Hypothèses pour la Construction et l'Exploitation Conjointe d'une Base Lexicale Sémantique Basée sur les Vecteurs Conceptuels

Didier Schwab, Mathieu Lafourcade, Violaine Prince

Hypothèses pour la construction et l'exploitation conjointe d'une base lexicale sémantique basée sur les vecteurs conceptuels.

Didier Schwab, Mathieu Lafourcade et Violaine Prince
{*schwab,lafourca,prince*}*@lirmm.fr*

LIRMM - Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.
http ://www.lirmm.fr/~{schwab, lafourca, prince}

Résumé

Dans le cadre de la représentation du sens en TALN, l'équipe TAL (Traitement Algorithmique des Langues) du LIRMM développe un système d'analyse des aspects thématiques des textes et de désambiguïsation lexicale basée sur les vecteurs conceptuels. Ces vecteurs cherchent à représenter un ensemble d'idées associées à tout segment textuel (mots, expressions, textes, ...). Associés à des informations de nature diverse, par exemple des fonctions lexicales, les vecteurs peuvent permettre de représenter la sémantique de ces segments. Dans cet article, nous exposons les hypothèses de départ que nous avons prises pour la construction d'une base lexicale sémantique. Nous présentons l'implémentation réalisée sous forme de système multi-agents et accessible en ligne ainsi que quelques expérimentations existantes ou en cours de réalisation.

Abstract

In the framework of meaning representation in NLP, the TAL research group at LIRMM develops a system for analysing thematic aspects of texts and for lexical disambiguation based on conceptual vectors. These vectors aim at representing sets of ideas related to any kind of textual segment (words, phrases, sentences, texts). Associated to various information, lexical functions for instance, vectors allow to represent the semantic of such textual segments. In this paper, we expose the hypothesis grounding the construction of a semantic lexical database. We present the overall organisation of such a system and an example of cooperation between agents in the context of a lexical item learning.

Mots clés : vecteurs conceptuels, représentation du sens, base lexicale sémantique, hypothèses de constructions, fonctions lexicales, systèmes multi-agents.

Keywords: conceptual vectors, sense representation, semantic lexical base, construction hypothesis, lexical functions, multi-agents systems.

1 Introduction

Dans le cadre de la représentation du sens en Traitement Automatique du Langage Naturel (TALN), l'équipe Traitement Algorithmique des Langues (TAL) du LIRMM développe actuellement un système d'analyse des aspects thématiques des textes et de désambiguïsation lexicale basée sur les vecteurs conceptuels. Ces vecteurs cherchent à représenter un

ensemble d'idées associées à tout segment textuel (mots, expressions, textes, ...). Associés à des informations lexicales, les relations sémantiques par exemple, les vecteurs peuvent permettre de représenter la sémantique de ces segments. Dans cet article, nous exposons les hypothèses de départ que nous avons prises pour la construction d'une base lexicale sémantique, c'est-à-dire les moyens utilisés pour fabriquer des objets appelés *acception* permettant de représenter le sens d'un maximum de termes de la langue. Ces hypothèses, au nombre de cinq, sont (I) *représentation du sens par une approche combinant approche thématique (vectorielle) et approche lexicale (relations sémantiques externes)*, (II) prise en compte des *relations sémantiques internes* (polysémie), (III) *génération automatique* des acceptions, (IV) réalisation d'une *analyse multi-sources* (à partir de dictionnaires, listes de synonymes, sites web, ...) et (V) *apprentissage permanent*. Ces hypothèses nous ont conduit à adopter une architecture multi-agents dont nous présentons ensuite succinctement l'implémentation accessible en ligne. Nous conclurons cet article par un exemple de collaboration entre agents ainsi que sur quelques expérimentations existantes ou en cours de réalisation.

2 Objectifs : approches polyvalentes

L'utilisation des vecteurs conceptuels se fait dans le cadre d'un système de TALN qui a pour objectif la représentation du sens et la création d'outils exploitant cette représentation. Les applications de ce système visent des domaines aussi variés que la recherche d'informations, la traduction ou le résumé automatique. Dès lors, nous développons un système qui doit être le plus générique et le plus évolutif possible. Il tente de représenter un maximum de termes du lexique des langues traitées, ce qui pose un problème d'acquisition des connaissances : un problème d'acquisition de ce lexique (acquérir un maximum d'*items lexicaux* [1] constituant ce lexique) mais aussi un problème d'acquisition du sens de ces items (en récupérant des données qui concernent le sens et en les exploitant). C'est cette double problématique que nous tentons de résoudre dans le cadre de ce projet.

3 Hypothèses de construction d'une base sémantique lexicale.

L'acquisition du sens des termes est basée sur cinq hypothèses fondamentales. C'est sur elles que reposent les architectures conceptuelles et implémentationnelles que nous avons choisies pour notre système.

3.1 Hypothèse I. Représentation du sens : approche combinant représentation thématique et informations lexicales

3.1.1 Vecteurs conceptuels

Nous représentons les aspects thématiques des segments textuels (documents, paragraphes, syntagmes, etc.) par des vecteurs conceptuels. Si les vecteurs ont été utilisés en informatique documentaire pour la recherche d'information dès la fin des années 1960 [Salton, 1968], leur emploi pour la représentation du sens est plus le fait du modèle LSI (*Latent Semantic Indexing* [Deerwester et al., 1990]) issu de l'analyse sémantique latente en psycholinguistique.

