Extraction automatique du sens d'une phrase en langue Française par une approche neuronale

Med Tayeb Laskri, Karima Meftouh

Groupe de Recherche en Intelligence Artificielle (GRIA/LRI) – Département d'informatique – Université Badji Mokhtar Annaba – BP 12 – Annaba 23000 – Algérie

Abstract

In this paper, we propose a connectionist model for the generation of an internal representation of the sense of a sentence in french language, based on the semantic cases. We use the back propagation algorithm in a Sample Recurrent Network (SRN). The sentence is analysed word by word. Every word is introduced to the network under semantic features. The task of the network consists of reading the sentence, and deciding the suitable semantic role for each word. The network successfully learned the case role assignment task. It has been experimented on several corpora and the got results are satisfactory.
The network has also been tested on a composed corpus of different sizes of sentences and the rate of generalisation approaches of the 92% was obtained.

Résumé

Dans cet article, nous proposons un modèle connexionniste pour la génération de la représentation du sens d'une phrase en langue française, basée sur les cas sémantiques, et nous utilisons pour cela l'algorithme de rétro propagation traditionnel dans un réseau simplement récurrent. La phrase est analysée mot par mot, chaque mot est introduit au réseau sous forme de traits, et la fonction du réseau est de déterminer le rôle sémantique approprié. L'apprentissage a été effectué d'une manière incrémentale : le réseau a été expérimenté sur plusieurs corpus et les résultats obtenus sont assez satisfaisants.
Le réseau a été également testé sur un corpus composé de phrases de différentes tailles et le taux de généralisation approche les 92%.

Mots-clés : Traitement automatique du langage naturel, Réseaux de neurones, Réseau simplement récurrent, Représentation distribuée, Représentation locale, Algorithme de rétro propagation, Cas sémantique.

1. Introduction

Depuis plusieurs décennies, l'étude sur le Traitement Automatique du Langage Naturel (TALN) a été dominée par l'approche symbolique. Cependant, durant ces dernières années un intérêt croissant est apparu quant à l'applicabilité des techniques connexionnistes pour le TALN et ceci vu les propriétés attractives des réseaux de neurones telles que leur résistance aux pannes et leurs capacités d'apprentissage et de généralisation. Plusieurs modèles connexionnistes ont été développés pour la résolution des anaphores, transformation active-passive, apprentissage de la syntaxe, l'interprétation sémantique et la traduction (Archambault 1995 ; Courant 1999, Elman 1993) . Parmi ces modèles, nous citons celui de Macclelland et Kawamoto (1986) identifiant l'assignation de rôles sémantiques comme une bonne tâche pour les réseaux de neurones (Mcclelland and St Jhon 1990).

La représentation basée sur les cas sémantiques est une technique de l'IA permettant la description du sens d'une phrase. L'idée est basée sur la théorie de Fillmore de 1968. Chaque acte est décrit par le verbe principal et un ensemble de cas sémantiques tels que : ***agent, instrument, patient.*** La

tâche consiste donc à décider quel constituant de la phrase joue quel rôle sémantique dans cette phrase.
Dans cet article, nous proposons un modèle connexionniste pour la génération de la représentation du sens d'une phrase en langue française, basée sur les cas sémantiques. Nous utilisons l'algorithme de rétro propagation classique dans un réseau simplement récurrent. Dans cette approche, les mots sont codés selon un ensemble de traits sémantiques et syntaxiques dont la définition sera donnée un peu plus loin. Les réseaux simplement récurrents (SRN) ont été énormément utilisés dans le traitement du langage naturel ; Surtout dans les travaux (Berg 1992 ; Elman 1990) qui ont démontré la puissance des SRN à apprendre des structures syntaxiques avec une performance remarquablement similaire à celle des humains.

