Approche stochastique

Sommaire

Résumé

Ce chapitre présente les approches stochastiques du problème de la compréhension. Le modèle théorique fondamental est détaillé en 2.2. Les applications classiques s'appuyant sur ce modèle sont exposées dans la section 2.3. La section 2.4 s'intéresse aux premiers systèmes de compréhension à base de réseaux bayésiens dynamiques qui ont inspiré les travaux présentés dans ce document.

2.1 Introduction

Les approches stochastiques de la compréhension du dialogue oral sont la principale alternative aux méthodes linguistiques à base de grammaires. Les méthodes stochastiques permettent de concevoir des systèmes adaptés aux spécificités de l'oral, facilement évolutifs, et robustes aux erreurs de transcription.

Basées sur le choix d'un modèle stochastique, adapté à la tâche visée, et l'apprentissage automatique de ses paramètres, ces méthodes réduisent les besoins en expertise humaine lors du développement d'applications tout en obtenant des résultats au moins comparables à ceux des méthodes à base de règles.

Le modèle théorique qui sous-tend les approches stochastiques de la compréhension est détaillé en 2.2. La section 2.3 s'intéresse à quelques systèmes de référence. L'approche à base de réseaux bayésiens dynamiques, qui a initié les travaux présentés dans ce document, est exposée dans la section 2.4.

2.2 Modèle théorique

L'approche stochastique de la compréhension est principalement basée sur le paradigme du *canal bruité* déjà utilisé pour formaliser le problème de reconnaissance de la parole (Jelinek, 1976). Ce paradigme est appliqué au problème de la compréhension sous deux hypothèses.

La première considère que le sens d'un message peut être exprimé par une séquence d'unités de sens en correspondance séquentielle avec les observations acoustiques (Pieraccini et al., 1993). Les unités sémantiques sont rassemblées dans un dictionnaire de *concepts*.

La seconde hypothèse consiste à considérer la représentation acoustique d'un message comme issue une séquence de concepts dégradée par un canal bruité de caractéristiques inconnues.

Sous ces deux hypothèses, le problème de la compréhension d'un message orale se ramène donc à un problème de décodage : il s'agit de déterminer la séquence conceptuelle $\hat{C}$ dont la probabilité *a posteriori* est maximale pour une séquence acoustique A observée.

Formellement, on a donc :

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, P(C|A) = \underset{C}{\operatorname{argmax}} \sum_{W} P(W, C|A) \tag{2.1}$$

où $W = w_1, ..., w_L$ est la séquence de mots composants le message.

Le théorème de Bayes permet de renverser le conditionnement pour obtenir :

$$\hat{C} = \underset{C}{\operatorname{argmax}} \sum_{W} P(A|W, C)P(W, C) \tag{2.2}$$

Sous l'hypothèse d'indépendance entre la séquence acoustique observée A et la séquence de concepts C, l'équation précédente devient :

$$\hat{C} = \underset{C,W}{\operatorname{argmax}} P(A|W)P(W,C) \tag{2.3}$$

où A ne dépend plus que de la séquence de mots W.

La probabilité $P(A|W)$ étant évaluée dans le cadre de la reconnaissance de la parole, l'enjeu de la compréhension est donc la résolution de l'équation :

$$\hat{C} = \underset{C}{\operatorname{argmax}} P(W,C) = \underset{C}{\operatorname{argmax}} P(W|C)P(C) \tag{2.4}$$

La probabilité $P(W|C)$ d'une séquence de mots sachant une séquence conceptuelle représente le modèle de réalisation lexicale. Elle est généralement estimée par des probabilités n-grammes de mots conditionnées par le concept associé au mot courant et l'on a :

$$P(W|C) \backsimeq \prod_{i=1}^{L} P(w_i|w_{i-1},...,w_{i-n},c_i)$$

La probabilité $P(C)$, probabilité *a priori* d'une séquence de concepts, représente le modèle sémantique. Son estimation est également réalisée par des probabilités m-grammes de concepts selon :

$$P(C) \backsimeq \prod_{i=1}^{L} P(c_i|c_{i-1},...,c_{i-m})$$

Cette modélisation classique est une chaîne de Markov d'ordre n où seules les n dernières observations sont utilisées pour la prédiction du mot ou du concept suivant (i.e. un bigramme est une chaîne de Markov d'ordre 2).

