UNE APPROCHE CONNEXIONNISTE POUR LE TRAITEMENT AUTOMATIQUE DES STRUCTURES SYNTAXIQUES DE LA LANGUE ARABE BASEES SUR LE FORMALISME NEO-KHALILIEN

Hadja Faîza Khellaf-Haned
Centre de Recherche Scientifique et Technique
pour le Développement de la Langue Arabe

Mohamed Tayeb Laskri
Groupe de Recherche en Intelligence Artificielle
Département d'Informatique
Facultés des Sciences-Université de Annaba

Résumé

Les techniques utilisant les réseaux de neurones tentent d'imiter la structure connexionniste du système nerveux pour en tirer les avantages concernant principalement les capacités d'apprentissage, de généralisation, de robustesse, de tolérance aux pannes et possibilité de traitement parallèle.

Cette étude entre dans le cadre de développement d'un système connexionniste pour l'analyse des structures syntaxiques de la langue arabe basées sur le formalisme linguistique néo-khalilien. Elle s'inscrit dans le domaine multidisciplinaire des sciences cognitives, intégrant ainsi la modélisation mathématique par les réseaux de neurones, les techniques informatiques du traitement automatique du langage naturel et les fondements d'une théorie linguistique.

Le système conçu intitulé "Neurokhall", est composé d'un réseau de neurones simplement récurrent, combiné avec une RAAM (Recursive Auto Associative Memory) pour le traitement des structures récursives. Le système accepte en entrée la phrase mot par mot sous forme de traits syntaxiques, et fournit en sortie les catégories syntaxiques sous la forme du schème générateur du niveau de la syntaxe de la théorie néo-khalilienne.

Mots clés

Intelligence artificielle - réseau de neurones - théorie néo-khalilienne - analyse syntaxique - schème générateur.

الملخص

تحاول التقنيات التي تستعمل شبكات الخلايا العصبية الاصطناعية تقليد هيكل الجهاز العصبي لاستخلاص المزايا المتعلقة أساسا بقدرات التعلم والتعميم، والمتانة، والقدرة على مواجهة الخلل، والقدرة على العلاج المتوازي.

وهذه الدراسة تدخل في إطار تطوير نظام مبني على شبكة الخلايا العصبية الاصطناعية لتحليل البنى التركيبية القائمة على النظرية الخليلية الحديثة، وذلك في الميدان المتعدد التخصصات للعلوم المعرفية حيث أنها تدمج التمثيل الرياضي، والتقنيات الإعلامية للعلاج الصوري للغة الطبيعية، وأسس النظرية اللسانية.

يتكون نظام « Neurokhall » من شبكة الخلايا العصبية المتراجعة إلى الوراء بطريقة بسيطة مركبة مع نظام RAAM (Recursive Auto Associative Memory) لعلاج البنى التركيبية المعقدة التي لها صفة الإطالة. ويمكن، في هذا النظام، إدخال الجملة كلمة بكلمة على شكل صفات تركيبية ليقدم في الأخير أصناف البنية التركيبية كما حددتها النظرية الخليلية الحديثة.

الكلمات المفاتيح

الذكاء الاصطناعي - شبكات الخلايا العصبية - النظرية الخليلية الحديثة - التحليل التركيبي - المثال.

Abstract

The techniques using neural networks attempt to imitate the connectionist structure of the nervous system to extract the advantages of the capacities of learning and generalisation (hardiness, tolerance to breakdowns, and the possibility of parallel treatment).

This survey enters in the setting of development of a connectionist system for the analysis of the syntactic structures of the arabic language based on the neo-khalilian linguistic formalism. It is undertaken in the multidisciplinary domain of the cognitive sciences, which integrates the mathematical modelisation by neural networks, the computer techniques of the automatic treatment of the natural language, and the basic concepts of a linguistic theory.

The conceived system untitled Neurokhall, is composed of a merely recurrent neurone network compound with a RAAM (Recursive Auto Associative Memory) for the recursive structure treatment. The system accepts in entrance the sentence word by word under the shape of syntactic features, and provides in exit the syntactic categories under the shape of the template of the level of the syntax of the neo-khalilian theory.

Key words

Artificial intelligence - neural networks - neo-khalilian theory - syntactic analysis - template.

1. Introduction

Issue des recherches pluridisciplinaires (neurobiologie, mathématiques, informatique, etc.), l'intelligence artificielle connexionniste connaît depuis plus de dix ans un essor important. Les applications à modélisation neuronale sont de plus en plus nombreuses. Au début de son essor, l'approche neuronale a été introduite pour les applications de bas niveau d'abstraction telles que le traitement du signal, la reconnaissance des formes, etc. Cependant, un regain d'intérêt est apparu quant à leur applicabilité aux tâches de haut niveau, traditionnellement dévolues à des traitements symboliques telles que le traitement automatique du langage naturel.

C'est justement dans cette optique que nous proposons un système connexionniste pour l'analyse syntaxique des structures de la langue arabe en se basant sur un formalisme linguistique développé par les anciens grammairiens arabes à savoir le formalisme néo-khalilien.

2. La notion de structure dans les systèmes connexionnistes pour le traitement automatique de la langue naturelle

L'approche connexionniste offre des avantages attractifs tels que la capacité d'apprentissage, le traitement parallèle et la résistance au bruit, etc. Néanmoins, l'utilisation des réseaux de neurones pour les tâches cognitives a suscité de vives critiques quant à la faiblesse des techniques de représentations, d'autant plus que dans le cas du traitement automatique du langage naturel, le problème de la représentation est rendu difficile par le nombre et la complexité des objets à décrire.