2. Les cas sémantiques

Dans un article célèbre (« The case for the case " 1968), Fillmore soutient que l'on peut identifier un ensemble de cas sémantiques. Ces cas permettent de mettre en évidence, à la manière des cas syntaxiques, les relations de sens qui existent dans une phrase simple, entre les groupes nominaux et le verbe considéré comme composant central. Les cas forment idéalement une liste unique, finie, petite en nombre, universelle et valide dans toutes les langues : cette liste permet de classer les verbes suivant leur structure argumentale. Le verbe est donc décrit en termes de cas sémantiques.

Le choix des cas influe énormément sur l'interprétation sémantique. Se limiter à une certaine ossature peut provoquer certaines ambiguïtés pour ce qui est des éléments restants de la phrase.

Exemple :

Pour le verbe VOLER, il faut indiquer les trois ossatures qui lui sont reliées pour distinguer le vol d'un oiseau, du vol d'un avion et de l'action de dérober (Sabah 1988 ; Sabah 1989). Aussi si l'on arrive à décider du nombre de cas possibles à utiliser, il reste toujours le problème de décision du rôle joué par un élément donné vis à vis du verbe.

Donc, pour pouvoir déterminer un ensemble de cas qui répondraient à toutes les exigences, il faut se préoccuper de l'étude d'un corpus pour tirer les cas sémantiques nécessaires.

Exemple :

Khaled ira au stade demain pour regarder le match.

La représentation sera comme suit :

Khaled	: Agent
ira	: Action
demain	: Temps
au stade	: Destination
pour regarder	: But
le match	: Objet

L'assignation de cas sémantiques semble être une bonne approche particulièrement bien adaptée au réseau de neurones du fait que les cas peuvent être représentés convenablement par des unités :

- Si l'on adopte une représentation locale, on associera un neurone à chaque cas sémantique.
- Dans le cas d'une représentation distribuée, un ensemble d'unités représentera un cas sémantique.

3. Description générale du système

Traduire une phrase en une représentation interne est une tâche fondamentale dans le TALN. Dans ce qui suit nous illustrons les représentations connexionnistes dans cette tâche. Le système proposé est un modèle connexionniste qui ayant certaines informations en entrée permet la construction d'une représentation interne du sens d'une phrase en langue française. La représentation interne est basée sur la théorie des cas sémantiques de Fillmore. La structure est basée sur les réseaux simplement récurrents de ELMAN (voir figure 1). Une copie de la couche cachée au temps (t) est sauvegardée et utilisée avec l'entrée actuelle au temps (t+1) comme entrée de la couche cachée. La tâche consiste à lire la phrase mot par mot et le réseau décide si le mot en entrée joue le rôle d'agent, d'action, de patient, d'instrument…
Par exemple dans la phrase : *le ballon a cassé la fenêtre*

Le sujet *le ballon* joue le rôle d'instrument de l'action *a cassé*.
La fenêtre joue le rôle de patient alors que l'agent de l'action est inconnu.

L'assignation des rôles thématiques dépend du contexte et aussi des propriétés sémantiques du mot. Dans la phrase :

L'enfant a mangé un gâteau aux pommes

Modifie le patient un gâteau

Alors que dans la phrase :

L'enfant a mangé un gâteau avec une fourchette

Joue le rôle de l'instrument de l'action a mangé

Dans d'autres cas l'assignation des rôles reste ambiguë.