[1] Les items lexicaux sont des mots ou des expressions qui constituent les entrées du lexique. Par exemple, ‹*voiture*› ou ‹*pomme de terre*› sont des items lexicaux. Dans la suite de cet article, par abus de langage, nous utiliserons parfois mot ou terme pour qualifier un item lexical. Nous noterons les items en minuscule et entre apostrophes (‹*vie*›) et les concepts en majuscules (VIE).

En informatique, et de façon presque concurrente, c'est à partir de [Chauché, 1990] que l'on a une formalisation de la projection de la notion, linguistique cette fois, de champ sémantique dans un espace vectoriel. À partir d'un ensemble de notions élémentaires données *a priori* dont nous faisons l'hypothèse, les concepts (dans notre expérimentation sur le français nous utilisons [Larousse, 1992] dans lequel sont définis 873 concepts), il est possible de construire des vecteurs (dits conceptuels) et de les associer à des items lexicaux. Les termes polysémiques combinent les différents vecteurs correspondant aux différents sens. Cette approche vectorielle est fondée sur des propriétés mathématiques bien connues sur lesquelles il est possible d'effectuer des manipulations formellement pertinentes auxquelles sont rattachées des interprétations linguistiques appropriées. L'hypothèse principale du thésaurus, que nous adoptons ici, est que cet ensemble constitue un espace générateur pour les termes et leurs sens. De manière plus générale, n'importe quel sens peut s'y projeter selon le principe suivant :

Soit $\mathcal{C}$ un ensemble fini de n concepts. Un vecteur conceptuel V est une combinaison linéaire d'éléments c_i de $\mathcal{C}$. Pour une idée A, le vecteur V_A est la description en extension des activations de tous les concepts de $\mathcal{C}$. Par exemple, les différents sens d'‹*existence*› peuvent être projetés sur les concepts suivants (les CONCEPT[intensité] sont ordonnés par valeurs décroissantes) : V‹*existence*› = (EXISTENCE[0.82], VIE[0.44], IDENTITÉ[0.38], ÉTAT[0.33], ...). En pratique, plus $\mathcal{C}$ est grand, plus fines sont les descriptions de sens mais plus leur manipulation est lourde. Il est clair que pour les vecteurs denses, ceux qui ont peu de coordonnées nulles, l'énumération des concepts activés est longue et la pertinence difficile à évaluer. En général, pour évaluer la qualité d'un vecteur, nous préférerons sélectionner les termes thématiquement proches, le *voisinage* (noté $\mathcal{V}$). Par exemple, pour ‹*vie*› : $\mathcal{V}$(‹*existence*›) : ‹*existence*›, ‹*exister*›, ‹*vivant*›, ‹*vie*›, ... Cette opération est réalisée à l'aide de la distance angulaire.

Soit $Sim(X,Y)$ une des mesures de *similarité* entre deux vecteurs X et Y, souvent utilisée en recherche d'information [Morin, 1999]. $Sim(X,Y) = \cos(\widehat{X,Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$ avec "·" désignant le produit scalaire. Nous supposons ici que les composantes des vecteurs sont positives ou nulles, la *distance angulaire* entre deux vecteurs X et Y est $D_A(X,Y) = \arccos(Sim(X,Y))$. Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et en pratique la mesure de l'angle entre les deux vecteurs. Nous considérons en général que pour une distance $D_A(X,Y) \leq \frac{\pi}{4}$ (45°), X et Y sont thématiquement proches et partagent plusieurs concepts. Pour $D_A(X,Y) > \frac{\pi}{4}$, la proximité thématique est considérée comme faible et aux alentours de $\frac{\pi}{2}$ (90°), X et Y n'ont aucune relation. On remarquera que ces seuils ne servent que d'indicateurs pour un réviseur humain et restent à la fois subjectifs et arbitraires. D_A est une distance, elle vérifie donc les propriétés de réflexivité, de symétrie et d'inégalité triangulaire. Nous obtenons, par exemple, les angles suivants [2] :

D_A(V(‹*locomotive*›),V(‹*locomotive*›))=0 (0°) D_A(V(‹*locomotive*›),V(‹*rhododendron*›))=1.15 (65°)
D_A(V(‹*locomotive*›),V(‹*locomotrice*›))=0.24 (14°) D_A(V(‹*locomotive*›),V(‹*train*›))=0.54 (31°)
D_A(V(‹*locomotive*›),V(‹*automotrice*›))=0.22 (13°) D_A(V(‹*locomotive*›),V(‹*guépard*›))=0,94 (54°)

Le premier résultat a une interprétation directe, ‹*locomotive*› ne peut être plus proche d'autre chose que de lui même. Les termes ‹*automotrice*› et ‹*locomotrice*› sont quasi-synonymes de ‹*locomotive*›, ce qui explique les deux résultats suivants. Le peu de rapport

[2] Les exemples sont extraits de `http ://www.lirmm.fr/~schwab`

entre ‹*locomotive*› et ‹*rhododendron*› explique l'écart entre leur vecteur respectif. Dans le dernier exemple, l'angle peu important entre ‹*locomotive*› et ‹*guépard*› au regard de celui entre ‹*locomotive*› et ‹*rhododendron*› se comprend si l'on se rappelle que D_A est une distance thématique et non une distance ontologique ou sémantique. Les deux items ont en commun de partager une idée de rapidité. On remarquera que les comparaisons entre les valeurs sont plus significatives que les valeurs elles-mêmes.