Les différents systèmes proposés se basent sur le réseau simplement récurrent de Elman (1990). L'utilisation des réseaux simplement récurrents fournit un contexte ou un historique. En effet, une copie de l'activation des unités de la couche cachée est présentée en entrée avec les unités de la couche d'entrée au temps *t-1*. Ainsi, le réseau traite les entrées de l'instant *t* en tenant compte des entrées de l'instant *t-1*.

Fodor et Pylyshyn soulèvent le point concernant la faiblesse des techniques connexionnistes à représenter des traits essentiels du langage tels que la productivité (capacité d'exprimer la diversité de la connaissance du domaine avec un minimum de ressources), la systématicité (la capacité de manipuler des représentations équivalentes de façon identique) et la compositionnalité et qui sont des traits essentiels de tout modèle cognitif. Ils ont également soulevé le problème de représentation des structures complexes par les réseaux de neurones.

Pollack (1990) a développé la RAAM (Recursive Auto-Association Memory) afin de représenter les structures récursives. La RAAM a la capacité de développer automatiquement, les représentations distribuées récursives avec l'entraînement d'un ensemble fini d'états en utilisant l'algorithme de rétropropagation (Rumelhart, 1986). La RAAM est composée d'un compresseur et d'un décompresseur, entraînés simultanément. La tâche du compresseur est de codifier un ensemble de « patterns » de dimension fixe en un seul pattern de même dimension. Cette compression est appliquée de manière récursive de bas en haut. La tâche du reconstructeur est de décoder les patterns de haut en bas afin de restituer la forme originale de l'arbre.

Durant cette dernière décennie, plusieurs systèmes connexionnistes dédiés aux différentes tâches du traitement automatique de la langue naturelle, ont vu le jour : Miikkulainen (1991), McClelland (1989), Chan (1994), Lin (1995), Jain (1996).

Berg (1992) avec son système XERIC a combiné un réseau simplement récurrent, la théorie X-Bar et une RAAM permettant ainsi d'analyser des phrases quel que soit leur longueur et leur complexité.

3. Le système Neurokhall

Le système Neurokhall proposé est dédié à l'analyse des structures syntaxiques de la langue arabe. Il combine un réseau simplement récurrent de Elman, une RAAM et le schème générateur du formalisme linguistique néo-khalilien.

3.1. Concepts de base

- Le système conçu, traite les phrases diacritisées. Ainsi, nous associons à chaque mot du lexique, mis à part sa chaîne consonantique, la chaîne de signe de diacritisation. Ceci permet de lever certaines ambiguïtés lexicales.
- Le système Neurokhall est basé sur une représentation semi-distribuée où chaque neurone représente un trait spécifique. Un mot du lexique sera ainsi représenté par un ensemble de neurones, et chaque neurone participe à la codification de plusieurs mots. C'est une représentation semi-distribuée dans laquelle la valeur de chaque neurone est directement interprétable. Les représentations locales ne sont pas assez efficaces pour représenter les éléments du langage vu leur complexité et leur diversité.
- Une condition nécessaire que doit respecter tout système connexionniste pour le traitement automatique du langage naturel, est qu'il doit être en mesure de traiter les phrases complexes. Le système Neurokhall, grâce à l'uniformité du template de représentation des structures et grâce au mécanisme de la RAAM, est capable de traiter n'importe quelle structure de la langue arabe quel que soit son niveau de récursivité et le type d'enchâssement. Aucune restriction n'est imposée au préalable.
- Le système Neurokhall est capable de représenter la structure syntaxique d'une phrase sous la forme du schème générateur modifié sans pour autant lui fournir une description de la grammaire utilisée, lui laissant ainsi le soin d'acquérir les règles grammaticales au cours de la phase d'apprentissage. Ainsi, le système se compose d'un réseau simplement récurrent, d'une RAAM (Recursive Auto Association Memory) et le template modifié.
- Un réseau simplement récurrent, une RAAM, et le template modifié de la structure syntaxique de l'arabe sont combinés afin de fournir un système, n'ayant à priori aucune limite quant à la longueur de la phrase ou la profondeur des imbrications ou des enchâssements.
- L'apprentissage est supervisé.

3.2. Description du lexique

La plupart des théories linguistiques courantes tendent à attribuer une part importante au lexique pour des phénomènes considérés comme étant purement syntaxiques. Cette approche lexicalisée implique la gestion d'un lexique volumineux et complexe. Cependant, elle simplifie la tâche de l'analyseur syntaxique. L'approche utilisée pour re-

présenter les mots du lexique est une approche lexicaliste, c'est à dire qu'il y aura autant d'entrées lexicales que de dérivées d'un mot (exemple : *kataba* (كتب), *kātib* (كاتب), *maktūb* (مكتوب), *maktab* (مكتب).

Dans le cas du système Neurokhall, chaque mot du lexique est défini par un ensemble de traits distinctifs dont chacun représente une propriété syntaxique ou sémantique. Chaque item sera affecté d'un trait positif (représenté par [+ trait]) ou négatif (représenté par [-trait]) dont la catégorie sera soit de nature catégorielle soit booléenne.

Dans le cas où le trait a une valeur booléenne, nous lui assignerons un neurone pour le représenter ; il aura la valeur 0 ou 1 selon l'absence ou la présence du trait. Dans le cas ou le trait possède des valeurs numériques, nous lui assignerons le nombre de neurones nécessaires pour représenter l'ensemble des valeurs.

