent également une meilleure généralisation. Néanmoins, ils sont moins précis et leur perplexité est plus importante. C'est pourquoi ils sont souvent combinés avec les n-grammes. La procédure de mise en oeuvre d'un modèle de langage dans ce cas consiste à à déterminer *a priori* l'ordre du modèle(n). Ensuite, les deux modèles (classes et grammes) sont appris séparément en utilisant le maximum de vraisemblance comme critère d'apprentissage. L'étape suivante consiste à combiner linéairement les modèles ou d'intégrer leurs caractéristiques respectives dans une architecture du type maximum d'entropie.

Cette approche est intéressante, mais si nous voulons repousser les limites de ces modèles et mettre à profit leurs informations lexicales et syntaxiques, il est préférable de les considérer comme un tout indivisible devant être appris dans une même et unique procédure. Dans l'approche que nous proposons, nous souhaitons d'abord que "l'ordre" du modèle soit déterminé automatiquement à partir des données d'apprentissage. Ensuite, comme un mot est influencé par son contexte lexical et syntaxique, nous proposons de prendre en compte directement cette influence linguistique dans l'estimation du mot à prédire.

Le formalisme des RBDs nous offre le cadre théorique et algorithmique pour réaliser cet objectif. Notre idée principale est de ne faire aucune hypothèse *a priori* sur la façon de représenter le langage mais plutôt de considérer toutes les données disponibles (mots et classes) comme des observations du système dynamique $\{W_t, C_t\}$, notre but ensuite est de trouver le modèle qui décrit ces observations au mieux (en terme de perplexité). De cette façon, nous laissons les données dicter ce qui influence la prononciation d'un mot. Dans la terminologie des réseaux Bayésiens, ceci est le problème d'*apprentissage structurel* : trouver la structure graphique (et sa paramétrisation numérique) qui explique "au mieux" les données. Il existe des algorithmes [6] qui tente de résoudre ce problème (assez difficile) et qui fonctionnent plus au moins bien selon les applications. Notre but dans cet article n'est pas d'utiliser de tels algorithmes (sachant que notre objectif ultime et de développer un algorithme d'apprentissage structurel qui soit bien adapté à la modélisation du langage). Notre but est plutôt de vérifier que notre approche peut effectivement concurrencer les n-grammes et n-classes classiques. Pour ce faire nous considérons un ensemble assez riche de structures graphiques que nous jugeons plausibles pour modéliser le langage. Ensuite, nous évaluons sur une application concrète les performances de plusieurs RBDs appartenant à cet ensemble en les comparant aux n-grammes et n-classes.

3.1. Structures graphiques utilisées

Pour définir un ensemble de structures graphiques plausibles pour la modélisation du langage, nous devons spécifier des hypothèses d'IC qui soient linguistiquement

informatives et faciles à interpréter. Pour ce faire, nous commençons par assouplir l'hypothèse d'IC des n-classes en considérant qu'un mot peut dépendre non seulement de la classe courante mais aussi d'autres classes dans un contexte limité dans le passé et/ou le futur. Pour incorporer les propriétés des n-grammes, nous autorisons aussi la dépendance d'un mot de son historique (lexical). Enfin, nous considérons qu'une classe peut dépendre non seulement de son historique mais aussi du mot courant et/ou d'autres mots dans un contexte limité dans le passé et/ou le futur. La figure 2 montre un exemple de RBD possible dans notre modélisation.

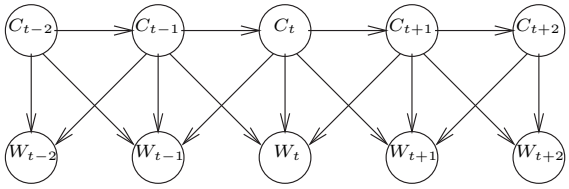


FIG. 2: RBD où chaque mot (resp. classe) dépend de la classe antérieure, présente et futur (resp. antérieure).