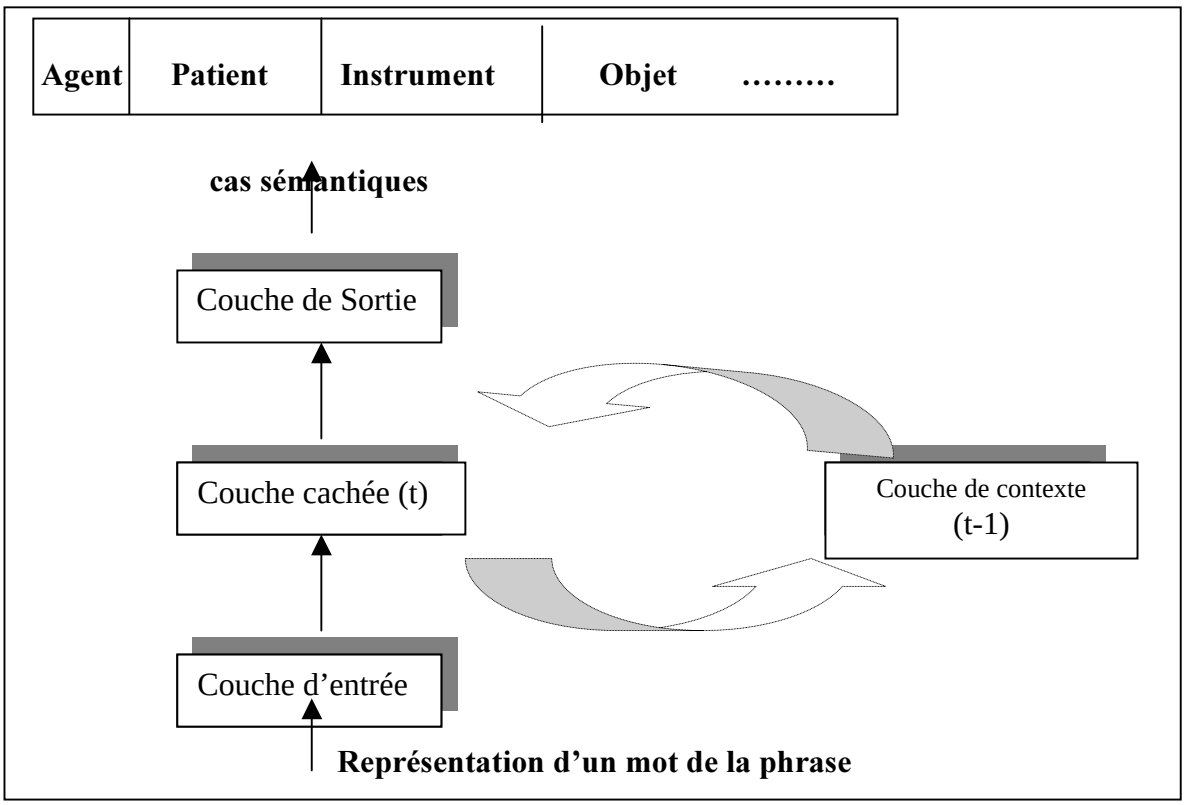

Figure 1 . Allure générale du système

4. L'interface du système

La tâche du système est de traiter une séquence de constituants qui représentent une phrase particulière et de déterminer pour chacun de ces composants le rôle sémantique approprié.

4.1. La couche de sortie

Cette couche représente l'ensemble des cas sémantiques utilisés. Chaque cas sémantique est représenté par une unité (représentation locale) qui sera activée si elle correspond au cas sémantique du mot en entrée du réseau. La couche de sortie compte 14 unités chacune correspondant à un cas sémantique. Selon l'état du neurone (actif ou inactif) on peut décider du rôle sémantique du mot. Ce tableau présente l'indice du neurone et le cas associé.

Neuron	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rôle	Action	Agent	Objet	Instrument	Manière	Patient	Lieu	Temps	Etat	Source	destination	Fournisseur	Bénéficiaire	But

Tableau 1. indices des neurones représentant les cas sémantiques (couche de sortie)

4.2. La couche d'entrée

Une représentation locale semble être inappropriée pour cette couche. En effet, choisir d'attribuer un neurone par mot entraînera une réduction du lexique utilisé. Dans le cas d'un lexique volumineux, la taille du réseau constituera un obstacle pour son implémentation (espace mémoire trop grand). Aussi, nous n'avons pas voulu fixer d'avance la taille de la phrase (nombre de mots qui la composent) ce qui explique notre choix de présenter la phrase en entrée composant par composant plutôt que de présenter la phrase entière. Le nombre de mots, variant d'une phrase à une autre, fait de l'estimation du nombre d'unités en entrée du réseau une tâche très difficile : Concevoir un réseau qui accepte uniquement des phrases de moins de 4 mots, par exemple, limite les performances du système ou bien mettre en œuvre un réseau qui accepte des phrases qui comptent au maximum 15 mots par exemple, semble plus englobant mais augmente le nombre d'unités en entrée inutilement.