On remarquera donc d'emblée que la formulation probabiliste du problème de la compréhension de la parole rend complexe une interprétation structurée des requêtes utilisateurs. La première hypothèse permettant l'expression sous forme d'un canal bruité impose la dépendance séquentielle des informations conceptuelles extraites. Or, il est clair que la sémantique d'une phrase présente très souvent des dépendances à long terme entre ses constituants. Les approches développées vont donc aborder ce problème de différentes manières : soit en l'occultant complètement par le biais d'une représentation purement "à plat" (qui peut se révéler suffisante pour un grand nombre d'applications), soit en composant différents niveaux de décodage permettant l'emboîtement des unités décodées (et devenant ainsi comparables à des grammaires non-contextuelles sémantiques probabilisées), soit finalement par le biais d'approches composites basées sur une structuration progressive de l'hypothèse.

2.3 Quelques applications

2.3.1 Le système CHRONUS

Un de premiers système basé sur le modèle stochastique présenté en 2.2 est le système CHRONUS (*Conceptual Hidden Representation Of Natural Unconstrained Speech*) (Pieraccini et al., 1991). Le décodeur conceptuel sur lequel repose CHRONUS utilise un modèle stochastique de type modèle de Markov Caché (*Hidden Markov Model*, HMM) dont les états représentent les concepts. Les séquences de mots associées à un concept donné sont également modélisées par un processus markovien représenté par un modèle de langage n-grammes conditionné par le concept. Les états du HMM conceptuel sont caractérisés par ces modèles de langage qui utilisent des techniques de repli (Riccardi et al., 1995) permettant d'attribuer des probabilités non nulles aux n-uplets non rencontrés dans les données observées.

Les paramètres du modèle sont appris à partir d'un corpus dans lequel les concepts sont associés à des portions de phrase. CHRONUS est évalué sur la tâche de renseignements dédiée aux voyages aériens ATIS (*Air Travel Information System*) décrite en détails dans (Dahl et al., 1994).

2.3.2 Le système CHANEL

Dans le système CHANEL (Kuhn et De Mori, 1995; De Mori, 1998), les règles d'interprétation sémantique sont apprises au moyen d'une forêt d'arbres de décision spécifiques appelés Arbres de Classification Sémantique (ACS). Contrairement à CHRONUS qui associe une séquence de mots à un concept, chaque ACS examine le message complet pour construire une partie de sa représentation sémantique. Chaque noeud d'un ACS est associé à une question binaire portant sur la présence ou non de séquences de mots prédéfinies. Ces questions sont générées et apprises automatiquement à partir de corpus dont les annotations préalables consistent uniquement en la liste des concepts présents dans chaque phrase. La sélection des questions s'appuie sur la maximisation de l'information d'un noeud à l'autre.

CHANEL est donc un système hybride dont l'originalité majeure est l'utilisation d'arbres de décision pour l'étiquetage sémantique.

2.3.3 L'approche HUM

Le modèle HUM (*Hidden Understanding Models*), proposé par (Miller et al., 1994), est inspiré des modèles stochastiques utilisés en reconnaissance de la parole et notamment des modèles de Markov cachés comme CHRONUS. Son objectif est de retrouver la structure sémantique la plus probable d'un message donné grâce à la résolution de l'équation (2.4).

La représentation sémantique est structurée sous forme d'arbres dont les nœuds non-terminaux sont les concepts abstraits et leurs composants, et les feuilles sont les mots du message. Un exemple d'arbre associé à une phrase de la tâche ATIS extrait de (Miller et al., 1994) est reproduit dans la figure 2.1.