La probabilité jointe associée à un modèle spécifique s'exprime comme :

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

où π_{w_t} (resp. π_{c_t}) est la réalisation des parents Π_{W_t} (resp. Π_{C_t}) de W_t (resp. C_t). Les paramètres du modèle sont donnés par les tables de probabilités conditionnelles (indépendantes du temps t) $P(w_t|\pi_{w_t})$ and $P(c_t|\pi_{c_t})$ que nous notons, pour $X_t \in \{W_t, C_t\}$, par :

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

Ces paramètres sont estimés en utilisant le critère du maximum de vraisemblance, ce qui donne :

$$\theta_{x,j,\mathbf{k}} = \frac{N_{x,j,\mathbf{k}}}{\sum_i N_{x,i,\mathbf{k}}} \tag{7}$$

où $N_{x,j,\mathbf{k}}$ est le nombre de réalisations $X_t = j, \Pi_{X_t} = \mathbf{k}$. Les schémas classiques de "discounting", de "smoothing" et de "back-off" restent applicables à ce type de modèles.

3.2. Expérimentations et évaluation

Les corpus d'apprentissage et de test sont extraits du journal *Le monde*. Nous avons utilisé 22M de mots pour l'apprentissage et 2M de mots pour le test. Le vocabulaire est constitué des 5000 mots les plus fréquents. Le corpus d'apprentissage a été étiqueté automatiquement par un ensemble de 200 classes syntaxiques définies manuellement [1]. Tous les modèles utilisés dans cette expérimentation ont été lissés par la méthode de "absolute discounting" [5].

La table 1 montre les perplexités que nous avons obtenues pour 14 réseaux Bayésiens dynamiques différents. Les modèles RBD1, RBD2 et RBD3 sont respectivement le bigramme, le biclasse et le triclasse. Afin d'atteindre l'objectif de trouver de meilleurs modèles de langage, nous avons pris le biclasse comme modèle noyau que nous avons enrichi incrémentalement par un contexte lexical et/ou syntaxique plus important. Nous avons introduit également la notion du contexte droit d'un mot. C'est le cas du modèle RBD6 qui intègre non seulement le contexte gauche de la classe d'un mot mais aussi son contexte syntaxique droit. On obtient ainsi une amélioration de $16,6\%$ par rapport à RBD4, ce qui prouve l'importance de la prise en compte du contexte droit. Il est vrai que linguistiquement parlant cela ne constitue nullement une surprise. En reconnaissance de la parole en revanche, il est difficile d'envisager la prise en compte du contexte droit mais cela peut être rendu possible grâce à un décodage multi-passes. D'autre part le modèle RBD5 montre que l'influence du contexte gauche est importante, c'est pourquoi son omission fait baisser les résultats de $7,8\%$. Une réduction considérable dans la perplexité est observée lorsque on fait dépendre un mot non seulement de son contexte syntaxique mais aussi lexical. Ainsi, RBD11 apporte une amélioration de $24,6\%$ par rapport à RBD4. Ceci confirme bien que l'historique lexical est indispensable et montre que l'historique syntaxique apporte une contribution indéniable.

TAB. 1: Perplexité des réseaux Bayésiens dynamiques

	structure (ou DJP) du RBD	Perplexité
RBD1	$\prod_t P(w_t\|w_{t-1})$	65.24
RBD2	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1})$	151.31
RBD3	$\prod_t P(w_t\|c_t)P(c_t\|c_{t-1}c_{t-2})$	130.00
RBD4	$\prod_t P(w_t\|c_t, c_{t-1})P(c_t\|c_{t-1})$	113.13
RBD5	$\prod_t P(w_t\|c_t, c_{t+1})P(c_t\|c_{t-1})$	121.98
RBD6	$\prod_t P(w_t\|c_{t-1}, c_t, c_{t+1})P(c_t\|c_{t-1})$	94.35
RBD7	$\prod_t P(w_t\|c_t, c_{t-1})P(c_t\|c_{t-1}, c_{t-2})$	97.19
RBD8	$\prod_t P(w_t\|c_t, c_{t+1})P(c_t\|c_{t-1}, c_{t-2})$	104.8
RBD9	$\prod_t P(w_t\|c_{t-1}, c_t, c_{t+1})P(c_t\|c_{t-1}, c_{t-2})$	81.06
RBD10	$\prod_t P(w_t\|w_{t-1}, c_{t-1}, c_t, c_{t+1})P(c_t\|c_{t-1})$	78.00
RBD11	$\prod_t P(w_t\|w_{t-1}, c_t)P(c_t\|c_{t-1})$	85.20
RBD12	$\prod_t P(w_t\|w_{t-1}, c_t)P(c_t\|c_{t-1}, c_{t-2})$	73.20
RBD13	$\prod_t P(w_t\|w_{t-1}, c_{t-1})P(c_t\|w_t)$	70.86
RBD14	$\prod_t P(w_t\|w_{t-1}, c_{t-1}, c_{t-2})P(c_t\|w_t)$	**63.67**