Ainsi un mot du lexique en mode de représentation connexionniste est un ensemble de neurones dont les valeurs sont soit à 1 (dans le cas où le trait serait présent) soit à 0 (dans le cas où le trait serait absent). Le nombre de neurones correspond au nombre total de traits pris en considération dans le système. Nous pouvons schématiser donc un mot du lexique comme étant un vecteur binaire de n cases (n correspondant au nombre de traits).

Les traits sont de type syntaxique et sémantique. Un trait élémentaire est le trait permettant de distinguer le type de *kalim* en arabe : le *ḥarf*, le nom et le verbe.

Pour les verbes, les traits syntaxiques spécifient essentiellement l'aspect (l'accompli, l'inaccompli et l'impératif), la flexion *raf'* (رفع), *naṣb* (نصب) ou *ǧazm* (جزم), la forme canonique d'un verbe de sa forme conjuguée, la voix du verbe (passive / active), le genre et le nombre, la sous-catégorisation spécifiant la relation qui existe entre un verbe, le régissant symbolisé par R et les termes régis symbolisés par T_i. La valeur de ce trait spécifie le nombre de termes régis (i =1, 2, 3, 4). Un trait est utilisé afin de spécifier les verbes exponentiels qui génèrent des constructions nominales comme le verbe transitif *kāna* et les éléments de sa classe (كان وأخواتها), les verbes de la classe de *ḥassiba* (حسب) (verbe bitransitif) ainsi que les verbes de la classe de *a'lama* (أعلم) (verbe triplement transitif).

Pour les noms, les traits considérés spécifient la catégorie (pronom, nom propre, nom commun, adverbe, adjectif), le type du pronom (personnel, démonstratif et relatif), *at- ta'rīf* (التعريف), *at- tanwīn* (التنوين), la désinence casuelle.

Pour la langue Arabe, *'inna* ainsi que tous les éléments de sa classe (إن وأخواتها) sont considérés comme des *ḥurūf*. Vu que les rôles de ces éléments sont différents des autres *ḥurūf*, il est impératif de les distinguer par un trait spécifique.

Le lexique est une composante centrale lors de l'analyse. Afin de lever certains cas d'ambiguïtés, nous proposons d'intégrer au lexique quelques traits sémantiques introduits par Chafe en 1970 tels que le trait animé, humain, concret, comptable et potent.

3.3. Description des structures syntaxiques

Le schème générateur RT_iD proposé dans la théorie néo-khalilienne est un modèle qui permet de couvrir toutes les structures syntaxiques de la langue arabe. L'élément central de ce schème est représenté par le régissant (*'āmil* /العامـــل). Au fait, le régissant joue un rôle prédominant pour l'ensemble de la structure. Il ne représente pas uniquement une position dans le schème mais cumule également le rôle de fonction car il conditionne également les termes régis (T_i) en déterminant leur nombre et le rôle de chacun d'eux.

L'utilisation du "template" RT_iD pour représenter la complexité et la diversité des structures de la langue arabe, est assez fructueuse car elle fournit une description économique représentant aussi bien le rôle syntaxique des différentes structures, que les fonctions de chaque élément de la structure.

Dans le cas des phrases simples, chaque position du template est occupée par une seule entité ou *kalima*.

Dans le cas des phrases complexes, les positions du schème générateur de la syntaxe peuvent être occupées soit par des entités simples ou des entités composées formées de lexies verbales dans le cas du régissant (R) ou par des lexies nominales dans le cas des termes régis (T_i).

Les deux niveaux d'analyse : celui de la tectonie et celui de la lexie, ne sont pas hiérarchiques mais s'imbriquent mutuellement l'un dans l'autre.

Le modèle linguistique néo-khalilien met en coopération cinq schèmes générateurs :

☞ Celui de la tectonie représenté par :

R	T_1	T_2	T_3	T_4	D

☞ Celui de la lexic nominale :

Préposition	Déterminant	Noyau	Désinence	Complément adnominal	Caractérisant
←	←	↔	→	→	→
2	1	0	1	2	3

☞ Celui de la lexie verbale de l'accompli :

Convertisseur	Exposant	Noyau	Pronom affixe
←	←	↔	→
2	1	0	1

☞ Celui de la lexie verbale de l'inaccompli :

Convertisseur	Exposant	Noyau	Désinence	Assertion énergique	Ajouts pronominaux
←	←	↔	→	→	→
2	1	0	1	2	3

☞ Celui de la lexie verbale de l'impératif :

Noyau		Assertion énergique	Ajouts pronominaux
← ↔	→	→	→
1 + 0	1	2	3

Afin de permettre l'intégration de ces schèmes, et pour pouvoir traiter les différents types d'imbrication que l'on pourrait rencontrer dans les phrases complexes, nous avons procédé à leur unification. L'idée principale est de mettre au point un template ou schème général rendant compte des différentes combinaisons possibles des structures syntaxiques que l'on pourrait rencontrer dans la langue arabe tout en respectant les règles générales de chaque schème ainsi que les fondements de base de la théorie néo-khalilienne.

Cette unification a été faite d'une manière progressive :

- **Simplification du schème générateur de la tectonie**

Hadj Salah (1979) a souligné que la formule $(R \rightarrow T_1), T_2$ (où la flèche qui relie T_2 au couple ordonné (R, T_1) indique la dépendance du binā') constitue au fait, un véritable schème générateur capable de caractériser tous les types de noyaux syntaxiques.