3.1.2 Informations lexicales : fonctions lexicales

Nous venons de le voir, les vecteurs conceptuels constituent une représentation thématique et non une représentation sémantique. En psycholinguistique, il est aujourd'hui bien accepté que le sens est issus à la fois du niveau des vocables et du niveau des thèmes [Kawamoto, 1993]. Des informations spécifiques de rapports de sens que les termes entretiennent entre eux sont donc indispensables à notre modèle. Ces rapports peuvent être modélisés sous la formes de fonctions lexicales à la Mel'čuk. Par exemple, ‹*autoriser*› entretient une relation d'antonymie avec ‹*interdire*› ou ‹*défendre*› et une relation de synonymie avec ‹*permettre*›.

3.2 Hypothèse II. Relations sémantiques internes d'un item lexical

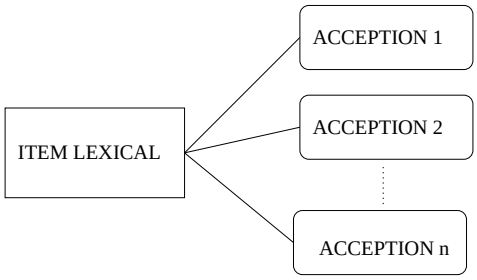
Les *relations sémantiques internes* sont les diverses relations qu'il existe entre les différents sens d'un même item lexical. Ces relations concernent ce que l'on appelle *l'aire sémantique* d'un mot, c'est-à-dire l'ensemble des significations qu'il est susceptible de prendre. On distingue la *monosémie* qui concerne les termes qui n'ont qu'un seul sens (‹*calame*›, ‹*cajou*›, ‹*neuroleptique*›, ...), et la *polysémie* qui concerne les termes qui ont plusieurs sens[3] (‹*louer*›, ‹*froid*›, ‹*frégate*›, ...).

Les relations sémantiques internes d'un item lexical, le fait qu'il puisse avoir plusieurs sens nous obligent à en tenir compte dans la construction de nos vecteurs conceptuels. Nous pensons qu'il est difficile de considérer qu'une quelconque tâche de désambiguïsation est possible sans que chaque sens d'un item lexical ne soit aussi affecté d'un vecteur conceptuel (chacun de ces vecteurs entrant dans la construction du vecteur global sans doute de façon non-linéaire).

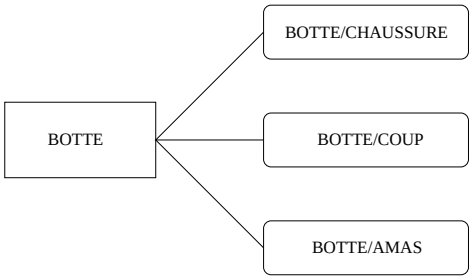
Nous appelons *acception* un sens particulier d'un item lexical admis et reconnu par l'usage. Il s'agit d'une unité sémantique propre à une langue donnée [Sérasset et Mangeot, 2001]. Par exemple, l'item lexical ‹*botte*› a au moins trois acceptions : la *« chaussure »*, l'*« amas de végétaux »* ou le *« coup en escrime »*. Contrairement aux items lexicaux, les acceptions sont donc monosémiques.

Nous avons tenu compte de ces deux premières hypothèses pour construire les objets acceptions. Les acceptions sont composées d'un certain nombre d'**informations linguistiques** comme la **morphologie** (composé des **catégories grammaticales** : *nom, pronom, adjectif, verbe, etc.*, du **genre** : *masculin, féminin, neutre* et le *nombre* : *singulier, pluriel*), la **fréquence en usage** (le nombre de fois ou au moins une évaluation où l'acception a

[3] Nous confondons volontairement ici l'homonymie et la polysémie. Le problème de leur différentiation est assez peu intéressant au niveau du TALN en général et de la désambiguïsation en particulier. Notre but est de trouver le meilleur sens pour un terme et non de savoir comment le terme a acquis ce nouveau sens ni quels sont les rapports qu'entretiennent ces sens entre eux. Nous parlerons dans la suite de cet article d'items polysémiques sans distinguer s'il s'agit d'une vraie polysémie ou d'un cas d'homonymie.



(a) Organisation générale de la représentation du sens pour un item lexical. Le vecteur conceptuel général d'un item est calculé à partir des vecteurs de chacune des acceptions de cet item.



(b) Organisation générale de la représentation du sens pour l'item lexical ‹*botte*›

FIG. 1 – Organisation générale de la représentation du sens pour un item lexical.

été rencontrée), **un vecteur conceptuel**, des **fonctions lexicales**, des **informations étymologiques**, des **gloses** (des informations que l'on trouve, par exemple, dans les dictionnaires de traduction ou de synonymie pour préciser le sens d'un mot).

3.3 Hypothèse III. Génération automatique

Notre objectif est donc de construire une base de stockage d'objets acception. La difficulté principale vient de la création de ces objets. Dans des dictionnaires francophones classiques comme le [Larousse, 2001] et le [Hachette, 2000], sur les 80000 items (mots communs, nom propres, expressions, ...) répertoriés, une majorité est polysémique. Dans notre expérience sur le français, pour un peu plus de 100000 entrées (mots communs, nom propres, expressions, ...), le taux de termes polysémiques est d'environ 60%. Le nombre moyen de définitions pour ces derniers étant d'un peu plus de 5, il faudrait indexer à la main plus de 400000 acceptions, ce qui paraît irréaliste.