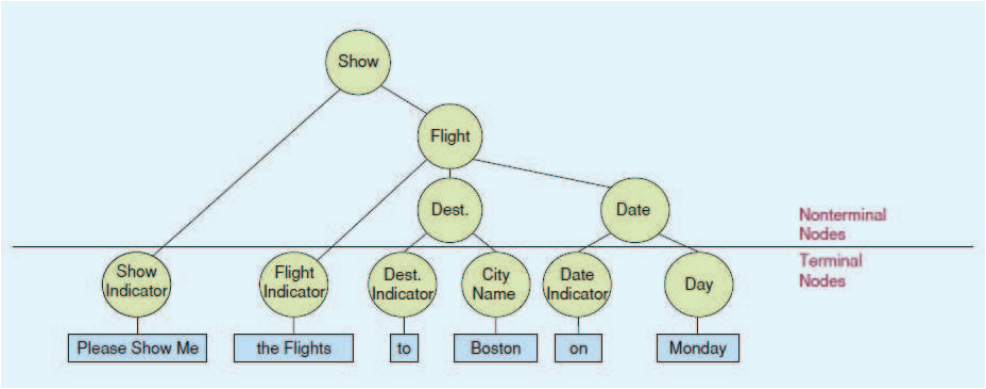

FIGURE 2.1 – *Exemple d'arbre associé à une phrase de la tâche ATIS*

Le modèle HUM est un modèle génératif basé sur deux composants stochastiques. Le premier composant, basé sur un modèle de langage sémantique, décide du sens à exprimer - i.e. *quoi dire* - tandis que le deuxième composant, basé sur un modèle de réalisation lexicale, sélectionne les séquences de mots adaptées à l'expression du sens - i.e. *comment le dire*.

Dans le modèle de langage sémantique, à chaque concept abstrait correspond un réseau de transition probabiliste contenant un état pour chacun de ses composants, un état d'entrée et un état de sortie. Un réseau est complet, autorisant ainsi tous les chemins sémantiques. Les probabilités de transition associées aux arcs du réseau sont obtenues en calculant $P(\acute{E}tat_n|\acute{E}tat_{n-1}, Concept)$. Les transitions entre états sont donc conditionnées par le contexte du concept courant qui permet de privilégier certains chemins par rapport à d'autres selon les paramètres collectés lors de l'apprentissage.

Les feuilles de l'arbre sémantique sont associées au modèle de réalisation lexicale qui repose sur les probabilités entre les mots, dans un contexte donné. Ces probabilités s'écrivent donc $P(Mot_n|Mot_{n-1}, Concept)$.

Le problème de la compréhension se ramène donc à la détermination du chemin le plus probable $\hat{T}$ dans la combinaison des réseaux sémantiques et lexicaux qui composent le message. Cette détermination repose sur la maximisation de la probabilité :

$$P(T) = \prod_{t \in T} \left[\begin{array}{ll} P(\acute{E}tat_n|\acute{E}tat_{n-1}, Concept) & \text{si} \quad t \in \{\text{Modèle de langage sémantique}\} \\ P(Mot_n|Mot_{n-1}, Concept) & \text{si} \quad t \in \{\text{Modèle de réalisation lexicale}\} \end{array} \right.$$

dont l'algorithme de calcul exacte est exponentiel par rapport à la longueur du message. Le calcul est donc réalisé par l'algorithme de Viterbi (programmation dynamique) aidé par la suppression des chemins de plus faibles probabilités (recherche en faisceau ou *bean search*).

2.3.4 Les systèmes HMM du LIMSI

Un premier système de compréhension stochastique est proposé par (Minker et al., 1996) et évalué par comparaison avec l'analyseur sémantique du LIMSI basé sur une grammaire de cas. Ce système utilise un modèle de Markov caché du premier ordre entrainé sur les données de la tâche ATIS. La représentation sémantique est voisine de celle utilisée par la grammaire de cas et permet l'extraction des valeurs associées aux concepts. Les performances de ce système restent modestes, incitant à l'introduction ultérieure d'informations contextuelles.