La productivité de ce template est assurée par le fait que chacune de ses positions peut se décomposer à son tour en R, T_1, T_2. Pour illustrer cela, considérons les deux séquences suivantes :

Cas où le régissant est complexe

	R			T_1	T_2						
	R	T_1	T_2								
#	A'lama	'Abdullahi	'Amran	Zaydan	qā'iman	##	ḫāfa	'Amrun	'an ya-ḥruǧa	Zaydun	ϕ	#

Cas de T_2 enchâssée

R	T_1	T_2		
		R	T_1	T_2
ḫāfa	'Amrun	'an ya-ḥruǧa	Zaydun	ϕ

A la structure globale du pivot, s'ajoute la position *hors binā'* (hors structure), celle du déterminant qui se décompose à son tour en R, T_1, T_2 dans le cas où il serait complexe.

Cas du déterminant composé

	R	T_1	T_2	D			
				R	T_1	T_2	
#	Sami'	tu	Zaydan	ya-qūlu	ϕ	ḍāka	#

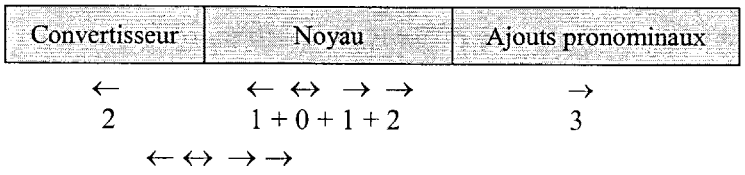
Il est important de noter que chaque position du template R, T_1, T_2 peut être occupée par la position vide symbolisée par ϕ.

A ce stade, nous considérons donc, que toute structure syntaxique se décompose en R, T_1, T_2, D.

- **Unification des trois lexies verbales**

Comme nous l'avons vu précédemment, la lexie verbale est gérée par trois schèmes générateurs (l'accompli, l'inaccompli et l'impératif). Les différentes positions des templates possèdent beaucoup de similarités. C'est pourquoi, par souci d'uniformité, nous avons superposé les trois schèmes générateurs de la lexie verbale, puis nous les avons unifiés, puis simplifiés.

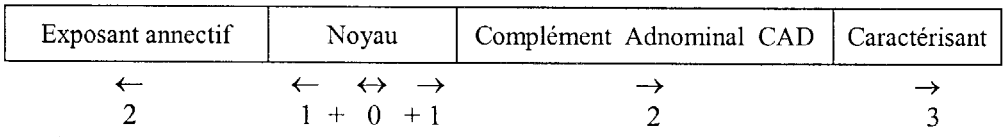
Ainsi le schème générateur de la lexie verbale obtenu est le suivant :



La fusion des positions 1, 0, 1, 2 est rendue possible car l'ensemble des informations contenues dans ces positions réside dans le lexique. En effet, la position de la flexion ainsi que celle de la marque ة sont exprimées dans le lexique sous forme de traits.

- **Simplification de la lexie nominale**

La lexie nominale a également été réduite à quatre positions comme suit :

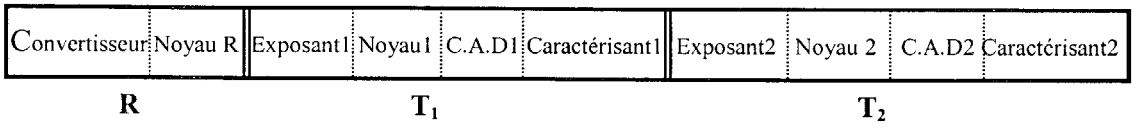


Les positions du *ta'rîf* et de la désinence casuelle, ont été intégrées dans la position du noyau, car ces informations sont exprimées dans le lexique sous forme de traits (voir trait TAN et trait CAS).

- **Intégration de la lexie verbale unifiée et de la lexie nominale dans le Template de la tectonie simplifié**

Après avoir unifié les trois lexies verbales et simplifié la lexie nominale, nous les intégrons dans le template général de la tectonie. Ainsi, le régissant est représenté par une lexie verbale et les termes régis sont des lexies nominales.

Voici d'après ce qui vient d'être dit, comment se présente le schème générateur qui sera utilisé dans le système Neurokhall :



3.4. Architecture du système Neurokhall

Après une étude détaillée sur les différents systèmes connexionnistes pour le traitement automatique du langage naturel, il en ressort que le modèle le plus adéquat est le réseau simplement récurrent (SRN) de Elman.

Ce réseau est composé de trois couches : une couche d'entrée qui reçoit les données du système, une couche cachée et une couche de sortie dont l'activation de ses unités sera considérée comme la réponse du réseau.

L'ensemble des neurones de la couche cachée n'a aucun lien avec l'utilisateur et agit par l'intermédiaire d'autres neurones. Un réseau disposant de neurones cachés est souvent plus puissant qu'un réseau qui n'en dispose pas.

La couche cachée à l'instant (*t-1*) est recopiée comme couche de contexte est présentée avec les unités de la couche d'entrée à l'instant *t*, ce qui fournit au réseau un contexte ou un regard en arrière. Cette spécificité est primordiale pour les systèmes de TALN car il faudrait préserver la nature séquentielle de l'analyse.

Néanmoins, cette structure de base n'est pas assez suffisante pour traiter les phrases complexes ou récursives pour lesquelles, il faudrait retenir l'information durant une longue période. D'ailleurs, la critique majeure des réseaux connexionnistes en matière de représentation est le problème de la compositionnalité.

Afin de pallier à cet handicap, nous proposons d'intégrer à l'architecture de base du système une RAAM afin de pouvoir traiter de manière dynamique les phrases complexes et ceci quel que soit le niveau d'imbrication.