Nous considérons qu'il est possible d'automatiser cette tâche grâce à un apprentissage basé sur des informations extraites de diverses sources (dictionnaires, listes de synonymes, indexations manuelles, recherches web, ...). Pour chaque item lexical, chacune de ces sources peut nous donner une ou plusieurs définitions. Il est possible d'extraire un certain nombre d'informations de ces définitions. Les plus riches sont celles extraites des dictionnaires, elles peuvent permettre d'obtenir des informations d'ordre morphologique et éventuellement étymologique. À partir du texte des définitions, il est aussi possible de calculer un vecteur conceptuel. L'objet qui regroupe toutes les informations qu'il est possible d'extraire d'une définition est appelé *lexie*. Sa structure interne est identique à celle d'une acception (cf. hypothèse I). L'essentiel de l'apprentissage principal se fait sur des dictionnaires à usage humain (dictionnaires classiques, de synonymes, d'antonymes, ...) ou machinaux. Dans la perspective où nous nous plaçons, dans un espace dimensionné en fonction d'une hiérarchie de concepts, un amorçage du système d'apprentissage est nécessaire. Il s'agit d'affecter des vecteurs à un nombre réduit d'entités qui sont choisies en fonction de leur fréquence en langue et/ou de leur polysémie. La taille du noyau est très réduite (un millier de termes environ) et les éléments de ce noyau considérés comme pertinents. À partir de ce noyau, le processus d'apprentissage peut débuter. La méthode d'analyse construit, à partir de vec-

teurs conceptuels déjà existants et de nouvelles définitions, de nouveaux vecteurs. L'idée est qu'à partir d'un noyau réduit d'items pertinents, un apprentissage sur des définitions permet de créer une cohérence entre les vecteurs et donc de générer une base d'items pertinents.

3.4 Hypothèse IV. Analyse multi-sources

L'analyse des définitions pose un certain nombre de problèmes quant à la lecture du sens. C'est le cas du métalangage, c'est-à-dire le langage utilisé pour structurer le dictionnaire. Ce dernier est aisément utilisable lorsqu'il s'agit de récupérer les catégories grammaticales des items mais certaines constructions de définitions sont difficilement compréhensibles sans compétence métalinguistique. Il existe un certain nombre de phrases qu'il convient d'analyser convenablement(*partie de*, *se dit de*, etc.) ou d'ambiguïtés que même un humain ne peut lever. Comment savoir, par exemple, que *en parlant* est du métalangage dans une définition de l'item ‹*aboyer*› comme « *Crier, en parlant du chien* »? Pour pallier de tels manques définitoires, nous utilisons diverses sources lexicales. Il s'agit de tempérer statistiquement les diverses incohérences locales. Ainsi, si une définition est mal formée (donc difficilement analysable correctement), une autre définition, mieux formée et provenant d'une autre source, pourra corriger l'effet de la première.

Aucune source ne peut couvrir l'intégralité du lexique non seulement parce que celui-ci est en constance évolution mais surtout parce que rechercher, trouver et finalement décrire de façon systématique chaque terme de la langue est une tâche au mieux extrêmement difficile. Il faudrait, en effet, connaître l'ensemble des formes mais aussi l'ensemble des sens que peuvent prendre ses formes dans n'importe quel domaine d'activité, dans n'importe quelle structure sociale et quel que soit le niveau de langage. L'utilisation de multi-sources permet donc aussi de maximiser la probabilité de récupérer une définition pour certains termes qui pourraient être trouvés dans certains dictionnaires et pas dans d'autres. Par exemple, ‹*liturgiste*› ne se trouve pas dans [Larousse, 2001] mais on le trouve dans [Hachette, 2000].

La figure 2 présente la structuration générale de la base sémantique lexicale. À partir de diverses sources, on peut obtenir des définitions pour un item lexical. Pour chaque définition, on crée une lexie qui est un objet qui regroupe les diverses informations contenues dans cette définition dont un vecteur conceptuel calculé à partir du texte de la définition. Ces lexies doivent alors être catégorisées (regroupées par sens) afin de fabriquer les diverses acceptions de l'item.

Prenons un exemple. Considérons trois sources, le *Petit Larousse* [Larousse, 2001], le dictionnaire *Hachette de la langue* [Hachette, 2000] et l'*Encyclopédie Club-internet*[4]. Pour l'item ‹*botte*›, nous trouvons les définitions suivantes :

botte.1 : #nf# Réunion de végétaux de même nature liés ensemble. (Une botte de paille, de radis, de fleurs) . *[Hach]*

botte.2 : #nf# En escrime, coup porté à l'adversaire avec un fleuret ou une épée. (Pousser, porter, parer une botte) (Botte secrète.). *[Hach]*

botte.3 : #nf# Chaussure de cuir, de caoutchouc ou de plastique qui enferme le pied et la jambe, parfois la cuisse. (Des bottes de cavalier) – Chaussure d'extérieur basse. (Botte d'hiver, de ski, de marche) . *[Hach]*