Le nombre de mots utilisés dans une phrase est donc dynamique.

Nous adoptons, au niveau de cette couche, une représentation distribuée établie manuellement. Cette tâche est très difficile car il faudrait prendre en considération plusieurs sortes d'informations : syntaxiques, sémantiques, thématiques, contextuelles … Avant de déterminer le nombre de neurones en entrée, il faut étudier séparément les différentes catégories des mots et leurs représentations.

4.2.1. Représentation des verbes

Pour la représentation des verbes, nous utilisons la classification de Shanck. Le verbe sera donc représenté par sa primitive. Dix (10) primitives définies par SHANCK ont été considérées et chacune d'elles est représentée par un neurone. Le neurone sera actif si le verbe appartient à cette primitive. Le tableau ci-dessous présente l'indice du neurone et la primitive associée.

Indice	1	2	3	4	5	6	7	8	9	10
Primitive	Atrans	Ptrans	Mtrans	Mbuild	Attend	Speak	Propel	Ingest	Expel	State

Tableau 2. indices des neurones associés aux primitives des verbes

Un autre point très important concernant le verbe est son «type». Le verbe peut être en «voix active», qui peut avoir un **agent** ou en « voix passive » , où l'**agent** est absent dans la phrase.

Exemple : *L'enseignant interroge l'étudiant* (voix active agent : *l'enseignant*)
L'étudiant est interrogé (voix passive objet : *l'étudiant*)

Cette information est très importante pour la détermination des rôles sémantiques. A cet effet un neurone est ajouté pour représenter la voix utilisée (voix passive ou active). Si le verbe est en voix passive le neurone est actif, sinon il est inactif.

4.2.2. Représentation des noms

Pour la représentation des noms, nous avons utilisé le modèle de Chafe pour la classification des noms. Chafe a définit une liste de traits sémantiques (marqueurs) qui représentent les propriétés du nom. D'après Chafe, un nom ne peut être qu'un : Animé, Humain, Féminin, unique, concret, comptable ou potent.

Exemple :
Un enfant [(+)Animé, (+)Humain, (-)Féminin, (-) unique, (+)concret, (+)comptable, (-) potent].
La chaleur [(-)Animé, (-)Humain, (+)Féminin, (+) unique, (-)concret , (-) comptable, (+) potent].

Nous avons constaté que ces traits n'étaient pas suffisants pour représenter tous les noms et par fois donnent la même représentation pour des mots différents qui ne peuvent pas jouer le même rôle dans le sens de la phrase.

Exemple :
Marseille [(-)Animé,(-) Humain,(+)Féminin,(+) unique, (+)concret,(-) comptable,(-) potent].
Lait [(-)Animé,(-) Humain,(-)Féminin,(+) unique, (+)concret,(-) comptable,(-) potent].
Pour résoudre ce problème nous avons augmenté la liste de Chafe de deux traits supplémentaires:

Consommable : ce trait caractérise les objets ou les produits qui peuvent être consommés ou rejetés par un être vivant.

Dimension : ce trait est réservé pour les noms indiquant une dimension spatiale ou temporelle ou un lieu.

Exemple : *Ville, matin …*

Chaque trait lui est associé un neurone. Si le nom vérifie le trait, le neurone correspondant est activé, sinon il est inactif. De la même manière, voici un tableau qui présente l'indice du neurone et le trait associé.