A cet ensemble (SRN + RAAM) est associé un template compact général afin de spécifier les sorties du système c'est à dire les structures syntaxiques (sous forme de template modifié) de la phrase présentée en entrée sous forme de traits syntaxiques.

Nous allons décrire dans ce qui suit les différents composants de l'architecture de Neurokhall.

Couche d'entrée

Afin de pouvoir communiquer les informations au réseau et d'en interpréter les réponses, un système de codage doit être mis en place. Celui-ci doit être judicieusement choisi pour éviter de tomber dans les excès. Soit trop d'informations sont communiquées au réseau engendrant ainsi beaucoup de bruits parasites, soit trop peu d'informations sont communiquées, alors le réseau ne serait pas en mesure de mener à bien la tâche qui lui a été attribuée.

La représentation locale ne semble pas très appropriée comme mode pour représenter les structures linguistiques vu les différents inconvénients qu'elles présentent. C'est pour cela, que nous avons opté dans le système Neurokhall pour une représentation semi distribuée dans laquelle, chaque mot est représenté par un ensemble de traits et chaque trait contribue à la représentation de plusieurs mots.

La couche d'entrée du système est composée par la description lexicale du mot ainsi que par les unités de contexte :

Codification du mot

L'entrée lexicale d'un mot est codifiée par 30 unités. Dans le cas où le trait est présent, l'unité correspondante sera activée à 1, et si le trait est absent, la valeur de l'unité sera à 0.

A cet ensemble d'unités, une unité supplémentaire, l'identificateur (ID), est ajoutée à l'entrée lexicale, mis à part sa description syntaxico-sémantique. Ainsi deux mots appartenant à la même catégorie seront différenciés uniquement par l'unité ID.

Unités de contexte

Vu la récurrence de l'architecture du réseau simplement récurrent, les unités de contexte à l'instant (*t-1*) sont injectées en entrée comme unités de contexte à l'instant *t*.

Initialement ces unités de la couche de contexte, dont le nombre est égale à celui de la couche cachée, sont initialisées à une valeur null prédéterminée par exemple 0.5.

Couche de sortie

La couche de sortie représente le template modifié décrit précédemment. Il est composé de 10 positions. Chaque position peut être occupée par une entité conforme à son rôle syntaxique ou vide. La position vide (ϕ) correspond à la valeur null. Il est important de signaler que l'un des trois noyaux (NoyauR, Noyau1, Noyau2) doit être différent de null, quelque soit le niveau d'analyse.

Un bit supplémentaire est ajouté au template modifié, afin de pouvoir distinguer le niveau d'enchâssement : la valeur 0 correspond à une structure simple et la valeur 1 correspond à une structure enchâssée.

Couche cachée

Dans le système Neurokhall, nous avons intégré une RAAM au réseau simplement récurrent afin de représenter les structures récursives. La RAAM est composée de : un encodeur, une couche cachée et un décodeur.

L'architecture générale d'une RAAM est :{2m, m, 2m}. Nous pouvons rappeler brièvement que le principe général de la RAAM était de pouvoir représenter les arbres binaires dans lesquelles un arbre pouvait contenir des sous arbres (notion de récursivité). Les m unités cachées servent à codifier les informations nécessaires pour recréer l'activation des 2m unités de sorties. Les valeurs des unités cachées seront présentées comme des non terminaux en entrée afin de pouvoir représenter des arbres binaires plus complexes.

Ce processus sera inversé d'une manière récursive, dans la partie du décodage, jusqu'à ce que les unités de la couche de sortie contiennent des terminaux.

Ce fonctionnement de la RAAM est universel, quelle que soit son application. Il en est de même pour le système Neurokhall.

3.5. Structure de base du réseau

La phrase à analyser est présentée au réseau mot par mot. En entrée, le système reçoit la description lexicale du mot courant associée aux valeurs de l'activation des unités de la couche cachée au pas précédent. Initialement, les unités de la couche de contexte sont initialisées avec une valeur "null" (exemple 0.5).

A la présentation du dernier mot de la phrase, et aprés l'entraînement en avant du réseau, les unités de la couche cachée auront développé la représentation interne de la phrase tout entière.

La structure de la phrase peut être montrée explicitement en utilisant la partie inférieure du réseau, c'est à dire la couche cachée, le décodeur et la couche de sortie. Ces trois éléments représentent la deuxième composante de la RAAM classique (la partie du décodeur). Celle-ci est capable de représenter toutes les phrases allant de la phrase simple jusqu'aux phrases les plus complexes.

Le décodage d'une phrase simple en utilisant le template de la syntaxe modifié, permet d'identifier distinctement le régissant (deux premières positions), le premier terme régi (les quatre positions suivantes) et le deuxième terme régi (les quatre dernières positions).

Pour les phrases complexes, chaque composant de base, c'est à dire le NoyauR, le noyau1 et le noyau2, peut être lui-même composé.

La puissance de ce modèle réside dans l'intensité des relations qui existent entre le template modifié de la syntaxe et les entrées lexicales des mots.

Un point important à souligner est que la taille du réseau est fixe. Chaque couche est définie par un nombre fixe d'unités et chaque mot du lexique est décrit par un vecteur de longueur fixe. La dynamique du système est assurée par les possibilités qu'offre le fonctionnement de la RAAM.