[4] http ://www.club-internet.fr

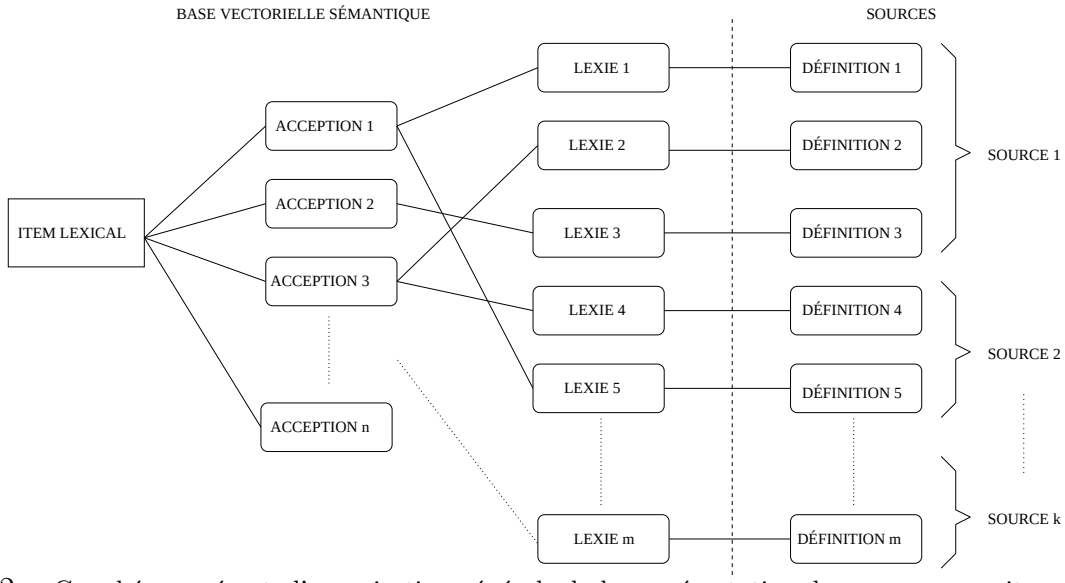


FIG. 2 – Ce schéma présente l'organisation générale de la représentation du sens pour un item lexical. À partir des informations données par les définitions, on crée des lexies. Ces lexies sont ensuite utilisées pour fabriquer les acceptions correspondant à chaque sens du mot.

botte.4 : #nf# (néerl. bote, touffe de lin) . Assemblage de végétaux de même nature liés ensemble : (Botte de paille. Botte de radis.) . *[Lar]*
botte.5 : #nf# (#ethym-it# botta, coup) . Coup de pointe donné avec le fleuret ou l'épée . *[Lar]*
botte.6 : #nf# (p.-ê. de bot) . Chaussure à tige montante qui enferme le pied et la jambe généralement jusqu'au genou : (Bottes de cuir) . *[Lar]*
botte.7 : #nf# (néerl. (bote) « touffe de lin »). Assemblage de végétaux, de même sorte, tenus par un lien. (Une botte de paille, d'asperges.). *[Club]*
botte.8 : #nf# (ital. (botta) « coup »). En escrime, coup de pointe à l'épée ou au fleuret. (Porter une botte.) #fig# Attaque vive . *[Club]*
botte.9 : #nf# Chaussure montant jusqu'à mi-jambe ou jusqu'au genou, enfermant parfois une partie de la cuisse. (Des bottes de pêche, d'équitation.). *[Club]*

Il semble clair que les sens {1,4,7} peuvent se regrouper pour former le sens d'*« amas »*, les sens {2,5,8} pour former le sens de *« coup »* et les sens {3,6,9} pour former l'acception *« chaussure »*. La figure 3 présente l'organisation générale de l'item ‹*botte*› telle qu'elle serait si celui-ci était défini par ces neuf définitions.

3.5 Hypothèse V. Apprentissage permanent

Pour analyser un certain nombre de documents, en particulier les journaux, il est souvent nécessaire de savoir à quoi correspondent des néologismes, qui sont certaines personnalités ou encore quel est le domaine d'activité d'une entreprise. Par exemple, dans un texte, l'usage du nom de l'entreprise ‹*Arcelor*› indique vraisemblablement un contexte axé sur le traitement de l'acier. Les diverses sources et en particulier le web par les serveurs d'informations (*Le Monde, Libération, ...*) peuvent permettre les informations nécessaires à la fabrications des structures appropriées.

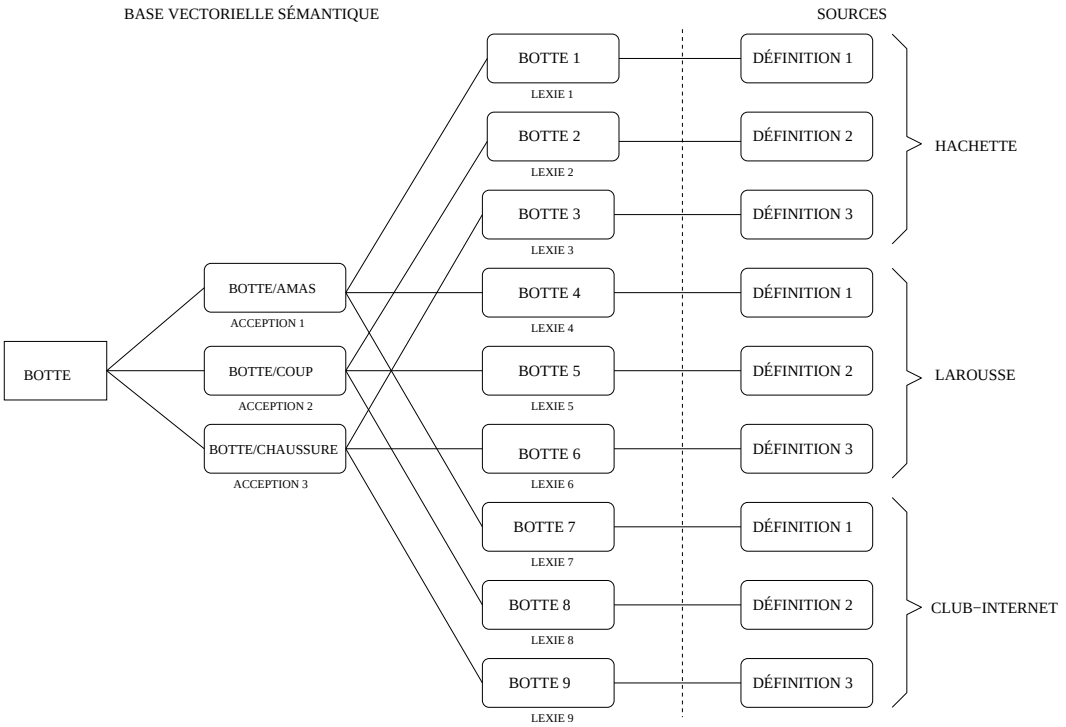


FIG. 3 – Ce schéma présente l'organisation générale de la représentation du sens pour un item lexical. À partir des informations données par les définitions, on crée des lexies. Ces lexies sont ensuite utilisées pour fabriquer les acceptions correspondant à chaque sens du mot.