Plus récemment, (Maynard et Lefèvre, 2002) présente un système de compréhension stochastique développé sur des données collectées grâce au système de dialogue ARISE du LIMSI (Lamel et al., 2000). Ce système de dialogue est dédié à la réservation téléphonique de billets de train et propose également des renseignements sur les horaires, les tarifs et les prestations. Le système stochastique proposé par (Maynard et Lefèvre, 2002) utilise une représentation sémantique à plat, spécifique à la tâche. Le dictionnaire sémantique défini comporte 64 concepts auxquels valeurs normalisées et modalité (affirmative, négative) peuvent être associées. Ainsi, un énoncé est représenté par une liste de triplets [mode, concept, valeur normalisée].

Le modèle stochastique développé est conforme au modèle génératif présenté au paragraphe 2.2, utilisant des bigrammes conceptuels ($m = 1$ dans le modèle du paragraphe 2.2) et conditionnant le mot courant par le concept courant ($n = 0$ dans le modèle du paragraphe 2.2).

La comparaison des performances de ce système à celles obtenues par l'analyseur à base de grammaire de cas du LIMSI met en évidence sa robustesse aux erreurs de reconnaissance. Il est également montré que le système reste performant lorsque l'on réduit la taille du corpus d'apprentissage.

2.3.5 L'approche par FSM

Suivant une proposition initiale de (Pereira et Wright, 1997), une stratégie pour la compréhension de la parole basée sur l'utilisation de machines à états finis (*Finite State Machine*, FSM) est présentée en détails dans (Raymond et al., 2006). Son fonctionnement est résumé dans la figure 2.2, par la description du système MEDIA du LIA. L'interprétation débute par une transduction pour laquelle les modèles de langages stochastiques sont implémentés sous forme de FSM qui émettent des constituants sémantiques. Ces constituants correspondent aux concepts définis par l'ontologie de la tâche. A chaque concept est attaché une chaîne de mots qui lui sert de support et à partir de laquelle la valeur peut être obtenue (e.g. la date, les noms propres ou les informations numériques).

Un FSM est construit pour chaque concept élémentaire. Ces FSM sont des transducteurs qui prennent des mots en entrée et proposent en sortie les concepts supportés par les locutions reconnues. Ils peuvent être définis manuellement pour les concepts

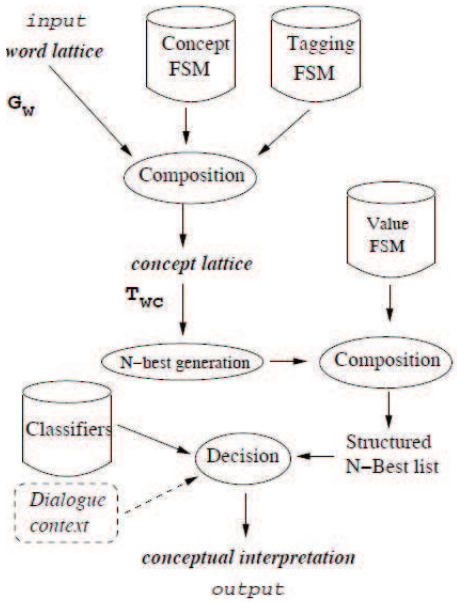

FIGURE 2.2 – *Exemple d'un système à base de FSM, le système* MEDIA *du* LIA

indépendants de la tâche (e.g. les dates ou les nombres) ou appris à partir des données fournies par un corpus d'apprentissage annoté. Tous ces transducteurs sont regroupés dans un unique transducteur, représentant leur union. Afin d'obtenir la meilleure séquence de concepts correspondant à la séquence de mots, un étiqueteur HMM, lui-même implémenté sous forme d'un FSM, est entraîné sur le corpus d'apprentissage. Enfin, une dernière étape de transduction est appliquée à chaque sous-séquence associée à un concept afin d'obtenir la valeur normalisée. Tous ces traitements peuvent être regroupés en appliquant les opérations adéquates sur les FSM intermédiaires. La disponibilité de l'AT&T FSM toolkit (Mohri et al., 2002) permettant l'implémentation de toutes les opérations usuelles sur les FSM est un grand atout de cette approche.