L'architecture de base du système Neurokhall peut être représentée par le schéma suivant :

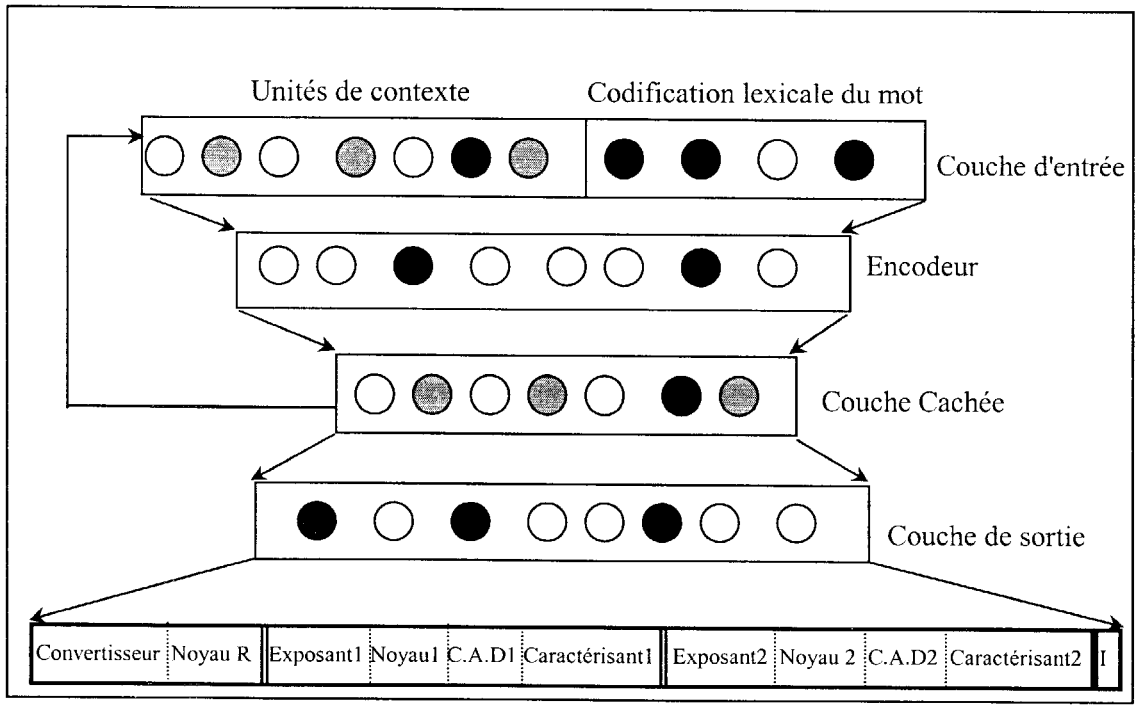


***Figure 1** : Architecture du système Neurokhall*

Le template modifié a été augmenté d'une position binaire (représentée par I pour indicateur) afin d'indiquer si le template correspond à une séquence simple (la valeur 0 sera attribuée à cette position) ou à une séquence composée (la valeur 1 sera attribuée à cette position).

3.6. Entraînement du réseau

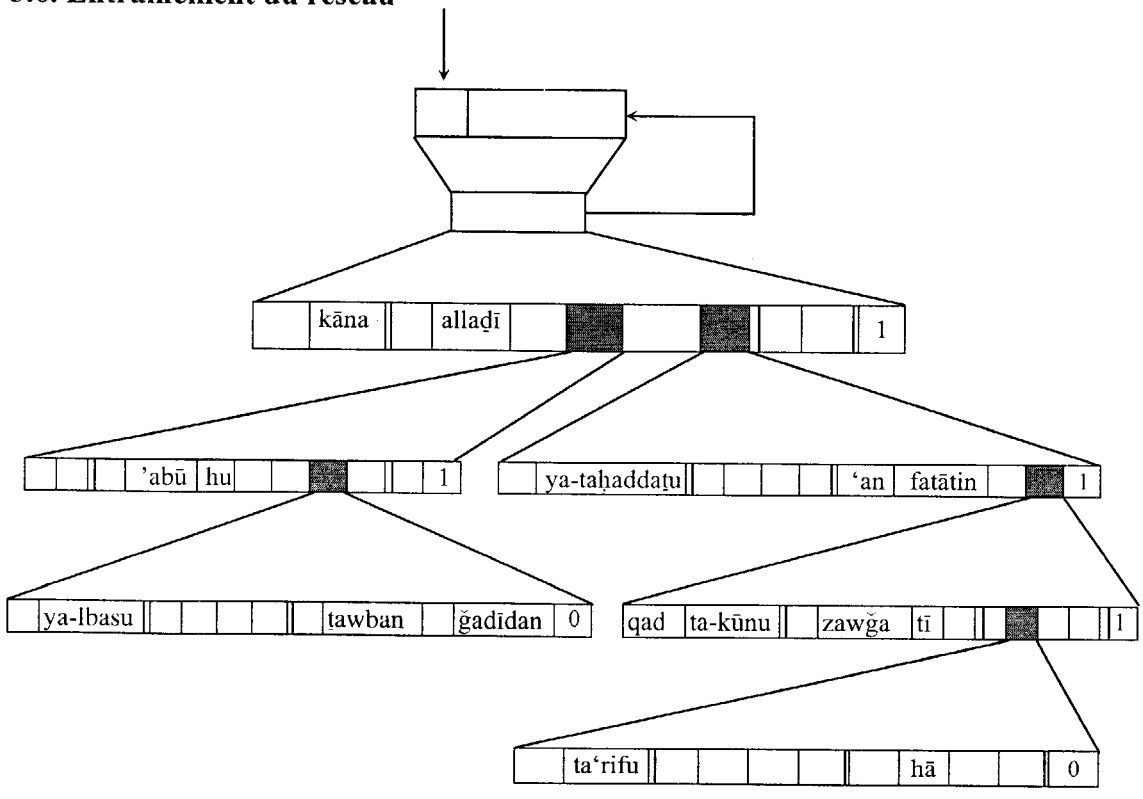


***Figure 2** : Déroulement du réseau pour la phrase :*
" كان الذي أبوه يلبس ثوبا جديدا يتحدث عن فتاة قد تكون زوجتي تعرفها"