Neurone	12	13	14	15	16	17	18	19	20
Trait	Animé	Humain	Féminin	Unique	Concret	Comptable	Potent	Consommable	dimension

Tableau 3. indices des neurones utilisés dans la représentation des noms

4.2.3. Les particules

Les particules en langue arabe n'ont pas un rôle sémantique dans la phrase, mais peuvent aider à déterminer certains cas et permettent ainsi de lever l'ambiguïté qui réside.

Exemple : *L'enfant mange le couscous. (patient)*
L'enfant mange avec la cueillere (instrument).

Ces deux phrases ont presque la même représentation si l'on ignore le rôle de la particule « avec ». Le réseau décide alors que «la cuillère » est patient.

Le mur est tombé : objet.
Il est tombé devant le mur : lieu.

De la même manière « devant » détourne le sens de « le mur » d'un objet à un lieu.

La gazelle court : agent.
Il court comme la gazelle : manière.

Pour aider le réseau à dégager les rôles sémantiques des mots, nous avons ajouté un ensemble de neurones où chacun représente une particule ou un groupe de particules. Le neurone, associé à la particule, est actif en même temps que les neurones qui représentent le mot qui la suit. L'indice du neurone et la particule associée sont résumés dans le tableau suivant :

Neurone	21	22	23	24	25	26	27	28	29	30
Les particules	Devant Derrière …		pour	dans	avec	de	Sur	au	par	avant apres ...

Tableau 4. Indices de neurones associés aux particules

La couche d'entrée compte ainsi 30 neurones.

6. Environnement d'entraînement

Nous avons présenté un réseau simplement récurrent qui apprend l'affectation des rôles sémantiques. L'apprentissage est supervisé vu que nous pouvons déterminer les sorties désirées (Jodouin 1994). L'apprentissage se fait de manière incrémentale (Elman 1993). Nous présentons d'abord au réseau des phrases simples telles que :

L'enfant joue. *L'enfant écrit le cours.*

Ensuite des phrases plus difficiles telles que :

L'enfant joue avec le ballon. *L'enfant écrit le cours dans le cahier.*

L'apprentissage est poursuivi jusqu'à ce que le réseau ait appris le corpus d'apprentissage à un certain seuil ou bien jusqu'à ce que ce taux ne progresse plus. Nous testons alors le réseau sur un corpus test composé de phrases ne faisant pas parti du corpus d'apprentissage a fin de mesurer sa capacité à généraliser sa connaissance. En phase de généralisation, lorsque le réseau rencontre une phrase non familière, il doit procéder à son identification avec une représentation qu'il a déjà rencontrée. En d'autres termes, le système traite le mot X dans la situation S car il sait comment traiter le mot Y dans la situation S et les mots X et Y sont utilisés de manières similaires dans plusieurs situations (Miikkulainen 1997).

7. Le traitement

Chacune des unités de la couche d'entrée et de la couche de contexte est connectée avec toutes les unités de la couche cachée, et chaque unité de la couche cachée est connectée avec toutes les unités de la couche de sortie.

Le traitement dans un réseau simplement récurrent s'effectue comme suit. Au début de chaque phrase, les unités de contexte sont toutes initialisées à zéro et les unités représentant le premier composant sont activées. L'activation est propagée des unités d'entrée et des unités de contexte vers la couche cachée et de la couche cachée vers la couche de sortie qui indiquera en conséquence le cas sémantique du composant courant de la phrase. La valeur de l'activation de la couche cachée est ensuite recopiée, sans autre traitement, dans la couche de contexte et les unités représentant le second composant sont alors activées. Ce processus de traitement se répète jusqu'à la fin de la phrase. A ce niveau, il y'a réinitialisation de la couche de contexte avant le traitement d'une prochaine phrase.