2.3.6 L'approche HVS

Le modèle HVS (*Hidden Vector State*), proposé par (He et Young, 2003, 2006), est dédié à l'analyse sémantique hiérarchique. Il est composé d'un modèle de Markov caché dans lequel chaque état représente un automate à pile de taille finie. La figure 2.3, issue de (He et Young, 2006), montre la séquence d'états du HSV correspondant à l'arbre d'analyse d'un message.

Les transitions entre états sont rassemblées dans des piles d'opérations d'entrée ou de sortie distinctes, limitées de façon à produire un espace de recherche calculable. Ce modèle est assez complexe pour capturer des structures hiérarchiques mais peut cependant être entraîné automatiquement à partir de données annotées sommairement.

Chaque état est représenté à l'instant t par un vecteur conceptuel c_t de dimension

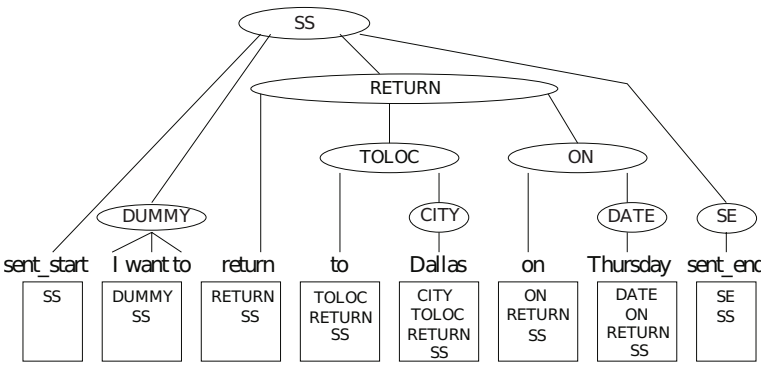

FIGURE 2.3 – *Arbre d'analyse d'un message et vecteurs d'états correspondants pour HVS*

D_t :

$$c_t = [c_t[1], c_t[2], \cdots, c_t[D_t]]$$

où $c_t[1]$ est le concept situé au sommet de la pile et $c_t[D_t]]$ est le concept racine (libellé "SS" dans la figure 2.3.

Pour une séquence de mots W, une séquence de vecteurs conceptuels C et une séquence d'opérations de sortie de la pile données, la probabilité jointe peut être décomposée sous la forme :

$$P(W, C, N) = \prod_{t=1}^{T} P(n_t|c_{t-1}) P(c_t[1]||[c_t[2 \cdots D_t]) P(w_t|c_t)$$

où n_t est le vecteur décrivant les opérations de modification de la pile, à valeurs dans l'intervalle $||[0, \cdots, D_t]||$ et $c_t[1] = c_{w_t}$ est le nouveau concept au sommet de la pile associé au mot w_t à l'instant t.

Le système nécessite donc l'apprentissage de trois tables de probabilités conditionnelles distinctes :

- $P(n|c)$, loi des opérations de sortie des concepts de la pile,
- $P(c_t[1]||[c_t[2 \cdots D_t])$, loi d'entrée d'un concept au sommet de la pile,
- $P(w|c)$, loi de génération des mots

Chacune de ces tables est estimée par un entraînement utilisant un algorithme de maximisation de la vraisemblance des paramètres des modèles (*Expectation-Maximization*, EM). De plus, pour garantir la calculabilité, les états non consistants avec le mot courant et ses concepts associés sont supprimés lors de l'apprentissage. Les tables obtenues sont ensuite utilisées pour produire les arbres d'analyse grâce à un décodage de Viterbi.