De plus, il est difficile de penser (notre expérimentation nous l'a montré) que les vecteurs deviendraient cohérents dès la première passe. Il est vraisemblable que des mots clés d'une définition n'aient pas encore été appris par le système lors de son analyse. La convergence des vecteurs vers une position quasi-stable ne pourra se faire que dans un nombre de cycles qu'il est impossible de déterminer à l'avance mais qui est fonction de la taille et de la richesse du vocabulaire utilisé dans les définitions. Ces deux raisons, la variabilité lexicale et l'impossibilité de véritablement stabiliser une base, nous ont conduits à considérer cette cinquième et dernière hypothèse : la base est en apprentissage permanent.

4 Vers une société d'agents

Nous venons de le voir, notre objectif est la création d'un véritable système permettant l'apprentissage d'informations sémantiques (thématiques par les Vecteurs Conceptuels et lexicales par les relations sémantiques) et leur exploitation. Il s'agit de récupérer des définitions, les analyser à l'aide des *lexies* et *acceptions* déjà calculés afin d'en fabriquer ou d'en réviser d'autres (cf. section 3). L'analyse simple des définitions peut ne pas toujours suffire à améliorer la cohérence des vecteurs (difficulté d'analyser certaines tournures, problèmes de métalangage, ...). Plusieurs solutions sont alors possibles comme par exemple, l'utilisation des relations sémantiques existant entre les items (synonymie [Lafourcade et Prince, 2001], antonymie [Schwab et al., 2002], hypéronymie, ...). Ces relations interviennent alors à deux niveaux : au niveau de la construction, pour fabriquer les acceptions et au niveau de l'exploitation puisqu'elles participent alors à une meilleure construction du sens.

L'exploitation de la base peut aussi être plurielle : utilisation pour la désambiguïsation sémantique, annotation, transfert lexical, traduction, recherche d'informations. À la fois pour l'exploitation et l'apprentissage, il est donc nécessaire de pouvoir facilement ajouter des modules apportant tel ou tel service. C'est une des raisons pour lesquelles notre vision de l'architecture nécessaire s'est rapidement rapprochée des systèmes multi-agents (SMA).

4.1 SMA et TALN

Les SMA sont issus de l'Intelligence Artificielle Distribuée (IAD) qui répartit l'intelligence dans des agents. Tout ou partie de cette intelligence est la conséquence de leurs interactions (phénomène d'émergence). Un agent est une entité physique ou virtuelle (virtuelle dans notre cas) capable d'agir sur son environnement (les autres agents), qui peut communiquer directement avec d'autres agents, qui possède des ressources propres, qui est capable de percevoir son environnement, qui possède des compétences et offre des services [Ferber, 1995]. Les SMA sont utilisés depuis longtemps dans le domaine des langues naturelles. Dès le début des années 70, des travaux en IAD ont été effectués sur la compréhension automatique de la parole (HEARSAY-II [Erman et al., 1980]). Plus récemment, [Lebarbé, 2001] utilise cette approche pour l'analyse syntaxique tandis que [Menézo et al., 1996] l'utilisent pour la détection d'erreurs. Parmi les autres travaux, certains se rapprochent de ce que nous voulons faire, une architecture modulaire permettant l'apprentissage de données et leur utilisation. On peut citer le système CARAMEL [Sabah, 1990] ou le système TALISMAN [Stéfanini et al., 1992].

Les systèmes SMA offrent deux avantages techniques importants qui nous ont conduit a choisir ce mode d'implémentation. (1) la *possibilité de distribuer sur plusieurs machines* : les systèmes TALN ont toujours été consommateurs de ressources systèmes importantes dues aux données à stocker (la taille du lexique est d'au moins 100000 entrées pour une langue comme le Français) et aux calculs souvent lourds. Chaque agent peut se trouver sur une machine. Ainsi, il pourra pleinement utiliser les ressources matérielles disponibles. La communication entre agents se fait alors par accès réseaux sur le modèle client-serveur. (2) La *modularité* qui est une des caractéristiques principales des architectures agent. Un service global résulte de la coopération de chaque agent qui effectue une sous-tâche moins complexe. Les avantages génie-logiciel sont nombreux : un développeur peut facilement créer un agent dans le langage informatique de son choix, pourra aisément le tester et l'améliorer indépendamment des autres. L'ajout d'un programme déjà existant est simplifié. Dans notre application, par exemple, il a été extrêmement facile de créer un agent d'analyse morpho-syntaxique qui n'est qu'une interface à l'analyseur SYGMART [Chauché, 1984].