8. Expérimentation et résultats

8.1. Conditions d'expérimentation

- Les paramètres du réseau :

Les réseaux de neurones sont caractérisés par un certain nombre de paramètres dont nous avons choisi les valeurs suivantes :

- Une entrée activée aura une valeur d'activation de 1 et une inactivée de 0.
- Une sortie sera considérée comme activée lorsque sa valeur est supérieure à 0.8 et inactivée lorsqu'elle sera inférieure à 0.2.
- Le coefficient d'apprentissage (pas) est fixé à 0.6. un pas d'apprentissage trop faible peut empêcher le réseau de sortir d'un minimum local.
- Intervalle d'initialisation des poids : après quelques essais, l'intervalle [-0.25 , +0.25]qui donnait de meilleurs résultats en temps d'apprentissage a été fixé.
- Arrêt de l'apprentissage : l'apprentissage s'arrête lorsque l'erreur RMS par neurone de sortie sur le corpus d'apprentissage est inférieure à 0.08
- Nombre de neurones de la couche cachée est fixé à 30 neurones. La taille de la couche cachée montre la capacité du réseau à généraliser : plus il y'a d'unités et moins le réseau généralise. Dans le cas inverse, le réseau n'apprendra rien

- Les caractéristiques du corpus :

Nous avons choisi d'utiliser, pour les échantillons d'apprentissage, des phrases provenant des élèves d'écoles primaires. Nous avons ajouté à ces échantillons un ensemble de phrases simples, afin de permettre au réseau un apprentissage progressif. Le corpus d'apprentissage comporte une centaine de phrases de tailles différentes. Les phrases sont différentes, porte des concepts différents et parfois contradictoires. Ces phrases comportent obligatoirement un seul verbe.

8.2. Les résultats

8.2.1. Temps d'apprentissage

Le point critique des réseaux de neurones est qu'ils sont trop gourmands en temps pendant la phase d'apprentissage. A cet effet le système a été testé avec des corpus de différentes tailles et en utilisant plusieurs seuils. Le tableau ci-après résume l'ensemble des résultats : le temps est exprimé en seconde.

Corpus	Corpus de 50 phrases		Corpus de 60 phrases		Corpus de 70 phrases		Corpus de 80 phrases		Corpus > 90 phrases	
Erreur	temps	taux	Temps	taux	Temps	Taux	Temps	taux	temps	Taux
0.20	8	73.31	11	79.42	20	82.95	27	85.85	32	88.42
0.19	10	73.95	14	79.42	24	84.24	33	86.81	40	90.03
0.18	13	73.95	17	79.42	31	86.17	41	87.78	49	90.35
0.17	16	74.91	21	80.70	40	87.13	53	88.74	63	90.67
0.16	21	74.91	29	81.02	51	87.45	67	89.06	80	91.31
0.15	26	75.56	32	80.70	64	87.13	86	90.99	104	91.31
0.14	31	77.49	39	81.67	82	87.45	110	92.28	136	92.28
0.13	39	77.49	49	81.99	109	87.78	141	92.28	177	94.53
0.12	50	78.45	62	82.63	145	87.78	179	92.28	230	94.85
0.11	56	79.74	80	82.95	194	87.78	227	92.60	304	96.14
0.10	83	79.09	109	82.31	257	87.45	296	92.6	423	96.46
0.09	106	79.42	151	82.63	336	89.06	391	92.60	616	96.78
0.08	132	79.09	208	83.60	423	89.06	543	92.60	821	97.10
0.07	161	80.70	301	84.56	537	88.74	766	92.92	1131	97.10
0.06	204	81.02	441	84.24	746	89.06	1252	92.92	1875	97.10
0.05	291	80.06	622	83.60	1159	89.71	2531	92.92	4013	97.10

Tableau 5. Temps d'apprentissage obtenus pour différents corpus en variant l'erreur

Nous remarquons que le taux d'apprentissage sur le corpus de plus de 90 phrases cesse d'augmenter à partir de l'erreur 0.08, ce qui justifie notre choix d'arrêter l'apprentissage à cette valeur.
Les graphes suivants montrent respectivement l'erreur calculée sur un mot, sur une phrase et sur le corpus.