Une version étendue du modèle HVS a été proposée récemment qui présente l'avantage de permettre les branchements gauches et droits lors du décodage (alors que la version initiale ne permet que les branchements gauches) (Jurcicek et al., 2008). Une représentation par modèle graphique (ces modèles sont décrits plus en détails dans le chapitre 7) est donnée dans la figure 2.4. On notera la complexité du modèle qui engendre des difficultés pour l'apprentissage de ses paramètres. De même, l'extension du

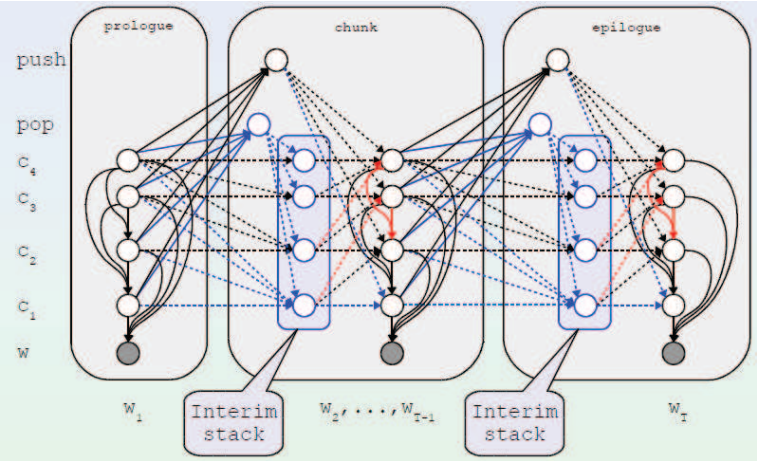

FIGURE 2.4 – *Représentation par modèle graphique du modèle HVS étendu avec insertion probabiliste (HVS-PP)*

modèle ne lui retire pas toutes ses limites ; ainsi le nombre de concepts pouvant être insérés simultanément (permettant le branchement droit) est codé dans la structure du modèle (limité à 3 dans le cas du modèle de la figure 2.4).

2.4 L'approche à base de réseaux bayésiens dynamiques

Un système de compréhension stochastique modélisé par des réseaux bayésiens dynamiques (*Dynamic Bayesian Networks*, DBN) est évoqué par (Bonneau-Maynard et Lefèvre, 2005) puis développé par (Lefèvre, 2006, 2007).

La représentation sémantique utilisée dans ce système est semblable à celle présentée dans (Maynard et Lefèvre, 2002) et détaillée en 2.3.4. Un énoncé est représenté par une liste de triplets [mode, concept, valeur normalisée].

Le modèle de compréhension intègre trois niveaux (mots W, concepts C et valeurs normalisées V), son objectif étant la recherche des séquences de concepts et de valeurs de probabilité *a posteriori* maximale selon :

$$\hat{C}, \hat{V} = \underset{C,V}{\operatorname{argmax}} P(C, V|W) = \underset{C,V}{\operatorname{argmax}} P(W|C, V)P(V|C)P(C) \tag{2.5}$$

La figure 2.5 présente le réseau bayésien dynamique utilisé par ce modèle pour deux mots successifs.

Le conditionnement par les valeurs, dont la liste est ouverte, augmente très sensiblement la complexité du modèle. Les auteurs résolvent ce problème en proposant un décodage en deux temps modélisé par un DBN à 2+1 niveaux selon :

$$\hat{C} = \underset{C}{\operatorname{argmax}} \sum_{V} P(W|C, V)P(V|C)P(C) \tag{2.6}$$