L'un des principaux objectifs de ce travail est de pouvoir traiter toutes les séquences de phrases possibles, sans aucune limite quant à leur niveau d'enchâssement. Nous ne connaissons pas à priori, la profondeur de la phrase à traiter. Afin de contourner ce problème, le réseau utilise, lors de la phase d'entraînement, des copies virtuelles de la portion du décodeur de l'architecture de base. L'utilisation des copies multiples pour émuler le réseau permet d'obtenir dans la couche de sortie, soit la codification du mot, soit la valeur « null ».

Les positions non étiquetées dans la couche de sortie correspondent à des positions vides.

4. Modélisation de l'apprentissage

L'apprentissage dans le système Neurokhal est supervisé. Généralement, l'apprentissage se fait sur une période relativement longue pendant laquelle les patrons d'entrée peuvert être présentés plusieurs fois au réseau. Cet apprentissage comprend quatre étapes de calcul :

- Initialisation des poids synaptiques à de petites valeurs aléatoires ;
- présentation du patron d'entrée et propagation de l'activation ;
- calcul de l'erreur ;

- calcul du vecteur de correction : à partir des valeurs d'erreur, nous déterminons quelle correction sera apportée aux poids du réseau.

Concernant la stratégie d'apprentissage, Elman dans ses articles (1990) (1991) souligne l'importance de commencer petit « *The importance of starting small* » lors de la phase d'apprentissage et ceci par analogie à l'apprentissage de l'être humain.

Nous pensons adopter la stratégie présentée par Elman, qui consiste à faire un apprentissage incrémentiel à plusieurs phases où les phrases traitées sont de complexité croissante. Les phases d'apprentissage sont les suivantes :

Phase 1 : consistè à entraîner le réseau sur un corpus composé exclusivement de phrases simples en ne tenant pas compte des phrases enchâssées.

Phase 2 : Le réseau sera ensuite exposé à un second corpus constitué de 25 % de phrases complexes et de 75% de phrases simples.

Phase 3 : Le troisième corpus sera constitué de 50% de phrases complexes et de 50% de phrases simples.

Phase 4 : Le dernier corpus sera constitué de 75% de phrases composées et de 25% de phrases simples.

Le choix de la base d'apprentissage est crucial. La RAAM doit être entraînée à tous les niveaux et à tous les cas de figure de la décomposition. Chaque position du template modifié doit faire l'objet d'un apprentissage.

Il est important de signaler à ce niveau, que le réseau de neurones développe, dans sa couche cachée, des représentations distribuées identiques pour des entités ayant des descriptions similaires.

5. Mise en œuvre du réseau Neurokhall

D'une manière générale, les performances des réseaux dépendent étroitement des paramètres servant à le définir. Bien qu'il n'existe pas de lois pour gérer les différents paramètres, il existe cependant des techniques permettant de mieux cerner les domaines de définition des paramètres et ceci en fonction du domaine traité.

Plusieurs types de paramètres sont à considérer :

5.1. Paramètres associés à l'architecture

- La taille de la couche cachée

Le système Neurokhall est basé sur une architecture fixée lors de la conception et non pas sur une architecture évolutive. La détermination du nombre de neurones de la couche cachée est une phase importante lors de la conception du réseau. En effet, les neurones de la couche d'entrée représentant les données du système, sont fixés et les neurones de sortie symbolisant le résultat sont connus également. Pour la couche cachée, chaque neurone supplémentaire permet de prendre en compte des profils spécifiques des neurones d'entrée. Un nombre important de neurones permettra de mieux coller aux données présentées mais diminue la capacité de généralisation du réseau.

Il n'existe à ce jour aucune méthode qui détermine le nombre de neurones adéquat. Cependant, plusieurs chercheurs ont essayé de tester quelques règles empiriques :

- Mikkulainen (1991) affirme que les meilleurs résultats sont obtenus lorsque le nombre de neurones de la couche cachée est égal approximativement à la moitié du nombre de neurones de la couche d'entrée.

- Wierenga et Kluytmans (1994) proposent que la taille de la couche cachée soit égale à celle de la couche d'entrée.

- Venugopal et Baets (1994) proposent que la taille de la couche cachée soit égale à 75% de la taille de la couche d'entrée.

- Shepard (1990) propose que la taille de la couche cachée soit égale à la racine carrée du produit du nombre de neurones dans la couche d'entrée et de sortie.

Notons que le dernier choix réduit le nombre de degrés de libertés laissé au réseau, et donc réduit la capacité d'adaptation sur l'échantillon d'apprentissage, au profit d'une plus grande stabilité.

- Le domaine d'initialisation des poids synaptiques

Un réseau reçoit, lors de sa création, des poids synaptiques habituellement choisis d'une manière aléatoire dans un intervalle déterminé.

5.2. Paramètres associés à l'apprentissage

En règle générale, l'algorithme de rétropropagation est la méthode la plus utilisée lors de la phase d'apprentissage. Plusieurs variantes de cette méthode sont apparues, mais le principe reste le même.