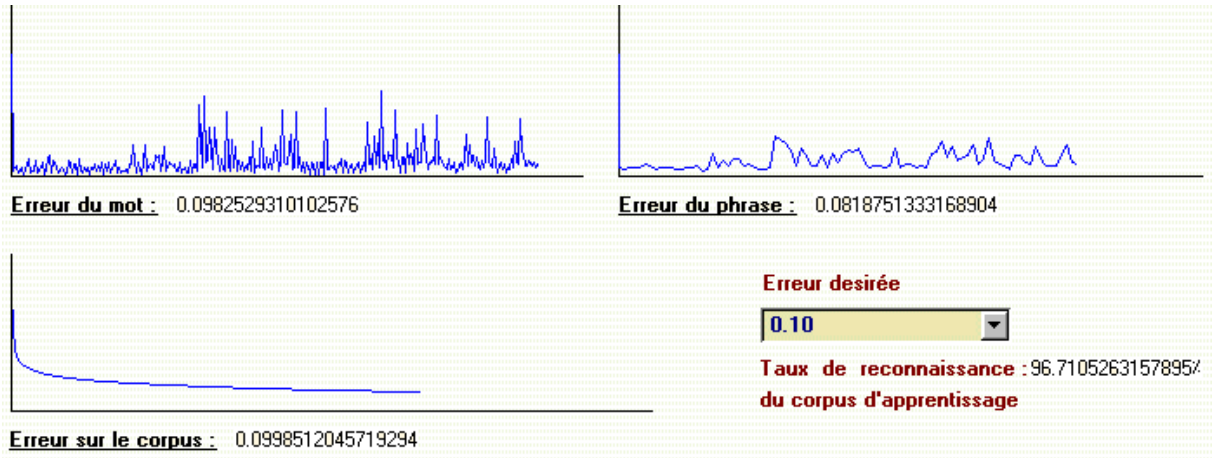

Figure 3. Graphes des erreurs calculées sur le mot, sur la phrase et sur le corpus

8.2.2. Généralisation

Pour tester le bon apprentissage du réseau, ce dernier a été testé avec différents corpus de test :

- Un corpus composé de phrase de deux mots
- Un corpus contenant des phrases de trois mots
- Un corpus de phrases composées de quatre mots
- Et finalement, un corpus général composé de phrases de tailles variables

Les résultats obtenus sont donnés dans le tableau suivant :

	Corpus de 2 mots	Corpus de 3 mots	Corpus de 4 mots	Corpus général
Taux de reconnaissance	100%	92.98%	96.82%	91.05%

Tableau 6. Taux de reconnaissance pour différents corpus

Ces résultats ont été reportés sur la figure suivante :