4.2 Exemple de société d'agents : implémentation

4.2.1 description du prototype

Nous avons donc réalisé un prototype[5] sous la forme d'un système multi-agents. Chaque agent possède un certain nombre de compétences, une mémoire qui lui est propre et peut interagir avec son environnement (les autres agents). La gestion des agents se fait par un référenceur. Lors de sa création, chaque agent adresse au référenceur son identifiant, son rôle, éventuellement sa langue[6], ainsi que la machine et le port sur lequel il écoute. Le

[5] Accessible en ligne à l'adresse `http ://www.lirmm.fr/~schwab`.

[6] certains agents peuvent être indépendants de la langue. C'est le cas, par exemple, d'agents dont les données ne sont pas de type lexical.

référenceur accepte la création de l'agent si aucun autre agent encore actif ne présente cet identifiant. Plusieurs types de communications sont possibles : (1) Communication directe entre agents : comme pour le *pair à pair* (*peer to peer*), un agent demande au référenceur l'adresse particulière d'un agent puis communique directement avec lui. C'est le cas, par exemple, d'un agent *contextualiseur* qui a besoin de récupérer très souvent les informations sur les lexies (cf 4.2.2) que stocke la *base*. (2) Communication par l'intermédiaire du référenceur : suivant la requête, les messages peuvent être envoyés à un agent, à tous les agents ayant un certain rôle, à tous les agents d'une langue ayant un certain rôle et même à tous les agents. En pratique, le message est envoyé au référenceur qui se charge de l'envoyer à ses destinataires. Chaque agent qui a reçu le message y répond en apportant une réponse ou en signifiant sa non-compétence dans ce domaine.

4.2.2 Agents implémentés

En Octobre 2003, plus d'une dizaine de types d'agents étaient implémentés.

- *Base lexicale* : ces agents implémentent l'architecture présentée en 3.4.
- *Contextualiseur* : cette sorte d'agents est capable de calculer le vecteur conceptuel d'un item en fonction de contextes sémantiques et morphologiques. En pratique, il fait une somme pondérée des vecteurs des *acceptions* du terme en fonction d'un vecteur contexte, d'informations morphologiques et statistiques (fréquence en corpus).
- *Analyseur morpho-syntaxique* : il s'agit de l'analyseur SYGMART [Chauché, 1984]. Pour un texte, il renvoie l'arbre morpho-syntaxique correspondant.
- *Analyseur sémantique* : avec l'aide de l'agent d'analyse morpho-syntaxique et l'aide de l'agent contextualiseur, l'agent d'analyse sémantique calcule le vecteur correspondant à un texte. Ce texte, dans le cas d'un apprentissage, est une définition de dictionnaire dont l'agent d'apprentissage souhaite le vecteur.
- *Catégoriseur* [Jalabert, 2003] qui catégorise les *lexies* pour fabriquer les *acceptions*.
- *Apprentissage* : Cet agent gère l'apprentissage des lexies. Il est directement aidé dans cette tâche par des agents d'analyse sémantique ainsi que par les agents extracteurs de définitions.
- *Agents extracteurs de définitions* : ces agents ont pour rôle de récupérer les définitions correspondant à des items et de les fournir à l'agent d'apprentissage.
- *Agents calculant les fonctions lexicales* : ces agents sont des experts des relations sémantiques comme la synonymie, l'antonymie, l'hypéronymie ou toute autre fonction lexicale.
- *Agents extrayant les fonctions lexicales* : ces agents extraient des dictionnaires spécialisés ou de définitions de dictionnaires classiques des relations sémantiques qui existent entre les items.

4.2.3 Exemple d'interaction entre agents, apprentissage du vecteur d'un item

Dans cette section, nous illustrons la collaboration entre agents par l'exemple de la fabrication d'objets *acceptions* et *lexies* d'un item lexical. L'agent d'*apprentissage* demande à un agent *extracteur de définitions* de lui fournir les définitions de cet item qu'il a récupérées en explorant le web ou des dictionnaires à usage humain. Le texte de chaque définition est donné à l'agent d'*analyse sémantique* (1) qui, à partir de l'arbre morpho-syntaxique obtenu grâce à l'agent *analyseur morpho-syntaxique* (2)-(3), calcule le vecteur conceptuel

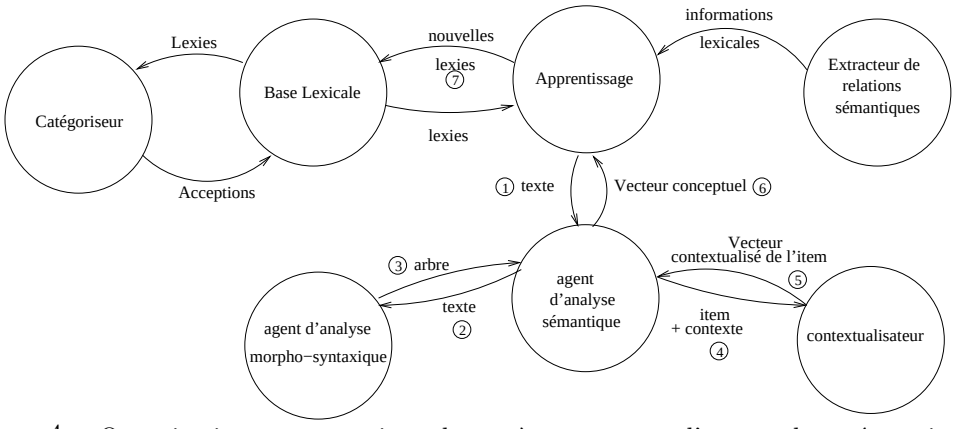


FIG. 4 – Organisation macroscopique du système au cours d'une analyse sémantique.

correspondant à la définition (6). L'agent *contextualiseur* (lui-même aidé par la *base de vecteurs*) et éventuellement des agents experts en *relations sémantiques* collaborent avec lui dans cette tâche (4)-(5). L'agent *apprentissage* récupère chaque vecteur des définitions. Les *agents extrayant les fonctions lexicales* lui permettent de compléter leurs informations pour construire les *lexies* qu'ils fournissent alors à la *base lexicale* (7). Parallèlement à ces opérations, l'agent *catégoriseur* explore la base et fabrique des *acceptions* à partir des *lexies* (8).