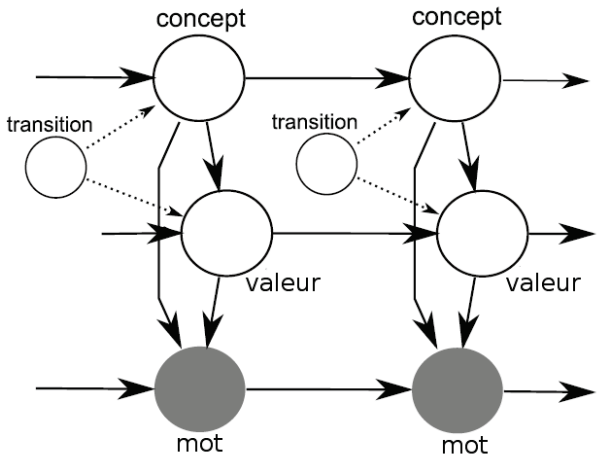

FIGURE 2.5 – *Modèle DBN à 3 niveaux*

$$\hat{V} = \underset{V}{\operatorname{argmax}}\, P(\hat{C}, V|W) = \underset{V}{\operatorname{argmax}}\, P(W|\hat{C}, V)P(V|\hat{C})P(\hat{C}) \tag{2.7}$$

Les concepts étant décodés par un premier DBN à partir des mots observés. Les valeurs sont obtenues par un second décodage utilisant les séquences de concepts produites lors de la première phase comme des observations. Les réseaux bayésiens dynamiques utilisés sont représentés figure 2.6

Les résultats encourageants fournis par ce système sur la tâche MEDIA 4 ont motivé les travaux présentés dans ce document, explorant la capacité de ces modèles à intégrer un système de compréhension de haut niveau.

2.5 Conclusion

Principales alternatives aux méthodes à base de règles, les méthodes stochastiques sont particulièrement adaptées à la tâche de compréhension de la parole. Dédiées à la modélisation de l'incertitude, ce sont par nature des méthodes robustes et peu dépendantes de l'application visée. Le coût de leur développement est essentiellement lié à la mise en forme de modèles théoriques par un ou plusieurs experts. Il est en cela très inférieur à celui des méthodes à base de règles. Les méthodes stochastiques étant basées sur l'apprentissage, leur emploi est cependant conditionné à la disponibilité de corpus d'entraînement de taille suffisante et cohérent avec l'espace sémantique visé.

Dans le contexte de la compréhension de la parole, méthodes à base de règles et méthodes stochastiques sont également dépendantes de la représentation sémantique choisie pour modéliser le sens des messages. Les travaux contemporains abordant cette problématique sont présentés dans le chapitre 3.

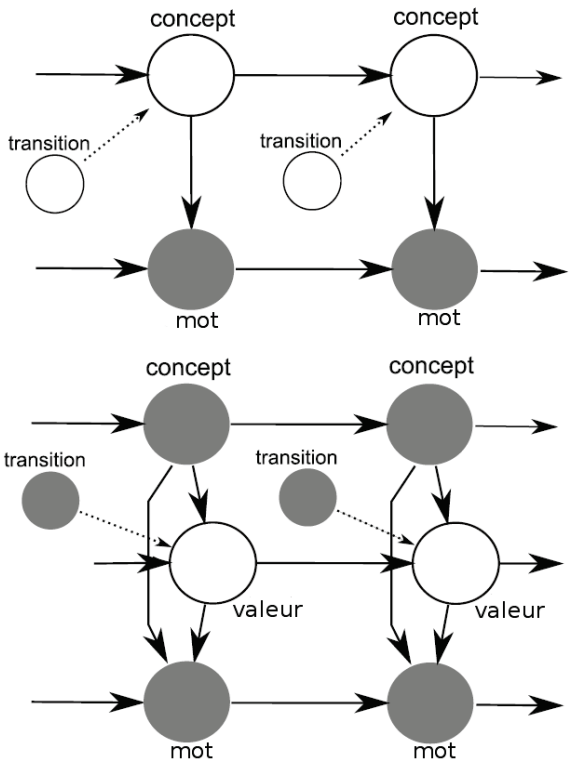

FIGURE 2.6 – *Modèle DBN à 2+1 niveaux*