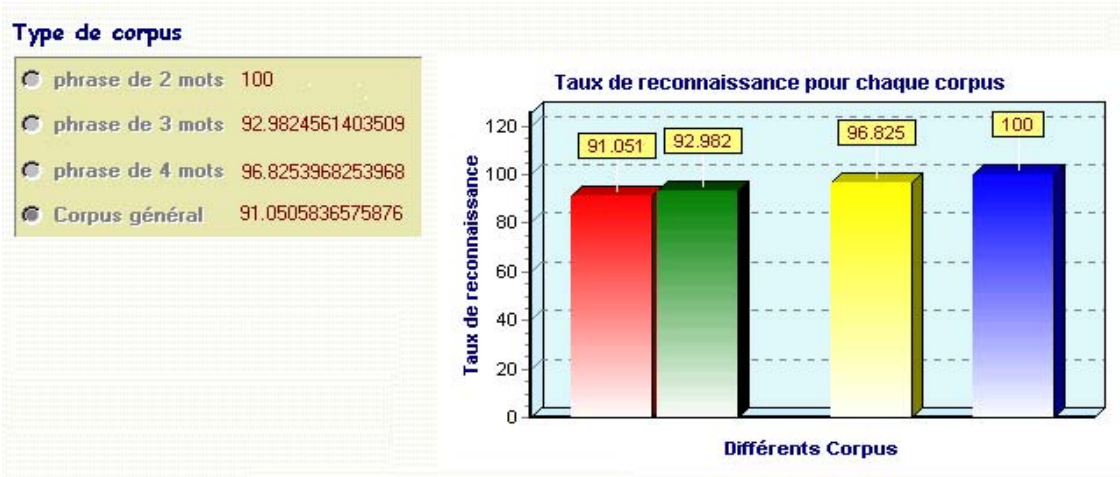

Figure 4. Taux de reconnaissance pour les différents corpus

9. Conclusion

Le travail que nous avons proposé concerne l'étude de l'utilisation des méthodes connexionnistes pour le traitement automatique du langage naturel et pour un système de compréhension en particulier.

Notre travail consistait donc a mettre au point un modèle connexionniste pour la génération de la représentation interne du sens d'une phrase écrite en langue Française. La même méthode a été également appliquée à la langue arabe ce qui nous permettra par la suite de réaliser un traducteur Arabe/Français des textes par une approche totalement connexionniste (Laskri and Mahdjoubi 1997 ; Meftouh and Laskri 2000).

La représentation du sens est basée sur la notion de cas sémantiques. L'apprentissage a été effectué d'une manière incrémentale : le réseau a été expérimenté sur plusieurs corpus et les résultats obtenus sont assez satisfaisants.

A l'issue de cette réalisation, des voies de recherche se dégagent :

- Dans un premier temps, étudier les modifications à apporter à ce modèle connexionniste pour le traitement de phrases récursives par intégration d'une RAAM (Recursive Auto Association Memory). La RAAM a été conçue spécialement pour la mise au point de structures de données complexes telles que les arbres, les listes et les piles (Pollack 1990).
- Et dans un second temps, intégrer le système de représentation du sens dans un système de traitement automatique du langage naturel totalement connexionniste y compris la phase de traduction automatique des textes.

Références

Archambault D. (1995). Proposition de réseaux neuromimétiques pour des traitements du langage naturel. *Thèse de doctorat, 3eme cycle*.

Berg G. (1992). A connectionist parser with recursive sentence structure and lexical desambiguisation. *Proceeding of American Association for Artificial Intelligence (AAAI)*, pp32-37

Courant Y. (1999). Analyse connexionniste robuste de la langue parlée, *DEA d'informatique, IMAG*

Elman J. (1990). Finding structure in time. *Cognitive science 14*, 179-211.

Elman J. (1993). Learning and development in neural networks: the importance of starting small. *Cognitive science 48*, 71-99.

Jodouin J.F. (1994). Les réseaux de neurones : principes et définitions

Laskri M.T. and Mahdjoubi R. (1997). Traitement automatique de la langue arabe en vue d'une traduction automatique vers la langue française. *1ère journées scientifiques et techniques, JST97, Francil de l'AUPELF.URF*. Avignon, France, 15-16 Avril 97

Mcclelland J.L. and St John M. and Taraban R. (1990). Sentence comprehension: APDP approach. *language and cognitive processes*, 287-335

Meftouh K. And Laskri M.T. (2000). *Generation of a sense of a sentence in Arabic language with connectionist approach*. *ACS/IEEE International Conference on Systems and Applications AICCSA'2001*. Beirut, Lebanon, June 26-29, 2001

Miikkulainen R. (1997). Natural language processing with Sub symbolic neural networks. *Neural Network Perspectives in Cognition and Adaptive Robotics*. A. Brown editor.

Pollack J. (1990). Recursive distributed representation., *Artificial Intelligence 46*, 77-105

Sabah G. (1988). *L'intelligence artificielle et le langage*, volume 1.

Sabah G. (1989). *L'intelligence artificielle et le langage*, volume 2.