5 Conclusion et perspectives

Dans cet article, nous présentons l'architecture d'une base lexicale sémantique générique utilisable par des applications de TALN nécessitant l'usage de la sémantique comme, par exemple, la recherche d'informations, la traduction automatique ou le résumé automatique. À chaque item lexical sont associées des acceptions qui rassemblent les informations correspondants aux différents sens que l'item peut prendre. Ces acceptions sont fabriquées automatiquement à partir de définitions issues de différentes sources. Cette architecture se fonde sur plusieurs hypothèses. Nous avons montré que nous devions représenter le sens par une approche combinant approche thématique (par vecteurs conceptuels) et approche lexicale (par une explicitation des relations sémantiques externes) afin de tenir compte des informations spécifiques sur les rapports de sens que les termes entretiennent entre eux. De même, il est nécessaire de tenir compte des relations sémantiques internes (polysémie). Il convient aussi de réaliser une analyse multi-sources afin d'obtenir un maximum d'informations sur le lexique pour ne pas passer à coté d'informations qui pourraient se révéler importantes ainsi qu'un apprentissage permanent pour pouvoir tenir compte des informations extraites par certaines sources dynamiques comme le web. Nous avons finalement présenté certains agents qui ont été implémentés et qui sont accessibles en ligne.

Dans la suite de nos travaux, nous allons continuer à travailler sur chaque agent afin d'améliorer leurs tâches respectives. De nouveaux seront ajoutés essentiellement en ce qui concerne d'autres fonctions lexicales comme les fonctions hiérarchiques (hypéronymie, holonymie, ...). Nous allons aussi travailler sur l'amélioration des modes de communication entre agents en analysant, par exemple, les moyens techniques qui permettraient aux agents de seconder un autre agent sans que celui-ci ait à demander de l'aide.

Références

[Chauché, 1990] Chauché J. (1990). Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TAL Information*, pages 17–24.

[Chauché, 1984] Chauché J. (1984). Un outil multidimensionnel de l'analyse du discours. In *Coling'84*, pages 11–15, Stanford University, California.

[Deerwester et al., 1990] Deerwester S. C., Dumais S. T., Landauer T. K., Furnas G. W., et Harshman R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6) :391–407.

[Erman et al., 1980] Erman L., Hayes-Roth F., Lesser V., et Reddy D. (1980). The hearsay-ii speech understanding system : Integration knowledge to resolve uncertainly. In *ACM Computing Surveys*, volume 12.

[Ferber, 1995] Ferber J. (1995). *Les systèmes multi-agents. Vers une intelligence collective.* InterEditions.

[Hachette, 2000] Hachette (2000). *Le Nouveau Petit Robert, dictionnaire alphabétique et analogique de la langue française*. Hachette.

[Jalabert, 2003] Jalabert F. (2003). Catégorisation de définitions et nommage de sens. Mémoire de dea, Université Montpellier II, LIRMM.

[Kawamoto, 1993] Kawamoto A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity : A parallel distributed processing account. *Journal of Memory and Language*, 32 :474–516.

[Lafourcade et Prince, 2001] Lafourcade M. et Prince V. (2001). Synonymies et vecteurs conceptuels. In *TALN'2001*, Tours, France.

[Larousse, 1992] Larousse (1992). *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse.

[Larousse, 2001] Larousse (2001). *Le Petit Larousse Illustré 2001*. Larousse.

[Lebarbé, 2001] Lebarbé T. (2001). Vers une plate-forme multi-agents pour l'exploration et le traitement linguistique. In *Traitement Automatique du Langages Naturel (TALN'2001)*, Tours, France.

[Menézo et al., 1996] Menézo J., Genthial D., et Courtin J. (1996). Reconnaissances pluri-lexicales dans celine, un système multi-agents de detection et correction des erreurs. In *NLP+IA96 : International Conference on Natural Language Processing and Industrial Applications.*, pages 174–180.

[Morin, 1999] Morin (1999). Extraction de liens sémantiques entre termes à partir de corpus techniques. Thèse de doctorat, Université de Nantes.

[Sabah, 1990] Sabah G. (1990). Caramel : Un système multi-expert pour le traitement automatique des langues. *Modèles linguistiques*, tome XII, fascicule 1.

[Salton, 1968] Salton G. (1968). *Automatic Information Organisation and Retrieval*. McGrawHill, New York.

[Schwab et al., 2002] Schwab D., Lafourcade M., et Prince V. (2002). Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. l'exemple de l'antonymie. In *TALN 2002*, volume 1, Nancy.

[Stéfanini et al., 1992] Stéfanini M.-H., Berrendonner A., Lallich G., et Oquendo F. (1992). Talisman : un système multi-agents gouverné par des lois linguistiques pour le traitement de la langue naturelle. In *Coling 1992*, Nantes.

[Sérasset et Mangeot, 2001] Sérasset G. et Mangeot M. (2001). Papillon lexical databases project : monolingual dictionaries and interlingual links. In *NLPRS 2001*, pages 119–125.