La vitesse d'apprentissage η figure parmi les paramètres d'apprentissage. Il est de pratique courante de choisir des petites valeurs pour η.

5.3. Paramètres associés à l'entraînement

La procédure d'entraînement est également soumise à des paramètres.

La taille du corpus d'entraînement est difficilement estimée. Un corpus relativement grand reflète plus correctement le problème, mais il n'est pas toujours nécessaire de disposer d'un corpus énorme pour garantir un bon apprentissage.

6. Conclusion

L'architecture du système Neurokhall proposée peut être directement implémentée afin de pouvoir en estimer les performances de généralisation.

Confronter des connaissances théoriques aux résultats pratiques semble le meilleur moyen de comprendre les divers phénomènes qui accompagnent l'apprentissage.

Le cadre de recherche de ce travail étant relativement large, de nombreux axes de recherches peuvent s'inscrire en complémentarité à savoir,

1. Par analogie au format X-bar de Chomsky, il serait intéressant d'unifier les deux templates proposés, celui de la lexie nominale et celui de la lexie verbale, afin de former un seul. Ce dernier, sera également composé à son tour de quatre positions qu'il restera à nommer. C'est la nature du noyau (verbal ou nominal) qui déterminera le rôle des trois autres positions.

2. Une autre perspective qui en ressort, est d'envisager l'intégration du système Neurokhall dans une architecture connexionniste globale dans laquelle chaque module sera constitué d'un réseau de neurones. Ainsi, dans le cadre du développement du projet

global, élaboré au sein du Groupe de Recherche en Intelligence Artificielle (G.R.I.A) à l'université de Annaba, nous envisageons de fusionner le système Neurokhall avec le système dédié à la génération d'une représentation interne du sens d'une phrase basée sur les cas sémantiques (Meftouh, 2000), en considérant les sorties du système Neurokhall comme entrées du second système.

Néanmoins, la question qui s'impose à ce niveau est de savoir comment faire communiquer les deux systèmes. Cette question est d'ailleurs posée à grande échelle.

3. Une autre approche consisterait à concevoir des systèmes hybrides dans lesquels l'approche symbolique et l'approche connexionniste seront intégrées. En effet, le langage est aussi bien symbolique que "sub-symbolique". Il est symbolique car il est constitué de symboles. Il est sub-symbolique car la notion d'ambiguïté, d'incertitude et de flou y sont également intégrées.

Si un système parvient à capturer toute la puissance de cette diversité, il sera alors capable de traiter le langage naturel à grande échelle.

BIBLIOGRAPHIE

Berg George, *Learning Recursive Phrase Structure : Combining Three Strengths of PDP and X - BAR Syntax*, IJCAI91 Workshop on Natural Language Processing in Sydney Australia,1991.

_____, *A Connectionist Parser with Recursive Sentence Structure and Lexical Disambiguation*, AAAI – 92 Americain Association for Artificial Intelligence, pp. 32-37, 1992.

Blair Alan D., *Scalinhg-up RAAMs*, Département informatique, université de Bandeis, janvier 1997.

Chafe Wallace, *Meaning and the Structure of the Language*, University Press Chicago, 1970.

Chan Samuel W.K. and Franklin James, *A Neural Network Model for Acquisition of Semantic Structures*, International Symposium on Speech Image Processing and Neural Networks Hong Kong, pp. 221-224, 1994.

McClelland J.L., Mark st Jhon and Roman Taraban, « Sentence Comprehension : a Parallel Distributed Processing Approach », Language and Cognitive Processes, pp. 287- 335, 1989.

Elman Jeffrey L., « Finding Structure in Time », Cognitive Science, 14, pp. 179-211, 1990.

_____, *Distributed Representations, Simple Recurrent Networks, and Grammatical Structure*, Département des sciences cognitives et de linguistique, 1991.

Fodor.J and Pylyshyn.Z, « Connectionism and Cognitive Architecture : A Critical Analysis », in Connections and symbols, pp. 3-71, Cambridge MIT Press, 1988.

Hadj Salah Abderrahmane, *Linguistique arabe et linguistique générale : Essai de méthodologie et d'épistémologie du 'ilm al-'Arabiyya*, Thèse de doctorat, Paris Sorbonne, Vol. 2, 1979.

Jain Jianchang Anil K. Mao K. Mohiuddin, *Artificial Neural Networks : A Tutorial, IEEE Computer Special Issue on Neural Computing*, mars 1996.

Koong H.C. Lin and Tung-Bo Chen and Von-Wun Soo, « Neural Network Learning and Encoding of Thematic Role Assignments in Parsing of Simple Chinese Sentences », Journal of Information Science and Engineering, vol.11, n° 1, pp.109-126,1995.

Meftouh Karima, *Une approche connexionniste pour la génération d'une représentation interne du sens d'une phrase basée sur les cas sémantiques appliquée à la langue arabe*, mémoire de Magistère, université de Annaba, 2000.

Miikkulainen Risto Michael G. Dyer, *Natural Language Processing with Modular PDP Network and Distributed Lexicon*, cognitive science, 1991.

Pollack Jordan B., *Recursive Distributed Representation*, Laboratoire de recherche en intelligence artificielle, université de l'état de l'Ohio, 1990.

Rumelhart De. and Hinton Wiliam, « Learning Internal Representations by Error Propagation », Explorations in the microstructure of cognition, Vol 1, pp. 318-362, Cambridge MIT Press, 1986.

Shephard GM, *Synaptic Organisation of the Brain Network*, Oxford University Press, 1990.