En poussant cette stratégie encore plus, nous sommes arrivés à concevoir un modèle qui non seulement est largement meilleur que le biclasse mais qui est meilleur que le bigramme aussi. En effet, le modèle RBD14, dont la structure graphique est montrée dans la figure 3, améliore de $57,9\%$ par rapport à RDB2 et de $2,4\%$ par rapport à RBD1 (le bigramme). Même si l'amélioration par rapport au bigramme n'est pas considérable (dans cette application), ce résultat (en plus des précédents) montre surtout que notre approche peut effectivement conduire à de nouvelles catégories de modèles de langage qui peuvent concurrencer les catégories classiques.

4. CONCLUSION ET PERSPECTIVES

Nous avons présenté une nouvelle approche pour la construction de modèles de langage fondée sur le formalisme des réseaux Bayésiens dynamiques. Cette approche possède plusieurs avantages par rapport aux techniques classiques. Tout d'abord, elle permet d'inférer le meilleur modèle pour le langage traité à partir des corpus d'apprentissage, le modèle tente ainsi d'expliquer au mieux

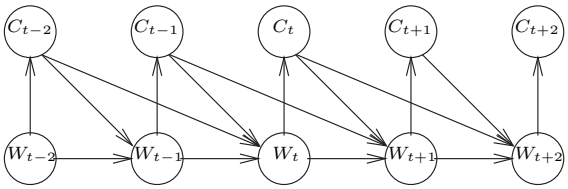


FIG. 3: RBD donnant une perplexité plus faible que le bigramme.

les données et n'est pas contraint par des hypothèses *a priori*. En outre, elle incorpore en une seule et même procédure toutes les unités linguistiques considérée dans la modélisation. Les modèles qui en résultent sont ainsi consistants et faciles à interpréter. Dans cet article, nous avons choisi de tester plusieurs type de RBDs pour évaluer le potentiel de notre approche. Les résultats montrent qu'elle est prometteuse et mérite d'être plus approfondie. D'autres expérimentations du même genre doivent évidemment être effectuées pour mieux analyser le comportement de cette technique. Mais notre objectif majeur est de développer un algorithme qui permet d'inférer *automatiquement* à partir des corpus d'apprentissage le meilleur RBD pour modéliser le langage traité. Ceci fera l'objet de nos futurs travaux.

RÉFÉRENCES

[1] K. Smaïli et A. Brun et I. Zitouni et J.P. Haton. Automatic and manual clustering for large vocabulary speech re cognition : A comparative study. In *European Conference on Speech Communication and Technology*, volume 4, pages 1795–1798, Budapest, Hungary, September 1999.

[2] M. Deviren et K. Daoudi. Structural learning of dynamic Bayesian networks in speech recognition. In *Eurospeech 2001*, volume 1, pages 1669–1673, Aalborg, Denmark, 2001.

[3] Nir Friedman et Kevin Murphy et Stuart Russell. Learning the structure of dynamic probabilistic networks. In *UAI'98*, volume 1, pages 139–147, Madison, Wisconsin, 1998.

[4] F. Jelinek et R.L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Pattern Recognition in Practice*, pages 381–397, Amsterdam, Holland, 1980.

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

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

[7] R. Rosenfeld. *Adaptive Statistical Language Modeling : A Maximum Entropy Approach*. PhD thesis, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213, April 1994.